

RESEARCH

Open Access



Inferring person-to-person networks of *Plasmodium falciparum* transmission: are analyses of routine surveillance data up to the task?

John H. Huber^{1*} , Michelle S. Hsiang^{2,3,4}, Nomcebo Dlamini⁵, Maxwell Murphy⁶, Sibonakaliso Vilakati⁵, Nomcebo Nhlabathi⁵, Anita Lerch¹, Rasmus Nielsen⁷, Nyasatu Ntshalintshali⁸, Bryan Greenhouse^{6,9} and T. Alex Perkins^{1*}

Abstract

Background: Inference of person-to-person transmission networks using surveillance data is increasingly used to estimate spatiotemporal patterns of pathogen transmission. Several data types can be used to inform transmission network inferences, yet the sensitivity of those inferences to different data types is not routinely evaluated.

Methods: The influence of different combinations of spatial, temporal, and travel-history data on transmission network inferences for *Plasmodium falciparum* malaria were evaluated.

Results: The information content of these data types may be limited for inferring person-to-person transmission networks and may lead to an overestimate of transmission. Only when outbreaks were temporally focal or travel histories were accurate was the algorithm able to accurately estimate the reproduction number under control, R_c . Applying this approach to data from Eswatini indicated that inferences of R_c and spatiotemporal patterns therein depend upon the choice of data types and assumptions about travel-history data.

Conclusions: These results suggest that transmission network inferences made with routine malaria surveillance data should be interpreted with caution.

Background

Concomitant with improved epidemiological surveillance, there is growing interest to leverage the collected data to infer transmission networks for a wide range of pathogens and to use those inferences to inform public health efforts. Past studies have incorporated temporal data [1] and spatial data [2–5] to estimate pairwise probabilities of transmission between individual cases

and to use those estimates to infer time-varying and spatially varying reproduction numbers, respectively. More recently, methods have been developed to incorporate this type of detailed, individual-level epidemiological data [6–8] to infer transmission networks for infectious diseases of humans, including severe acute respiratory syndrome [9] and tuberculosis [10], and of animals, such as rabies [11] and foot-and-mouth disease [12].

In addition to the diseases for which these methods have been applied to date, there is a growing need to apply similar methods to malaria in near-elimination settings. As incidence of malaria declines within a country, transmission becomes more heterogeneous in space

*Correspondence: jhuber3@nd.edu; taperkins@nd.edu

¹ Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

and time [13]. Focal areas of high transmission, known as ‘hotspots’, pose a serious risk of fuelling resurgence if left untargeted, potentially reversing decades of progress towards elimination [14]. To this end, granular estimates of when and where transmission occurs are needed, as spatially aggregated estimates may obscure important heterogeneities of practical relevance to control efforts [15]. In addition to characterizing details of local transmission, measurement of progress towards malaria elimination hinges on correct classification of cases as imported (i.e., acquired outside the country) or locally acquired [16, 17], which is a byproduct of estimating transmission networks.

Previous work on malaria has made progress on the use of individual-level epidemiological data to infer transmission networks and reproduction numbers of *Plasmodium falciparum*, the parasite primarily responsible for human malaria in many regions of the world. Churcher et al. [18] used temporal data to estimate the proportion of imported cases needed to confidently estimate the reproduction number under control, R_c , below one and thereby provide evidence of controlled, non-endemic malaria transmission. Reiner et al. [6] then built upon this work by incorporating spatial data and inferring an individual-level transmission network of *P. falciparum* in Eswatini. More recently, Routledge et al. [19, 20] used related approaches to infer transmission networks and R_c of *Plasmodium vivax* in El Salvador and China.

As the adoption of these methods increases, in particular for malaria, care should be taken to assess how the epidemiological setting and the inclusion or exclusion of certain data types might affect the accuracy of transmission network inferences, as well as resultant inferences about epidemiological quantities, including R_c and spatiotemporal variation therein. A recent study by Campbell et al. [21] noted that epidemiological data alone were generally insufficient to reconstruct transmission networks of other pathogens, ranging from *Mycobacterium tuberculosis* to SARS-CoV. Although falciparum malaria was not considered in that analysis, its long serial interval [22] calls into the question the utility of epidemiological data for this purpose, though this has been largely unaddressed in past studies. Furthermore, past transmission network inferences for malaria have relied on various types of epidemiological data, ranging from the timing of symptom onset [19, 20] to more detailed spatiotemporal data [6]. Each study incorporated travel-history information into transmission network inferences and considered these data to be perfectly accurate, assuming that all cases that reported travel were imported. However, travel history may be an imperfect indicator of importation owing to errors in recall [17] and the fact that travel to

an area of ongoing transmission alone does not guarantee that an individual was infected there [17, 23]. *Plasmodium falciparum* transmission network inferences are likely to be sensitive to the choice of data types [24], and failure to evaluate the sensitivity of transmission network inferences to choices about data types and different assumptions about possible errors in travel-history data could lead to apparently confident, though ultimately incorrect, assessments of *P. falciparum* transmission risk in near-elimination settings.

Here, a Bayesian method for inferring transmission networks based on temporal, spatial, and travel-history data for individual malaria cases is used to characterize the sensitivity of transmission network inferences to the inclusion of different data types and to different assumptions about the accuracy of travel histories. This method builds upon previous work by leveraging individual-level epidemiological data to obtain posterior estimates of transmission networks and model parameters in a way that can accommodate different assumptions about errors in travel histories. After establishing a proof-of-concept of the inference method on simple test cases, the method was applied to real-world surveillance data from Eswatini and additional simulated data sets to understand how the inclusion or exclusion of different data types and different assumptions about travel-history error affect the ability to infer transmission networks and estimate transmission metrics, namely R_c .

Methods

Bayesian framework for estimating transmission linkages

The goal was to obtain probabilistic estimates of a transmission network N that defines transmission linkages among a set of known cases. The transmission network is defined as a directed, acyclic graph comprised of a set of directed edges represented as $N = \{N_{i,j}\}$ for all i, j . Each $N_{i,j}$ indicates that case i is hypothesized to contain parasites that are the most direct observed ancestors of the parasites contained in case j . In addition, at least one edge denoted $N_{u,j}$ must exist in the network, indicating that the parasites contained in case j have no ancestors among the parasites contained in any known local case and are instead contained in some unknown case u from some source population s , such that it is denoted u_s . To illustrate this terminology, an example transmission network is depicted in Fig. 1.

To estimate N , the algorithm used spatial, temporal, and travel-history data about all cases, denoted as X_s , X_t , and X_h , respectively. It did so within a Bayesian statistical framework, meaning that it sought to estimate the joint posterior probability density,

$$\Pr(N, \Theta | X_t, \vec{X}_s, X_h) = \frac{\Pr(X_t, \vec{X}_s, X_h | N, \Theta) \Pr(N, \Theta)}{\Pr(X_t, \vec{X}_s, X_h)}, \quad (1)$$

of the transmission network defined by N and the model parameters Θ conditional on the data \vec{X}_s , X_t , and X_h . The first term in the numerator of Eq. (1) is the likelihood of N and Θ conditional on the data. The second term in the numerator is the prior probability of N and Θ . The term in the denominator is the probability of the data, which is an intractable quantity to calculate directly given that it would require evaluation of an extremely high-dimensional integral over N and Θ . To address this, a Markov chain Monte Carlo algorithm was used to draw random samples of N and Θ from the posterior distribution specified in Eq. (1).

The most critical piece of the inference framework is the likelihood, which was defined as a function of each case j as

$$\mathcal{L}(N, \Theta | X_t, \vec{X}_s, X_h) = \prod_j \Pr(X_{t,j}, \vec{X}_{s,j}, X_{h,j} | N_{i,j}, \Theta). \quad (2)$$

Below, the probability of the data associated with each known case j as a function of different assumptions that are possible about how case j is connected to the rest of the transmission network is defined.

Scenario 1: Local transmission between known cases i and j

When case i contains parasites that are immediate ancestors of the parasites contained in case j , its contribution to the likelihood is represented as

$$\begin{aligned} \Pr(X_{t,j}, \vec{X}_{s,j}, X_{h,j} | N_{i,j}, \Theta) &= \Pr(X_{t,j} | N_{i,j}, \Theta) \\ &\quad \Pr(\vec{X}_{s,j} | X_{t,j}, N_{i,j}, \Theta) \\ &\quad \Pr(X_{h,j} | N_{i,j}, \Theta), \end{aligned} \quad (3)$$

which is the product of the probabilities of the temporal, spatial and travel-history data given the network and model parameters. This formulation assumes that those data are generated independently for each individual, with the exception of a dependence of the spatial data on the temporal data.

Probability of the temporal data

To characterize the time elapsed between two cases resulting from local transmission, a model of the generation and serial intervals for *P. falciparum* malaria by Huber et al. [22] was used. The generation interval represents the time between infection of a primary and secondary case, whereas the serial interval represents the time between detection of those cases. Because the timing of infection

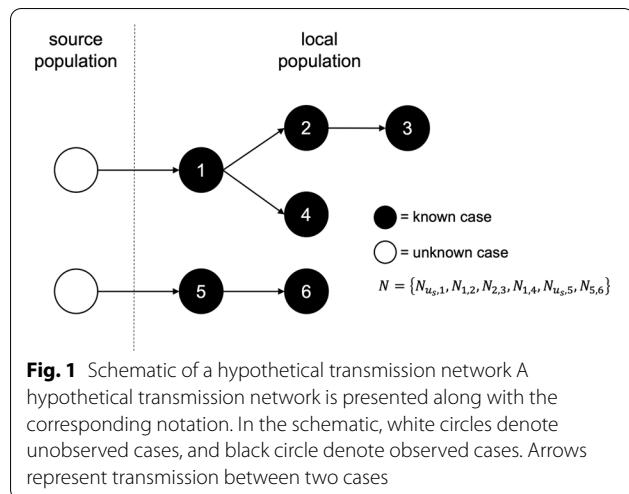


Fig. 1 Schematic of a hypothetical transmission network A
A hypothetical transmission network is presented along with the corresponding notation. In the schematic, white circles denote unobserved cases, and black circles denote observed cases. Arrows represent transmission between two cases

per se (i.e., an infectious mosquito inoculating a susceptible human) is typically unknown, this study focused on the serial interval as the most apropos temporal quantity relating cases.

In deriving the probability of a given length of the serial interval, Huber et al. [22] convolved a discrete random variable representing variability in the generation interval (*GI*) with a discrete random variable representing variability in the time between infection with *P. falciparum* and detection by surveillance, i.e., the infection to detection period (*IDP*). That framework yielded

$$\begin{aligned} \Pr(SI_{i,j} = -a + b + c) &= \sum_a \sum_b \sum_c \\ \Pr(IDP_i = a) \Pr(GI_{i,j} = b) \\ \Pr(IDP_j = c) \mathbb{I} \\ (x = -a + b + c), \end{aligned} \quad (4)$$

as the probability of a serial interval of length $SI_{i,j}$. The algorithm allowed for different models of the serial interval depending upon differences in the *GI* and *IDP* for different types of primary and secondary cases. For instance, the mean *GI* for a primary infection receiving treatment was 48.4 days, compared to 101.6 days for an untreated primary infection. Furthermore, symptomatic cases were assumed to present in a clinic some number of days after infection as informed by empirical data from Zanzibar with a mean of 16.6 days [22]. For an asymptomatic infection, it was assumed that detection occurred through active surveillance at a randomly drawn day among all days where its asexual parasitaemia exceeds a detection threshold, resulting in a mean of 69.8 days [22]. The choice of *IDP* for both the primary and secondary case informs the probability of two cases separated in time by $SI_{i,j} = X_{t,j} - X_{t,i}$ days.

Probability of the spatial data

Following Reiner et al. [6], it was assumed that a simple two-dimensional Wiener diffusion process determines the location of secondary cases relative to the location of their associated primary case. It follows that, for a given diffusion coefficient D with units $\text{km}^2\text{day}^{-1}$ and generation interval $GI_{i,j}$, the two-dimensional location $\vec{X}_{s,j}$ of the secondary case j is described by a bivariate normal distribution with probability density

$$f(\vec{X}_{s,j}|\vec{X}_{s,i}, D, GI_{i,j}, N_{i,j}, \Theta) = \frac{1}{2\pi\sigma^2(GI_{i,j})} e^{-\frac{\|\vec{X}_{s,j} - \vec{X}_{s,i}\|}{2\sigma^2(GI_{i,j})}}, \quad (5)$$

$$f(\vec{X}_{s,j}|D, SI_{i,j}, N_{i,j}, \Theta) = \int f(\vec{X}_{s,j}|\tilde{X}_{s,i}, D, N_{i,j}, \Theta) f(\tilde{X}_{s,i}|\vec{X}_{s,j}, D) d\tilde{X}_{s,i}, \quad (8)$$

where $\sigma^2(GI_{i,j}) = DGI_{i,j}$. This formulation assumes that each spatial dimension is independent, that the variance scales linearly with the generation interval, and that movement is isotropic across a continuous landscape. By making the spatial scale of transmission dependent upon time, the algorithm assumed that a longer generation interval provides a longer period of time over which movement of the primary case could occur. This permitted transmission linkages farther apart in space as the length of the generation interval increased. Because mosquito movement is more restricted and could lead to shorter transmission distances than would be obtained using the two-dimensional Wiener diffusion process, this study evaluated the sensitivity of the inferences to this assumption in the Supplement by using a time-invariant exponential kernel [25].

One complication to Eq. (5) is that the generation interval $GI_{i,j}$ is unobserved and, therefore, cannot take on a fixed value. Instead, data about the serial interval $SI_{i,j}$ must be used to inform the generative model for $\vec{X}_{s,j}$. To do so, the algorithm takes advantage of the property of normal random variables that the sum of two or more random variables is itself a normal random variable [26]. This property allows for the recasting of Eq. (5) as a function of SI rather than GI by computing the appropriate σ^2 as

$$\sigma^2(SI) = \int \sigma^2(GI) \Pr(GI|SI) dGI, \quad (6)$$

which is effectively a weighted sum of the spatial variances associated with a given GI proportional to the probability that the generation interval is exactly GI days

given that the serial interval was observed to be SI days. This results in

$$f(\vec{X}_{s,j}|\vec{X}_{s,i}, D, SI_{i,j}, N_{i,j}, \Theta) = \frac{1}{2\pi\sigma^2(SI_{i,j})} e^{-\frac{\|\vec{X}_{s,j} - \vec{X}_{s,i}\|}{2\sigma^2(SI_{i,j})}}, \quad (7)$$

as the probability density of the spatial data that was assumed.

In the event that case i has missing spatial data, one cannot compute the spatial likelihood of Eq. (7). To address this, a latent unobserved quantity $\tilde{X}_{s,i}$ was defined, which represents the unknown location of case i . The algorithm then integrated over the uncertainty in $\tilde{X}_{s,i}$,

to compute the probability density of case j with known spatial location $\vec{X}_{s,j}$ arising from case i with unknown spatial location $\tilde{X}_{s,i}$. Equation (8) is computed as the product of the probability density of the location of a known case j conditional on an unknown location $\tilde{X}_{s,i}$ and the probability density of spatial separation $\vec{X}_{s,j} - \tilde{X}_{s,i}$ conditional on the diffusion coefficient D for all $X_{s,i}$. Because it was assumed that movement is isotropic, Eq. (8) is a two-dimensional Gaussian integral, simplifying to

$$f(\vec{X}_{s,j}|D, SI_{i,j}, N_{i,j}, \Theta) = \frac{1}{4\pi\sigma^2(SI_{i,j})}. \quad (9)$$

In the event that case j has missing spatial data and case i has known spatial data, the latent unobserved quantity becomes $\tilde{X}_{s,j}$. The algorithm then integrates over the uncertainty in $\tilde{X}_{s,j}$ and calculates $f(\tilde{X}_{s,j}|D, SI_{i,j}, N_{i,j}, \Theta)$ using Eq. (8–9).

Probability of the travel-history data

Although it was assumed in this scenario that a person's infection was locally acquired, the model must still be capable of explaining the travel-history data $X_{h,j}$. Thus, τ_l is the probability that case j reported travel (i.e., $X_{h,j}=1$) even though they were not infected during that period of travel, such that

$$\Pr(X_{h,j}|N_{i,j}, \Theta) = \begin{cases} \tau_l, & X_{h,j} = 1 \\ 1 - \tau_l, & X_{h,j} = 0 \end{cases}. \quad (10)$$

In the event that case j has missing travel-history data, the travel-history likelihood of Eq. (10) cannot be computed. To address this, a latent unobserved quantity $\tilde{X}_{h,j}$,

which represents the unknown travel history of case j , was defined. The algorithm then sums across the uncertainty in $\tilde{X}_{h,j}$,

$$\Pr(X_{h,j} = \text{NA}|N_{i,j}, \Theta) = \Pr(\tilde{X}_{h,j} = 1)\tau_l + (1 - \Pr(\tilde{X}_{h,j} = 1))(1 - \tau_l), \quad (11)$$

to compute the probability that case j was locally acquired given an unknown travel history. In Eq. (11), $\Pr(\tilde{X}_{h,j} = 1)$ was computed as the proportion of cases with a positive travel history among all cases with known travel-history data.

Taken together with the probabilities of the temporal and spatial data described above, the product of these three probabilities constitutes the entirety of the contribution of a case j infected by a known local case i to the overall likelihood of N and Θ .

Scenario 2: Importation of local case j from source population s

In the event of $N_{u_s,j}$, the contribution of such a case to the overall likelihood of N and Θ is represented as the product of the probabilities of its temporal, spatial and travel-history data under similar assumptions as in Scenario 1. The key difference in this scenario is that there is no information about the unknown source case that gave rise to case j .

Probability of the temporal data

Because the person containing parasites that are the direct ancestors of those in case j is unobserved and does not have an $X_{t,i}$, the probability of the temporal data as described in Scenario 1 cannot be computed. It is important though to obtain a probability comparable to that from Scenario 1 as a reference point for determining whether it is more likely that a given case arose from some other known local case or from an unknown case u_s from source population s . To do so, the algorithm considers the variable \tilde{X}_{t,u_s} , which is a latent variable describing the timing of when u_s would have been detected, had it been detected.

Because u_s is not observed, it was considered to be asymptomatic and untreated. The algorithm then calculated the probability of the timing of a known case j arising from an unknown case u_s as

$$\Pr(X_{t,j}|N_{u_s,j}, \Theta) = \int \Pr(X_{t,j}|\tilde{X}_{t,u_s}, N_{u_s,j}, \Theta) \Pr(\text{SI} = X_{t,j} - \tilde{X}_{t,u_s}) d\tilde{X}_{t,u_s}, \quad (12)$$

by integrating over uncertainty in \tilde{X}_{t,u_s} . This is represented as the product of the probability of the timing of a known case j conditional on an unknown time of detection \tilde{X}_{t,u_s} and the probability of the serial interval $X_{t,j} - \tilde{X}_{t,u_s}$ for all \tilde{X}_{t,u_s} . Equation (12) does not distinguish between symptomatic and asymptomatic cases j because the calculation is identical; only the serial interval distributions differ.

Probability of the spatial data

Without an \tilde{X}_{t,u_s} for the unobserved case u_s , the algorithm lacked information on the serial interval between it and case j . Consequently, the probability from Eq. (7) could not be used in that particular form. Instead, the spatial variance was computed as a function of the diffusion coefficient alone, yielding

$$\sigma^2(D) = \int DGI \Pr(GI) dGI. \quad (13)$$

Equation (13) integrates across all possible generation intervals and simplifies to $D\mathbb{E}[GI]$, the product of the diffusion coefficient and the expectation of the generation interval.

This spatial variance was applied to the unobserved latent variable \tilde{X}_{s,u_s} , which represents the unknown location of the unobserved case u_s . The algorithm integrated over uncertainty in \tilde{X}_{s,u_s} to compute the probability density,

$$f(X_{s,j}|D, N_{u_s,j}, \Theta) = \int f(X_{s,j}|\tilde{X}_{s,u_s}, D, N_{u_s,j}, \Theta) f(\tilde{X}_{s,u_s}|X_{s,j}, D) d\tilde{X}_{s,u_s}, \quad (14)$$

of the location of a known case j arising from an unknown source case u_s with unknown location \tilde{X}_{s,u_s} . This is represented as the product of the probability density of the location of a known case j conditional on an unknown location \tilde{X}_{s,u_s} and the probability density of spatial separation $X_{s,j} - \tilde{X}_{s,u_s}$ conditional on the diffusion coefficient D for all \tilde{X}_{s,u_s} . As in Eq. (9), Eq. (14) was treated as an evaluation of the Gaussian integral, evaluating to

$$f(X_{s,j}|D, N_{u_s,j}, \Theta) = \frac{1}{4\pi D\mathbb{E}[GI]}. \quad (15)$$

In Eq. (15), D is the diffusion coefficient and $\mathbb{E}[GI]$ is the expectation of the generation interval.

Probability of the travel-history data

The travel history $X_{h,j}$ was considered to be a binary variable with a value of 1 indicating a presumed malaria importation due to reported international travel to an area with known malaria transmission within the past eight weeks but excluding the minimum incubation period of one week prior to the data of presentation. After defining the probability τ_s that $X_{h,j} = 1$ conditional on $N_{u_s,j}$, it follows that

$$\Pr(X_{h,j}|N_{u_s,j}, \Theta) = \begin{cases} \tau_s, X_{h,j} = 1 \\ 1 - \tau_s, X_{h,j} = 0 \end{cases}, \quad (16)$$

which constitutes the contribution of the travel history of such a case to the overall likelihood of N and Θ . If the travel history of case j is unknown, an analogous calculation to Eq. (11) is made using τ_s .

Bayesian inference

Markov Chain Monte Carlo algorithm

To avoid evaluating the high-dimensional integral over N and Θ , samples of N and Θ were drawn from their posterior distribution defined by Eq. (1) using a Metropolis-Hastings Markov chain Monte Carlo (MCMC) method [27, 28]. To begin the chain, N and Θ were initialized to $N^{(1)}$ and $\Theta^{(1)}$, and each subsequent step i in the chain was denoted $N^{(i)}$ and $\Theta^{(i)}$. At each step, states N' and Θ' were proposed with $\Pr((N^{(i)}, \Theta^{(i)}) \rightarrow (N', \Theta'))$. Proposed states were accepted with probability

$$\alpha_{\text{update}} = \min \left[1, \frac{\pi(N', \Theta') \Pr((N', \Theta') \rightarrow (N^{(i)}, \Theta^{(i)}))}{\pi(N^{(i)}, \Theta^{(i)}) \Pr((N^{(i)}, \Theta^{(i)}) \rightarrow (N', \Theta'))} \right], \quad (17)$$

where $\pi(N, \Theta)$ is the product of the likelihood $\Pr(\vec{X}_s, X_t, X_h|N, \Theta)$ of N and Θ conditional on the data and the assumed prior probability $\Pr(N, \Theta)$ of N and Θ . After a random draw R from a uniform distribution, the chain was updated according to

$$N^{(i+1)}, \Theta^{(i+1)} = \begin{cases} N', \Theta', & R \leq \alpha \\ N^{(i)}, \Theta^{(i)}, & R > \alpha \end{cases}. \quad (18)$$

To reduce the probability of the chain becoming stuck at a local maximum, this study employed Metropolis-coupled Markov chain Monte Carlo (MC³) [29]. Implementing MC³ involved running multiple chains in parallel, with $\pi_c(N, \Theta)$ in chain c raised to the power β_c according to

$$\beta_c = 1 + \lambda(c - 1), \quad (19)$$

where $\lambda > 0$ is a temperature increment parameter that governs the degree to which each chain is ‘heated’. As a result of setting $\beta_1 = 1$, $\pi_1(N, \Theta)$ is directly proportional

to the joint posterior distribution and is referred to as the master or ‘cold’ chain. This algorithm effectively flattens the likelihood in the heated chains by setting $\beta_c > 1$, allowing them to explore the parameter space more freely and to encounter alternative high-density regions more readily than the cold chain would alone. At a pre-defined frequency, two randomly selected chains i and j were allowed to swap parameter sets according to a swap probability

$$\alpha_{\text{swap}} = \min \left[1, \frac{\pi(N^{(j)}, \Theta^{(j)})^{\beta_i} \pi(N^{(i)}, \Theta^{(i)})^{\beta_j}}{\pi(N^{(i)}, \Theta^{(i)})^{\beta_i} \pi(N^{(j)}, \Theta^{(j)})^{\beta_j}} \right], \quad (20)$$

where $\pi(N, \Theta)$ is the same as it was in Eq. (17). A swap into the master chain only occurred if it was from one of the two randomly selected chains and $R \leq \alpha_{\text{swap}}$. This analysis recorded a total of 100 million samples from the posterior distribution, discarding the first 50 million samples as burn-in and thinning the chain every 10,000 samples between each recorded sample.

Proposals

Proposals made by the MC³ algorithm involved changes to the parameters (i.e., D , τ_s , and τ_l) and changes to the transmission network topology. Each proposal occurred with a fixed probability, where the sum of these proposal probabilities was equal to one.

Proposals to change parameters involved updating D , τ_s , or τ_l . To update the value of D , a new value was drawn from a normal distribution with mean set to the current value of the parameter and variance set to 2.5. Values of D proposed must be strictly non-negative, so any proposed D that was less than zero was rejected and assigned $\alpha_{\text{update}} = 0$. Similarly, new values of τ_s and τ_l were chosen according to normal distributions with means set to their current parameter value and variance set to 0.25. Because τ_s and τ_l are probabilities, any proposed value that fell outside the range [0,1] was rejected and assigned $\alpha_{\text{update}} = 0$.

Changes proposed to the network topology involved the addition or removal of an ancestor from a randomly selected node. The algorithm assigned a uniform probability of proposing case a as an ancestor to a randomly selected case i , such that proposals to the network topology are uninformed by spatial and temporal data. Each proposed ancestor was chosen from the set of ancestors that would ensure that the network remained acyclic. Furthermore, the proposal probability of removing case a as an ancestor to a randomly selected case i was defined as

$$\Pr(\text{remove}a) = \bar{A}_i^{-1}, \quad (21)$$

where \bar{A}_i represents the size of set A_i of all ancestors to case i . Proposed changes to the network are then accepted according to Eq. (17).

Prior assumptions

Strong priors were placed on τ_s and τ_p , because it was assumed that travel histories were mostly, but not completely, accurate. The algorithm used a beta-distributed prior on τ_s , with parameters $\alpha_{\tau_s} = 12$ and $\beta_{\tau_s} = 3$, which resulted in a mean of 0.8 and a variance of 0.01 for this prior distribution. The algorithm also used a beta distributed prior on τ_p , with parameters $\alpha_{\tau_p} = 3$ and $\beta_{\tau_p} = 12$, which resulted in a mean of 0.2 and a variance of 0.01. A uniform prior on D over the interval $[10^{-3}, \infty)$ and an even prior across all possible network configurations were assumed, meaning that those prior probabilities cancelled out in eqs. (17) and (20).

Assessing convergence

For D , τ_s , and τ_p , convergence was assessed using the Gelman-Rubin statistic [30], with values below 1.1 indicating convergence. For the transmission network N , convergence was assessed by calculating correlation coefficients of case-level probabilities across five chains from independent realizations of the MC³ algorithm, for a total of 10 pair-wise comparisons across the five chains. The two case-level probabilities considered were the posterior probability that each case was infected by an unknown case u_s from a source population and the posterior probability that each case j was infected by each other case i . Higher values of these correlation coefficients provided stronger support for convergence.

Results

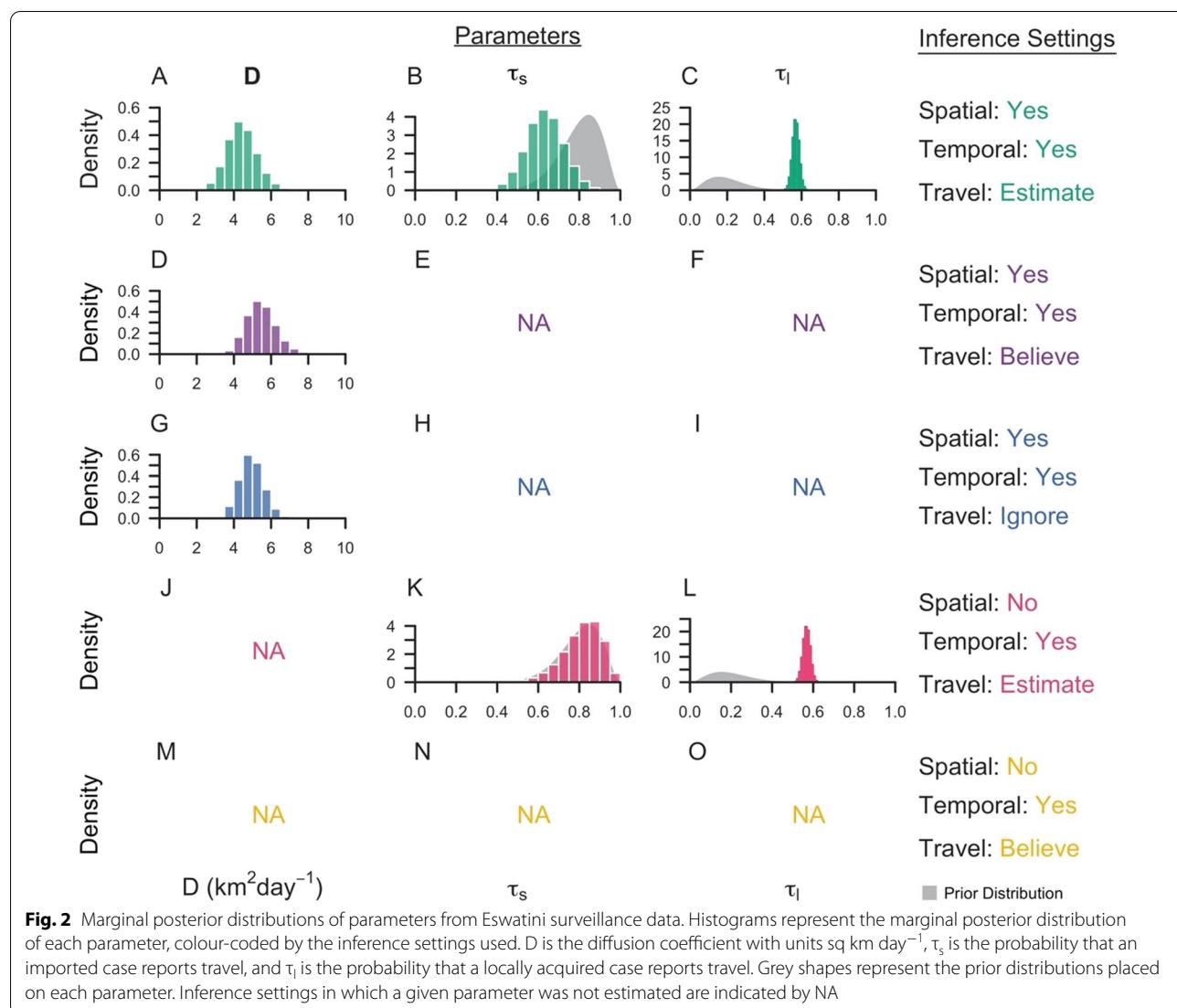
To establish proof-of-concept, this study first applied the inference method on three simple test cases and evaluated how well the inferences recovered the true transmission networks. Then, the method was applied to surveillance data collected in Eswatini during 2013–2017. The focus was less on understanding malaria epidemiology in Eswatini and more on understanding how epidemiological conclusions change with the inclusion or exclusion of different data types and different assumptions about travel histories. These inference settings used: (1) spatial and temporal data while estimating the accuracy of the travel history (default setting); (2) spatial and temporal data while believing the travel history; (3) spatial and temporal data alone; (4) temporal data while estimating the accuracy of the travel history; and, (5) temporal data while believing the travel history. To validate the inferences based on data from Eswatini,

simulated data was generated using posterior parameter estimates obtained from the data from Eswatini and evaluated the ability of our inference method to recover the true transmission networks along with the underlying parameters on those simulated data. Finally, a simulation sweep across different epidemiological settings was performed to determine the range of conditions under which our inference method yielded reliable estimates of transmission. A full description of the analyses and additional results can be found in the Supplement.

Application to Eswatini surveillance data

The method was applied to surveillance data collected in Eswatini during 2013–2017. Under the default inference setting, the median posterior diffusion coefficient D , which quantifies the spatial spread of transmission, was estimated to be $4.40 \text{ sq km day}^{-1}$ (95% Credible Interval: $2.93\text{--}6.13 \text{ sq km day}^{-1}$) (Fig. 2A). This corresponded to a median inferred transmission distance of 13.0 km ($0.0130\text{--}64.8 \text{ km}$), a median inferred serial interval of 45 days ($-37\text{--}148$ days) (Fig. 3A, B), and median estimates of τ_s , the probability that an imported case reported travel, of 0.63 ($0.46\text{--}0.81$) compared to the prior distribution mean of 0.80 and τ_p , the probability that a locally acquired case reported travel, of 0.57 ($0.53\text{--}0.61$) compared to the prior distribution mean of 0.20 (Fig. 2B, C). That the 95% credible interval for τ_s contained 0.50 indicated that the inference algorithm found limited use of travel-history data in discriminating between imported and locally acquired cases, because that implies that imported cases have equal probabilities of reporting or not reporting travel. The algorithm estimated the proportion of imported cases to be 0.046 , corresponding to $R_c=0.95$. Mapping risk of importation and local transmission across Eswatini under the default inference setting, the algorithm estimated consistently low risk of importation throughout the country and transmission hotspots in the northeastern part of Eswatini, close to the border with Mozambique (Fig. 4A, B).

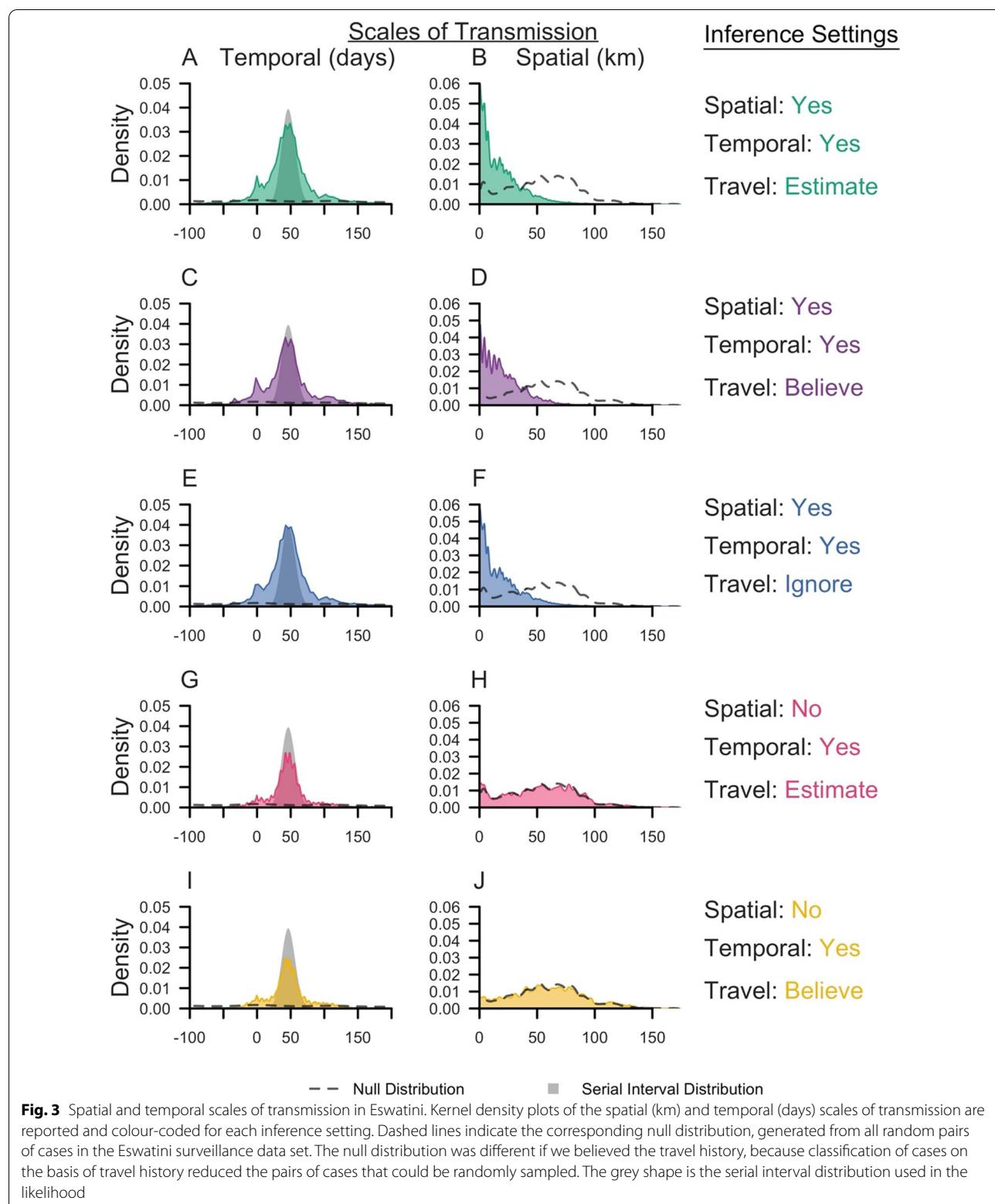
Parameter estimates and transmission network inferences differed under other inference settings. When the travel history was believed, a larger median transmission distance was estimated (Fig. 3D). This increase in the spatial scale of transmission can be attributed to clusters of cases with positive travel histories located near metropolitan areas. By forcing those cases to be imported, the algorithm tended to infer transmission across longer distances to explain the origins of the remainder of cases that did not report travel and were thereby inferred to be locally acquired. With respect to time, all five inference settings produced consistent serial interval estimates, although the inclusion of spatial data allowed for a wider range of transmission linkages in time (Fig. 3A, C,

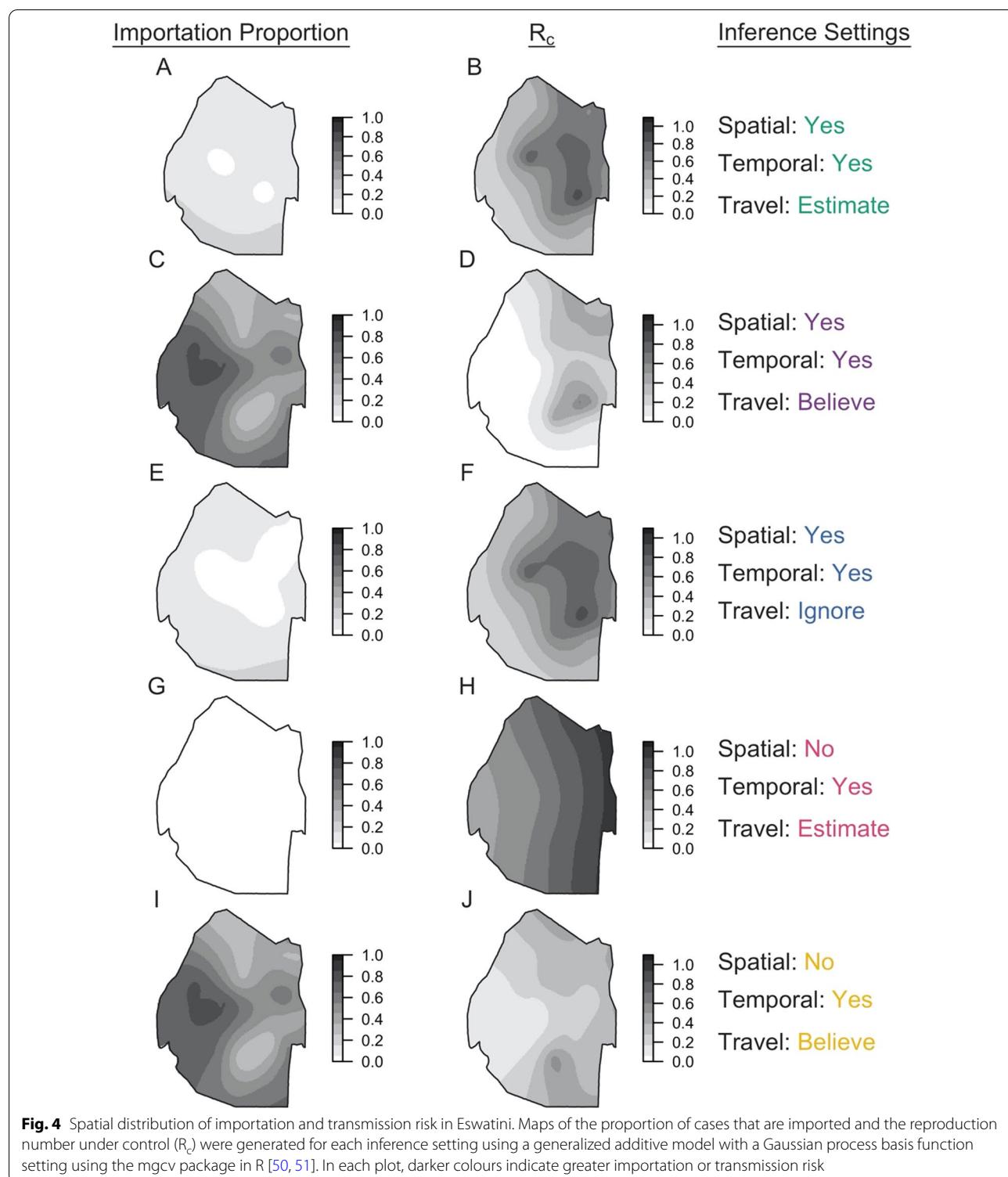


E). Finally, in the absence of spatial data, the model estimated higher predictive power of travel histories in identifying imported cases (τ_s : 0.83, [0.60, 0.95]), though the travel history was consistently found to be uninformative for identifying locally acquired cases (τ_l : 0.57, [0.53, 0.60]) (Fig. 2K, L).

Classification of cases as imported or locally acquired, key information for control programmes, was sensitive to the choice of inference setting. The proportion of cases classified as imported was most sensitive to different assumptions about the accuracy of the travel histories (Fig. 4, left column; Fig. 5). Believing the travel history yielded high estimates of importation in western Eswatini (Fig. 4C, I), whereas estimating or ignoring the travel history yielded low, relatively homogeneous estimates of importation risk (Fig. 4A, E, G).

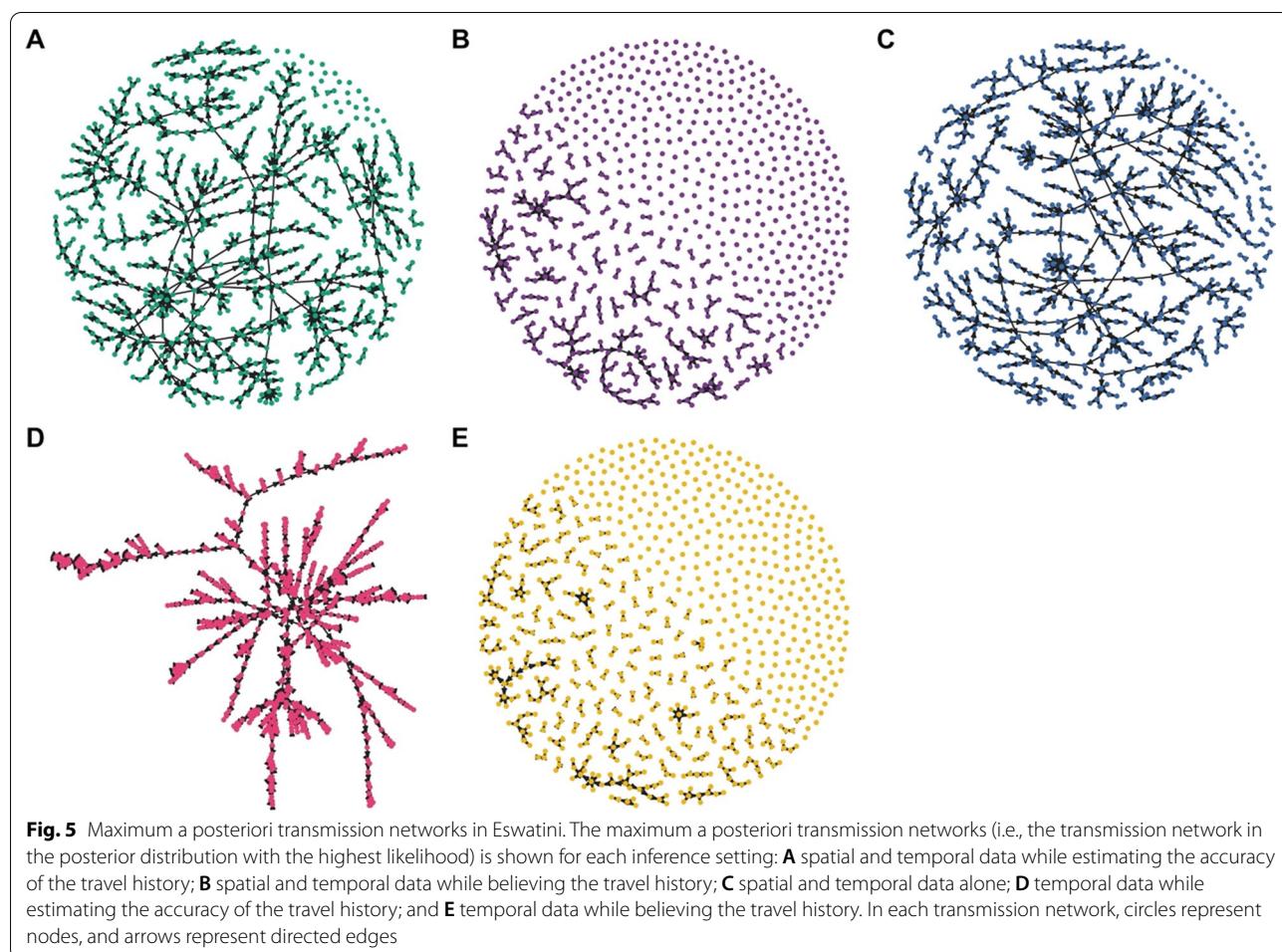
For instance, using temporal data and estimating the accuracy of the travel history produced probabilities of importation that ranged 0.0045–0.0053, suggesting that nearly all cases resulted from local transmission (Figs. 4G, 5D). Estimates of the spatial distribution of R_c depended most on the choice of which data types were included (Fig. 4, right column). Notably, inclusion of spatial and temporal data produced a consistent spatial distribution of relative transmission risk, with transmission hotspots in northeastern Eswatini (Fig. 4B, D, F). However, believing the travel history reduced the magnitude of transmission that was inferred from a median R_c of 0.95 (Figs. 4B, 5A) under default settings to 0.41 (Figs. 4D, 5B). Omitting spatial data changed the spatial distribution of transmission. Estimating the accuracy of the travel history yielded high transmission estimates





(median R_c : 1.00) in eastern Eswatini (Fig. 4H), whereas believing the travel history inferred hotspots of transmission (median R_c : 0.42) in southern Eswatini (Fig. 4J). Believing the travel history led to slightly different

median estimates of R_c (0.41 vs 0.42) depending upon whether spatial data were included, because the travel histories were unknown for 36 cases included in the analysis. As part of the inference procedure, the



algorithm classified these cases as imported or locally acquired, and including spatial data caused a greater number of cases to be inferred to be imported.

Validation of inferences from Eswatini

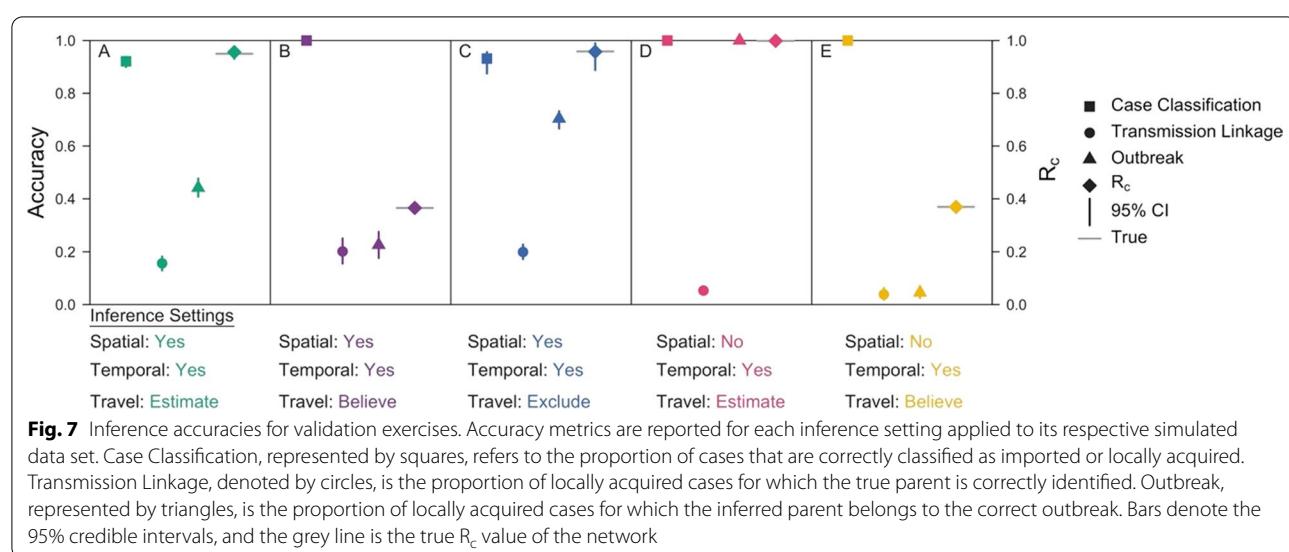
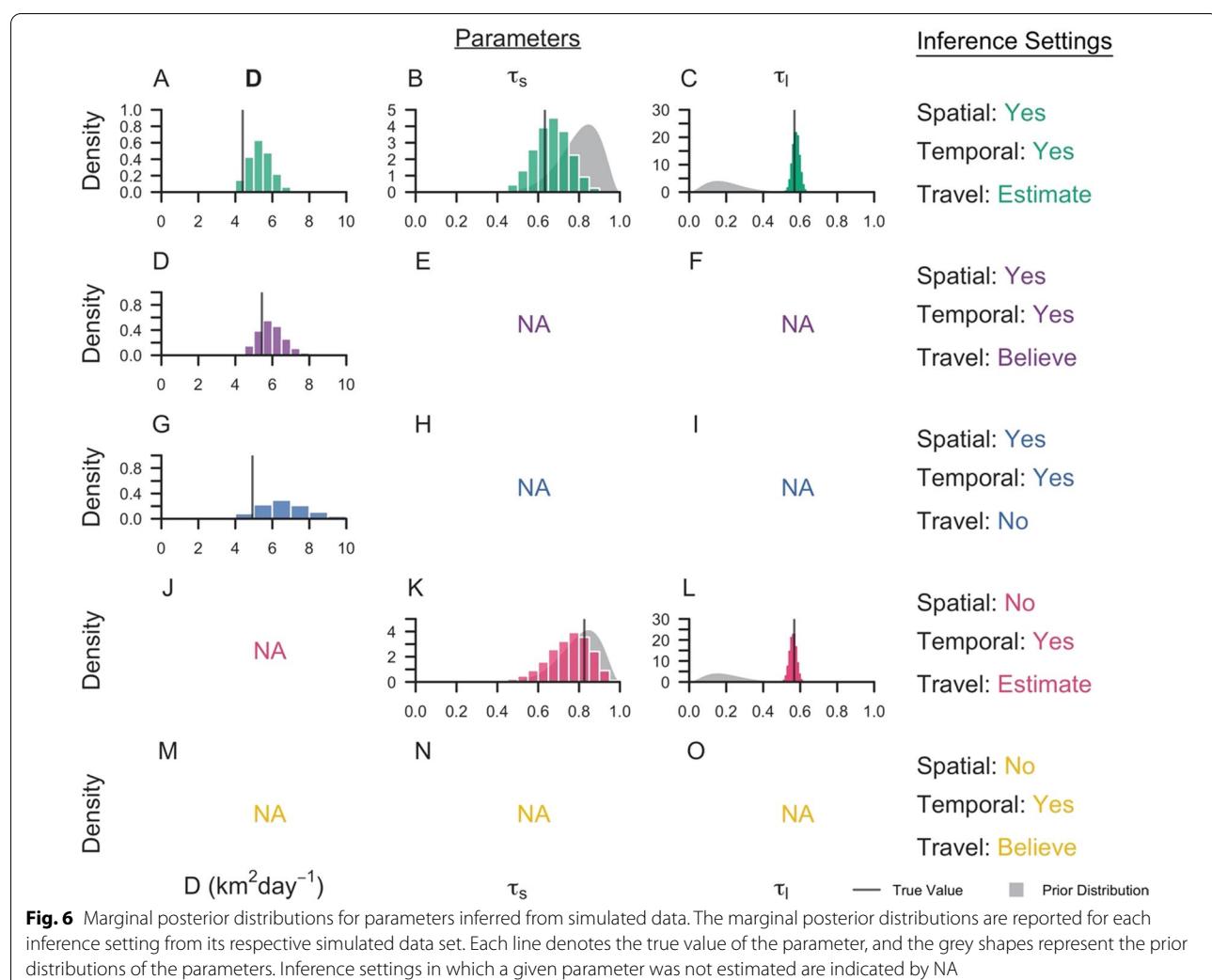
Reconciling the different inferences under different inference settings in Figs. 2, 3, 4 and 5 was challenging

because the true, underlying network and parameters were unknown. Using median posterior estimates from the Eswatini data under each inference setting, data was simulated to assess the ability of the inference method to recover the underlying parameters and transmission networks (Table 1). It was observed that the model was able to estimate the diffusion coefficient D , τ_s , and

Table 1 Characteristics of simulated data generated using the branching process model

Inference setting			Network Size	Number of outbreaks	Prop. imported	D	τ_s	τ_l
Space	Time	Travel						
Yes	Yes	Estimate	775	43	0.046	4.40	0.63	0.57
Yes	Yes	Believe	775	492	0.59	5.44	1	0
Yes	Yes	No	775	36	0.039	4.93	NA	NA
No	Yes	Estimate	775	1	0.0013	NA	0.83	0.57
No	Yes	Believe	775	489	0.58	NA	1	0

A description of the simulated data used in the inference exercises are reported for each of the five inference settings. The total number of nodes in the network, the number of distinct outbreaks, the proportion of cases that are imported, and the underlying parameters are provided



τ_l reasonably well, depending on the inference setting (Fig. 6).

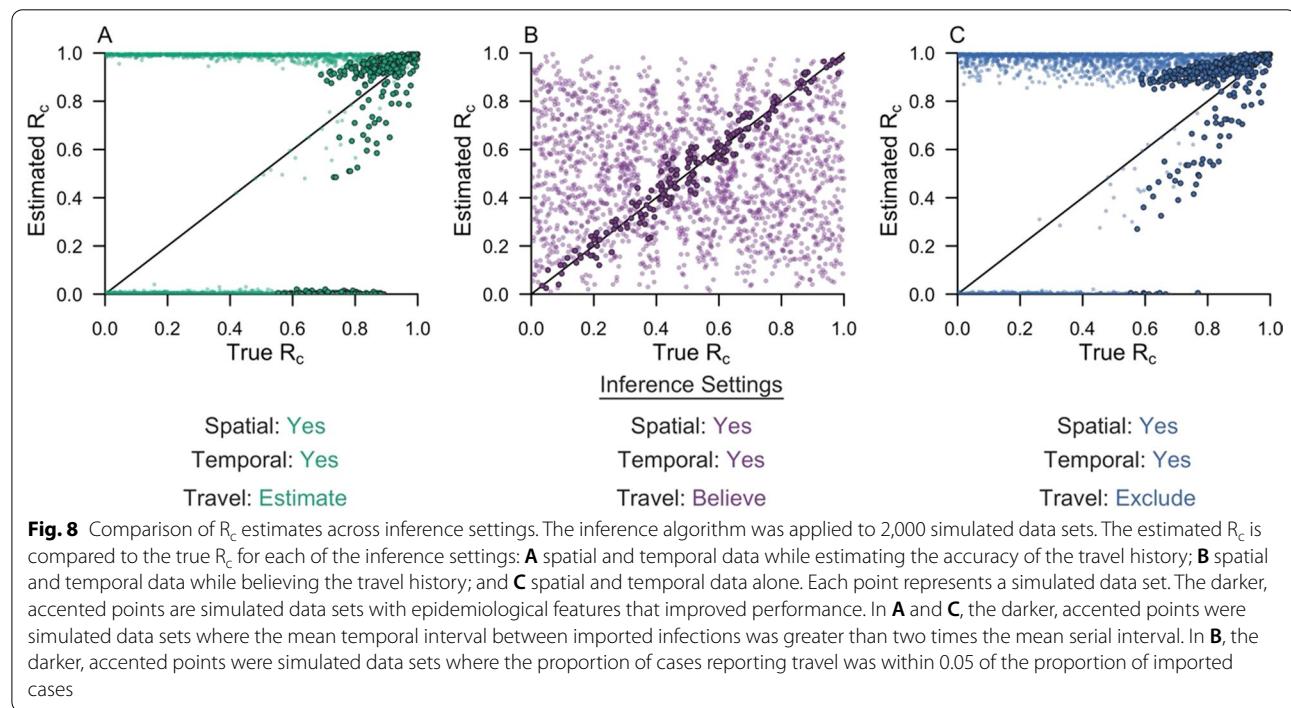
The overall accuracy of classifying cases as imported or locally acquired was close to one (Fig. 7). Though seemingly promising, these high accuracies masked a tendency to overclassify cases as locally acquired, because many more cases were simulated to be locally acquired than imported. For example, under the default inference setting, the accuracy of correctly classifying imported cases was 0.15 (0.051–0.26). Similarly, the accuracies of identifying the parent of each transmission linkage were poor, despite simulating under the assumptions of the model, with accuracies ranging from 0.038 (0.017–0.063) when using temporal data and believing the travel history to 0.20 (0.16–0.25) when incorporating spatial and temporal data and believing the travel history (Fig. 7, circle points). This suggests that, as the number of cases increases within a fixed space–time window, the information content of routinely collected epidemiological data for inferring transmission chains decreases and the method becomes incapable of correctly estimating the transmission network. Nevertheless, under some settings, the method was able to capture higher-order summaries of the network, such as case classification and R_c (Fig. 7, square and diamond points).

Simulation sweep

Validation of the inference algorithm revealed that its performance varied across simulated data sets. When

applied to a series of simple test cases in which the transmission networks were small and in an optimal spatiotemporal arrangement, the inference method was able to reconstruct the transmission network and correctly estimate R_c (Additional file 1: Fig. S2). When applied to larger transmission networks in which outbreaks overlapped in space and time, performance of the inference method was poor (Fig. 7). This indicated that the performance of the inference algorithm depends on the epidemiological setting to which it is applied. To address this observation, 2,000 simulated data sets were generated in which the proportion of imported cases, the spatiotemporal window over which imported cases were distributed, the diffusion coefficient, and the accuracies of the travel history (i.e., τ_s and τ_l) were varied (Additional file 1: Table S2). Then, the inference algorithm was applied under three different inference settings, and the accuracy of reconstructing each transmission network was quantified. The three inference settings used: (1) spatial and temporal data while estimating the accuracy of the travel history (default setting); (2) spatial and temporal data while believing the travel history; and, (3) spatial and temporal data alone (Additional file 1: Table S1).

The accuracy of reconstructing transmission networks depended upon both the inference setting used and the epidemiological features of the simulated data. When the algorithm used spatial and temporal data and estimated the accuracy of the travel history or excluded it, the accuracy of reconstructing transmission networks



depended on the relative proportion and temporal distribution of imported cases (Additional file 1: Fig. S9 and S11). As the temporal window over which imported cases are distributed increased, the accuracy of identifying the true parent and the true outbreak of each locally acquired case increased. With an increasing temporal window, outbreaks within the transmission network became relatively more focal in time, which made the likelihoods of alternative transmission linkages more readily distinguishable. More accurate estimates of R_c under these inference settings similarly depended on the temporal window over which imported cases were distributed (Fig. 8A, C). When the mean temporal interval between imported infections was greater than two times the mean length of the serial interval (i.e., approximately 100 days), the estimates of R_c improved, although the algorithm generally overestimated it. The estimates of τ_s and τ_l also improved under these epidemiological settings (Additional file 1: Fig. S12), providing further support that the inference method can reasonably infer transmission networks under select settings. Furthermore, as the proportion of imported cases increased and R_c decreased, the accuracy of identifying the correct outbreak of each locally acquired case decreased (Additional file 1: Fig. S9 and S11). This pattern reflected the relationship between R_c and the size of individual outbreaks. As R_c decreased, the size of individual outbreaks decreased, and consequently, the probability that the inferred parent of a locally acquired case belonged to the same outbreak decreased (Table 2).

By contrast, when the travel history was believed, the accuracy of reconstructing transmission networks depended most strongly on the accuracies of the travel history. As the probability of reporting travel increased, the accuracy of classifying imported cases increased, and the accuracy of classifying locally acquired cases decreased (Additional file 1: Fig. S10). Under this inference setting, the estimate of R_c depended only on the proportion of cases that reported travel. When the proportion of cases that reported travel matched the proportion of cases that were imported, R_c was correctly estimated (Fig. 8B).

Table 2 Definitions of estimated parameters

Parameter	Definition
D	Diffusion coefficient (sq km day^{-1})
τ_s	Probability that an imported case reports travel
τ_l	Probability that a locally acquired case reports travel

Discussion

The results show that, in many settings, analyses based on routinely collected surveillance data may not be capable of reconstructing individual-level transmission networks of falciparum malaria and inform estimates of the reproduction number under control, R_c . Using simulated data similar to the Eswatini surveillance data that were analysed, the inference algorithm correctly identified transmission linkages less than 25% of the time. This inaccuracy can be primarily attributed to the inherently limited information content of spatiotemporal data on *P. falciparum* for this purpose. Its characteristically long serial interval [22] means that an appreciable number of cases presenting within a short timeframe are difficult to link to each other based on their timing, even in a relatively facile test case in which the generative process assumed in the likelihood function matched that used to simulate the data. The inability to reconstruct transmission networks using routine surveillance data has been observed for other inference algorithms when applied to pathogens, such as *Mycobacterium tuberculosis* and *Klebsiella pneumoniae*, with similarly long serial intervals, providing further evidence that the limitations noted in this study may be generally inherent to the epidemiological data, rather than the method per se [21].

Under most simulated scenarios and assumptions about the accuracy of travel-history data, the algorithm overestimated the number of locally acquired cases, leading to overestimates of R_c . Crucially, the simulation sweep demonstrated that routinely collected surveillance data was most informative of individual-level transmission networks and R_c when local outbreaks were highly focal in time. Otherwise, while the algorithm was able to reconstruct the true transmission network with modest accuracy, it tended to misclassify truly imported cases as locally acquired, thereby overestimating R_c . Taken together, these results suggest that analyses may need to leverage additional data types beyond routinely collected surveillance data to infer transmission chains and inform fine-scale estimates of *P. falciparum* transmission in many near-elimination settings. For other purposes and at broader spatial scales, however, routinely collected surveillance data still have practical value, because the spatial distribution of cases can reveal epidemiological risk factors relevant for targeted interventions [31, 32].

Although this study was able to reach some general conclusions about the inference algorithm, the inferences were highly sensitive to which data types were included and which assumptions were made about the accuracy of travel-history data. Applying the algorithm to surveillance data from Eswatini, it was observed that inferred patterns of transmission depended on which

data types were included. With the inclusion of spatial data, the inferences captured a spatial pattern of transmission consistent with another analysis from Eswatini [33] with data from a different time period. Assumptions about the travel history appeared to have a strong influence on the overall magnitude of transmission that was inferred, due to the direct relationship between R_c and the proportion of imported cases [16]. As a result, believing the travel history, and thereby treating it as perfectly accurate as in previous approaches [6, 18–20], could bias R_c estimates if there are errors in travel-history data. A study comparing community travel surveys to mobile-phone data in Kenya found that travel histories considerably underestimated the volume of travel, suggesting high rates of false negatives in community travel surveys [34]. Believing the travel history may underestimate the number of imported cases and overestimate R_c . Accounting for inaccuracy in travel-history data is therefore important, and studies pairing community travel surveys with mobile-phone data could be used to inform prior distributions on the likely accuracy of travel-history data [34, 35].

The method that was used only considered a single spatial model to infer transmission linkages and assumed complete observation of cases, both of which are factors that could have affected our inferences based on the Eswatini surveillance data. The diffusion model that was used to represent spatial dispersion of parasites assumed that movement is isotropic in space and did not consider landscape features, such as heterogeneity in human population densities and environmental factors that may affect mosquito ecology. A study analysing self-reported movement patterns in Mali, Burkina Faso, Zambia, and Tanzania found that gravity and radiation models of spatial dispersion fit the data well, although the appropriateness of each model depended on the type of traveller, the travel distance, and the population size of the destination considered [36]. Although a variety of spatial kernels could have been used in the analysis, the conclusions reached are expected to be robust to the choice of spatial kernel, because the spatial kernel used in the likelihood matched that used to simulate the data. Regarding the representation of *P. falciparum* infections in the data set from Eswatini, there are asymptomatic and mild infections that are unlikely to have been recorded in the surveillance system yet may comprise a substantial proportion of malaria infections within Eswatini [13]. Accordingly, it is possible that the assumption of complete observation of cases could have biased R_c estimates, likely downward due to the fact that missing cases will tend to make offspring numbers appear smaller than they actually are [37, 38]. Even so, the conclusions about

the sensitivity of transmission network inferences to the choice of data types and assumptions about travel-history data are expected to be robust to these limitations of the study. This further reinforces the conclusion of the need for caution in attempting to reconstruct person-to-person transmission networks from routine surveillance data [39], because incomplete observation of cases would lead to greater inaccuracies in our transmission network inferences beyond what was noted in the study.

Given that some of the limitations of this approach may be inherent to the information content of these data types in this system, one potential avenue for improving inferences of fine-scale patterns of *P. falciparum* transmission could involve the integration of additional data streams. For example, mobile-phone data [40], high-resolution friction surfaces [41], and other anisotropic surfaces, such as transport networks, could more realistically characterize mobility patterns and allow quantification of the effects of spatial model misspecification, whereas travel-history information that details the dates, duration and location of each trip that has been used in programmatic contexts [31] could more accurately identify importation events. Additionally, the inclusion of pathogen genetic data, which has the potential to provide a more direct signal of parasite movement, could complement traditional epidemiological data [42]. Diverse genetic markers of *P. falciparum* have been characterized in near-elimination settings, such as Eswatini [43], and have been successfully used to identify imported cases in Bangladesh [44] and Namibia [35]. There is also scope for further methodological development, such as relaxing the assumption of complete observation of infections and incorporating an underlying mechanistic model of transmission (as in Lau et al. [8]; Guzzetta et al. [45]). Incorporating an underlying mechanistic model would relax the uninformative prior assumption on all possible transmission networks, ruling out transmission networks that are epidemiologically implausible and accounting for spatial differences in transmission potential and the rate of importation due to different epidemiological and demographic factors. This approach would also permit an estimate of the serial interval distribution and seasonal variation therein directly from the data rather than borrow estimates from the literature [22, 46, 47]. To this end, this study envisions that leveraging the strengths of this method along with other, complementary methods could strengthen inferences based on routinely collected epidemiological data and open up new possibilities to make use of even more data types, such as serological data, prevalence surveys and pathogen genetic data [42, 48, 49].

Conclusions

This study revealed limitations of analyses of routinely collected surveillance data for the inference of individual-level transmission networks of *P. falciparum*. It identified a tendency to overestimate local transmission using routinely collected surveillance data, especially when outbreaks overlapped in space and time. Using both real data from Eswatini and simulated data, this analysis identified strong sensitivities of the inferences to the epidemiological setting, the choice of data types included, and assumptions about the accuracy of travel-history data. The results indicated that using spatial and temporal data and believing travel histories yielded the most plausible estimates of transmission when applied to the Eswatini surveillance data. However, the simulation sweep demonstrated that the accuracy of the inferences strongly depended on the accuracy of the travel-history data when the travel-history data were assumed to be accurate. These sensitivities to the choice of data types and assumptions about the accuracy of travel-history data could have important programmatic implications if outputs of transmission network inferences are operationalized. Although this study was specific to *P. falciparum*, the results of the analyses indicate that future studies inferring transmission networks of *P. falciparum*, or any pathogen, should carefully consider the epidemiological setting and the choice of data types and assumptions that inform the model and should validate them using simulated data.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12936-022-04072-2>.

Additional file 1. Additional figures and tables.

Acknowledgements

The authors thank the National Malaria Elimination Programme as well as Non-tokozo Mgadi and Deepa Pindolia from the Clinton Health Access Initiative for their support in the collection of the surveillance data used in this study. We also thank Brooke Whittemore for her support in data management.

Authors' contributions

JHH, TAP and BG conceived of the study, MSH, ND, SV, NN, NN, and BG curated the data. JHH, MSH, MM, AL, BG, and TAP performed the formal analysis. Funding was acquired by BG and MSH. Investigation was by JHH, BG and TAP. Methodology was developed by JHH, MM, AL, RN, BG, and TAP. Project administration was by JHH and TAP. JHH worked with the software. TAP and BG supervised the project. JHH, BG and TAP wrote the original draft of the manuscript. All authors read and approved the final manuscript.

Funding

JHH acknowledges support from a National Science Foundation Graduate Research Fellowship and a Richard and Peggy Notebaert Premier Fellowship. BG and TAP received support from a grant from the Bill and Melinda Gates Foundation (OPP 1132226 to BG). MSH received support from NIAID (AI101012). The funders had no role in study design, analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The code and simulated data to reproduce the analyses can be found at https://github.com/johnhuber/SpaceTime_Networks. The data collected from Eswatini contains sensitive household locations and are unable to be shared due to institutional review board restrictions.

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from the Eswatini Ministry of Health, the University of California, San Francisco, and the University of Notre Dame (IRB 19-06-5408). All analyses were performed in accordance with relevant guidelines and regulations. Written informed consent was obtained from participants or a parent or guardian for children less than 18 years of age. From September 2015 until the end of the study, written assent was additionally obtained for children aged 12–17 years. All data were analyzed anonymously.

Consent for publication

Written informed consent was obtained from participants or a parent or guardian for children under 18 years of age. From September 2015 until the end of the study, written assent was additionally obtained for children aged 12–17 years. All data were analysed anonymously.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA. ²Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, USA. ³Malaria Elimination Initiative, Global Health Group, University of California, San Francisco, CA, USA. ⁴Department of Pediatrics, University of California, San Francisco, CA, USA. ⁵National Malaria Elimination Programme, Ministry of Health, Manzini, Eswatini. ⁶Department of Medicine, University of California, San Francisco, CA, USA. ⁷Department of Integrative Biology and Statistics, University of California, Berkeley, CA, USA. ⁸Clinton Health Access Initiative, Eswatini Country Office, Mbabane, Eswatini. ⁹Chan Zuckerberg Biohub, San Francisco, CA, USA.

Received: 17 July 2021 Accepted: 31 January 2022

Published online: 21 February 2022

References

- Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol*. 2004;160:509–16.
- White LF, Archer B, Pagano M. Estimating the reproductive number in the presence of spatial heterogeneity of transmission patterns. *Int J Health Geogr*. 2013;12:35.
- Métras R, Baguelin M, Edmunds WJ, Thompson PN, Kemp A, Pfeiffer DU, et al. Transmission potential of Rift Valley Fever virus over the course of the 2010 epidemic in South Africa. *Emerg Infect Dis*. 2013;19:916–24.
- Backer JA, Wallinga J. Spatiotemporal analysis of the 2014 Ebola epidemic in West Africa. *PLoS Comput Biol*. 2016;12:e1005210.
- Salje H, Cummings DAT, Lessler J. Estimating infectious disease transmission distances using the overall distribution of cases. *Epidemics*. 2016;17:10–8.
- Reiner RC, Le Menach A, Kunene S, Ntshalintshali N, Hsiang MS, Perkins TA, et al. Mapping residual transmission for malaria elimination. *ELife*. 2015;4:e09520.
- Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, et al. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *Proc Natl Acad Sci USA*. 2017;114:2337–42.
- Lau MSY, Gibson GJ, Adrakey H, McClelland A, Riley S, Zelner J, et al. A mechanistic spatio-temporal framework for modelling individual-to-individual transmission—With an application to the 2014–2015 West Africa Ebola outbreak. *PLoS Comput Biol*. 2017;13:e1005798.

9. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol.* 2014;10:e1003457.
10. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol.* 2014;31:1869–79.
11. Molentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc Biol Sci.* 2014;281:20133251.
12. Ypma RJF, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics.* 2013;195:1055–62.
13. Sturrock HJW, Hsiang MS, Cohen JM, Smith DL, Greenhouse B, Bousema T, et al. Targeting asymptomatic malaria infections: active surveillance in control and elimination. *PLoS Med.* 2013;10:e1001467.
14. Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, Takken W, et al. Hitting hotspots: spatial targeting of malaria for control and elimination. *PLoS Med.* 2012;9:e1001165.
15. Bejon P, Williams TN, Nyundo C, Hay SI, Benz D, Gething PW, et al. A micro-epidemiological analysis of febrile malaria in Coastal Kenya showing hotspots within hotspots. *ELife.* 2014;3:e02130.
16. Cohen JM, Moonen B, Snow RW, Smith DL. How absolute is zero? An evaluation of historical and current definitions of malaria elimination. *Malar J.* 2010;9:213.
17. Cohen JM, Le Menach A, Pothin E, Eisele TP, Gething PW, Eckhoff PA, et al. Mapping multiple components of malaria risk for improved targeting of elimination interventions. *Malar J.* 2017;16:459.
18. Churcher TS, Cohen JM, Novotny J, Ntshalintshali N, Kunene S, Cauchemez S. Measuring the path toward malaria elimination. *Science.* 2014;344:1230–2.
19. Routledge I, Chev   JER, Cucunub   ZM, Rodriguez MG, Guinovart C, Gustafson KB, et al. Estimating spatiotemporally varying malaria reproduction numbers in a near elimination setting. *Nat Commun.* 2018;26(9):2476.
20. Routledge I, Lai S, Battle KE, Ghani AC, Gomez-Rodriguez M, Gustafson KB, et al. Tracking progress towards malaria elimination in China: Individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. *PLoS Comput Biol.* 2020;16:e1007707.
21. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? *PLoS Pathog.* 2018;14:e1006885.
22. Huber JH, Johnston GL, Greenhouse B, Smith DL, Perkins TA. Quantitative, model-based estimates of variability in the generation and serial intervals of *Plasmodium falciparum* malaria. *Malar J.* 2016;15:490.
23. Marshall JM, Bennett A, Kiware SS, Sturrock HJW. The hitchhiking parasite: why human movement matters to malaria transmission and what we can do about it. *Trends Parasitol.* 2016;32:752–5.
24. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol.* 2019;15:e1006930.
25. Routledge I, Unwin HTJ, Bhatt S. Inference of malaria reproduction numbers in three elimination settings by combining temporal data and distance metrics. *Sci Rep.* 2021;11:14495.
26. Lemons DS, Langevin P. An introduction to stochastic processes in physics: containing "On the theory of Brownian motion" by Paul Langevin, translated by Anthony Gythiel. Baltimore: Johns Hopkins University Press; 2002. p. 110.
27. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys.* 1953;21:1087–92.
28. Hastings WK. Monte Carlo sampling methods using markov chains and their applications. *Biometrika.* 1970;57:97.
29. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics.* 2004;20:407–15.
30. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci.* 1992;7:457–72.
31. Cohen JM, Dlamini S, Novotny JM, Kandula D, Kunene S, Tatem AJ. Rapid case-based mapping of seasonal malaria transmission risk for strategic elimination planning in Swaziland. *Malar J.* 2013;12:61.
32. Hsiang MS, Ntuku H, Roberts KW, Dufour M-SK, Whittemore B, Tambo M, et al. Effectiveness of reactive focal mass drug administration and reactive focal vector control to reduce malaria transmission in the low malaria-endemic setting of Namibia: a cluster-randomised controlled, open-label, two-by-two factorial design trial. *Lancet.* 2020;395:1361–73.
33. Sturrock HJ, Cohen JM, Keil P, Tatem AJ, Le Menach A, Ntshalintshali NE, et al. Fine-scale malaria risk mapping from routine aggregated case data. *Malar J.* 2014;13:421.
34. Wesolowski A, Stresman G, Eagle N, Stevenson J, Owaga C, Marube E, et al. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. *Sci Rep.* 2015;4:5678.
35. Tessema S, Wesolowski A, Chen A, Murphy M, Wilheim J, Mupiri A-R, et al. Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. *ELife.* 2019;8:e43510.
36. Marshall JM, Wu SL, Sanchez CHM, Kiware SS, Ndhlovu M, Ou  draogo AL, et al. Mathematical models of human mobility of relevance to malaria transmission in Africa. *Sci Rep.* 2018;8:7713.
37. Blumberg S, Lloyd-Smith JO. Inference of R₀ and transmission heterogeneity from the size distribution of stuttering chains. *PLoS Comput Biol.* 2013;9:e1002993.
38. Blumberg S, Lloyd-Smith JO. Comparing methods for estimating R₀ from the size distribution of subcritical transmission chains. *Epidemics.* 2013;5:131–45.
39. Robert A, Kucharski AJ, Gastanhaduy PA, Paul P, Funk S. Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data. *J R Soc Interface.* 2020;17:20200084.
40. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. *Science.* 2012;338:267–70.
41. Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature.* 2018;553:333–6.
42. Wesolowski A, Taylor AR, Chang H-H, Verity R, Tessema S, Bailey JA, et al. Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Med.* 2018;16:190.
43. Roh ME, Tessema SK, Murphy M, Nhlabathi N, Mkhonta N, Vilakati S, et al. High genetic diversity of *Plasmodium falciparum* in the low-transmission setting of the Kingdom of Eswatini. *J Infect Dis.* 2019;220:1346–54.
44. Chang H-H, Wesolowski A, Sinha I, Jacob CG, Mahmud A, Uddin D, et al. Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. *ELife.* 2019;8:e43481.
45. Guzzetta G, Marques-Toledo CA, Ros   R, Teixeira M, Merler S. Quantifying the spatial spread of dengue in a non-endemic Brazilian metropolis via transmission chain reconstruction. *Nat Commun.* 2018;9:2837.
46. Salje H, Lessler J, Paul KK, Azman AS, Rahman MW, Rahman M, et al. How social structures, space, and behaviors shape the spread of infectious diseases using Chikungunya as a case study. *Proc Natl Acad Sci USA.* 2016;113:13420–5.
47. Guzzetta G, Vairo F, Mammone A, Lanini S, Poletti P, Manica M, et al. Spatial modes for transmission of chikungunya virus during a large chikungunya outbreak in Italy: a modeling analysis. *BMC Med.* 2020;18:226.
48. Weiss DJ, Lucas TCD, Nguyen M, Nandi AK, Bisanzio D, Battle KE, et al. Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *Lancet.* 2019;394:322–31.
49. Greenhouse B, Smith DL, Rodr  guez-Barraquer I, Mueller I, Drakeley CJ. Taking sharper pictures of malaria with CAMERAs: combined antibodies to measure exposure recency assays. *Am J Trop Med Hyg.* 2018;99:1120–7.
50. Wood SN. Generalized additive models: an introduction with R. Boca Raton, FL: Chapman & Hall/CRC; 2006.
51. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.