Epidemics 3 (2011) 109–118

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

The distribution of *Plasmodium falciparum* infection durations

Michael T. Bretscher [a,b,*], Nicolas Maire [a,b], Nakul Chitnis [a,b], Ingrid Felger [a,b], Seth Owusu-Agyei [c], Tom Smith [a,b]

[a] *Swiss TPH, Basel, Switzerland*
[b] *University of Basel, Basel, Switzerland*
[c] *Kintampo Health Research Center, Kintampo, Ghana*

ARTICLE INFO

Article history:
Received 26 November 2010
Revised 7 March 2011
Accepted 12 March 2011
Available online 28 March 2011

Keywords:
Plasmodium falciparum
Malaria
Duration of infection
Immigration-death model
Statistical analysis
Cohort study

ABSTRACT

Objectives: The duration of untreated *Plasmodium falciparum* infections in naturally exposed human populations is of interest for rational planning of malaria control interventions as it is related to the duration of infectivity. The extent of variability in duration is relevant where transmission is seasonal, and for the planning of elimination efforts. Methods for measuring these quantities from genotyping data have been restricted to exponential models of infection survival, as implied by constant clearance rates. Such models have greatly improved the understanding of infection dynamics on a population level but likely misrepresent the within-host dynamics of many pathogens. Conversely, the statistical properties of the distribution of infection durations, and how these are affected by exposure, should contain information on within-host dynamics.

Methods and results: We extended existing methods for the analysis of longitudinal genotyping data on *P. falciparum* infections. Our method simultaneously estimates force of infection, detectability, and the distribution of infection durations. Infection durations are modeled using parametric survival distributions. The method is validated using simulated data, and applied to data from a cohort study in Navrongo, Northern Ghana. Distribution estimates from exponential, Weibull, lognormal, and gamma models are compared with the distribution of durations in malariatherapy data.

Conclusions: The Weibull model fitted the data best. It estimated a shorter mean duration than the exponential model, which gave the worst fit. The distribution estimates appeared positively skewed when compared with the distribution of durations in malariatherapy data, suggesting that a significant proportion of infections is cleared shortly after inoculation. We conclude that malariatherapy data, the most important source of information on *P. falciparum* within-host dynamics, may not be representative of the actual processes in natural populations, and should be used with care. Further, conclusions from transmission models assuming exponential infection survival may be biased.

Introduction

The duration of *Plasmodium falciparum* infections is related to the duration of potential infectivity of the host. It affects the magnitude of transmission from humans to the mosquito population and gains special significance in settings where malaria transmission is seasonal: the fraction of infections surviving a transmission-free dry season constitutes the founder population for the new transmission season. Rational planning of a malaria control or elimination therefore profits from accurate measurements of infection duration. How much variation there is in the duration of natural infections is largely unknown, but important for similar reasons. A case study suggests that single infections may in extreme cases last up to 8 years (Szmitko et al., 2009). Current knowledge about within-host dynamics and the distribution of *P. falciparum* infection durations comes mostly from malariatherapy data (Sama et al., 2006a): before the arrival of suitable antibiotics, infection with malaria was a common method to treat neurosyphilis.

Analysis of such data is facilitated by the fact that the start- and end-points of every infection are approximately known, and that therefore standard methods of statistical survival analysis can be applied. A comparison of various parametric survival distributions suggested that the Gompertz and Weibull distributions gave the best fit to these data, followed by the gamma, lognormal, and exponential distributions (Sama et al., 2006a). An average duration of approximately 200 days was found. Infection durations much shorter or longer than the mean were rare.

However, malariatherapy data may not accurately mirror the situation in naturally exposed populations: the patients were immunologically naïve, infected with syphilis, and did not have multiple concurrent infections. Moreover, the *P. falciparum* strains

* Corresponding author at: London School of Hygiene and Tropical Medicine, London, UK. Fax: +44 20 7637 4314.
E-mail address: michael.bretscher@lshtm.ac.uk (M.T. Bretscher).

1755-4365 © 2011 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.epidem.2011.03.002

used for therapy were selected by physicians for optimal curative properties as well as for low clinical virulence (McKenzie et al., 2008). Thus it is not clear whether the distribution of infection durations would be the same in human hosts who have experienced high malaria transmission throughout their lives, and possibly have multiple concurrent infections caused by wild-type parasite strains.

A valuable source of information about infection dynamics in natural populations is cohort data on malaria infection, obtained using DNA-based diagnostic methods. These have the advantage that infecting clones can be distinguished on the basis of highly polymorphic genetic markers, such as Merozoite Surface Protein 2 (msp2). However, analysis of such data is not straightforward using standard techniques because detection of *P. falciparum* clones is imperfect, even when using polymerase chain reaction (PCR). Several dedicated statistical methods, allowing for imperfect detection, have been developed (Nagelkerke et al., 1990; Smith et al., 1999; Smith and Vounatsou, 2003; Sama et al., 2005, 2006b). The assumption of a constant clearance rate of infections is common to all these approaches. This has a long tradition and there are practical reasons for doing this: a constant clearance rate means that the rate at which infections are cleared is independent of the age of an infection, which implies an exponential distribution of infection durations. This simplifies the required mathematics enormously since it is not necessary to keep track of the age structure of the infection population. From a biological point of view, however, exponential survival of infections seems not very plausible. The study of how durations of infection are distributed is the quest for a statistical description of one important aspect of within-host dynamics. Such analyses may yield information which can be used to validate process-based within-host models.

In continuation of Smith et al. (1999), Smith and Vounatsou (2003), and Sama et al. (2005, 2006b) we have developed a method to analyze molecular cohort data and measure parameters of infection dynamics. We use a more complete dataset from the study analyzed by Sama et al. (2006b). The main findings of Sama et al. (2006b) are therefore briefly explained: using the statistical model described earlier in Sama et al. (2005), Sama et al. studied seasonality and age dependence of the following parameters of infection dynamics: λ (the force of infection, FOI), μ (the clearance rate, which is the inverse duration of infection and implies an exponential distribution of durations) and q (the detectability parameter, denoting the probability of detecting a specific clone in a blood sample, given it is present). In total, Sama et al. compared twelve different model parameterizations with respect to goodness of fit, using a longitudinal, age-stratified dataset from Navrongo, Northern Ghana. One of the main findings was that the detectability of infections declines with host age in a very pronounced way, suggesting an effect of cumulative exposure, a proxy for acquired immunity, on parasite densities. Contrary to expectation, the duration of infection was hardly affected by host age. The present analyses consider the sensitivity of these results with respect to the assumption of constant clearance rates, and provide estimates of the distribution of infection durations.

Methods

Study design and sample collection

A one year longitudinal study of malaria infection was conducted in the Kassena-Nankana district, in the upper East region of Ghana (Owusu-Agyei et al., 2002; Falk et al., 2006; Sama et al., 2005, 2006b). The malariological situation in this area is characterized by very high prevalence and multiplicity of infection (Owusu-Agyei et al., 2002; Binka et al., 1994), and year-round transmission with seasonal variation in transmission intensity (Sama et al., 2006b). A total of 349 individuals of all ages were followed up over one year in 2-monthly intervals. New births were recruited during the follow-up so as to ensure that the age distribution remained the same throughout the study. Blood was collected on ISOCodeStix™ PCR template preparation dipsticks (Schleicher & Schuell, Dassel, Germany).

Genotyping

DNA was eluted from ISOCodeStix™ and screened for presence of *P. falciparum* by PCR. Processing of stix and PCR conditions have been described in detail before (Felger et al., 1999). In brief, samples that tested positive for presence of *P. falciparum* were subjected to PCR using primers specific for the msp2 locus. Different alleles were distinguished on the basis of length polymorphisms, by means of automated capillary electrophoresis technology. The obtained data files were further processed using the GeneMapper® software and an in-house generated software, which facilitates identification of known alleles from the raw output of GeneMapper® and transforms the data into different formats suitable for data management and statistical analysis.

Data preparation

Only data of those participants who were present at all survey rounds were included in the analysis. This reduced the number of individuals in the dataset to 216. Failure or success to detect a strain was denoted by 0 or 1, respectively. The resulting 63 possible sequence types containing at least one positive test result were numbered from 1 to 63, using their binary value (e.g., 000010 is sequence 2). This yielded a frequency distribution of binary patterns for every host, to which statistical models could be fitted. The possibility of re-infection of a host with the same genotype was ignored for all modeling analyses. This assumption was justified by the high diversity of msp2 alleles in the population.

Models of infection dynamics

A selection of process-based statistical models, similar to the ones presented in Sama et al. (2006b), were devised and compared to the data. In the models, three main processes are assumed to determine frequencies of the different binary patterns in each human host: acquisition, clearance, and detection of infections. Given mathematical models for each of the three, a likelihood can be calculated as explained in the following section. The simplest possible model represents each process by a single parameter: the force of infection λ (no. of infections acquired per person year), the duration of a clonal infection (in the simplest case modeled as an exponentially distributed random variable, with scale parameter equal to the inverse clearance rate, $1/\mu$), and the detectability q (the probability of detecting any present *falciparum* clone in a blood sample by PCR). Such a simple model is not able to capture several important characteristics of real data, such as seasonality in transmission or changes in detectability with increasing immunity of the host. These have been shown by Sama et al. (2006b) to be present in the Navrongo dataset and need to be incorporated into a model in order to yield unbiased parameter estimates.

As a starting point for our analysis, we use the best fitting exponential model from Sama et al. (2006b), which was fitted to a partial dataset from the same study: the FOI parameter $\lambda(t)$ was modeled as a function of season alone, meaning that for every two-month season a separate parameter λ_i was estimated. The resulting pattern of seasonal transmission was assumed to have repeated since the birth of every host. We extended the work of Sama et al. (2006b) to allow the use of four parametric survival distributions for modeling of the clearance of infections: these are the exponential, Weibull, gamma and lognormal distributions (Table 1). Except for the exponential distribution, which is characterized by a single (scale) parameter, these distributions require two parameters. In the following we will refer to these as “scale” and “shape” parameters, ignoring possible distribution-specific names. Because the best-fitting model of Sama

Table 1
Survival distributions. Selected properties of the different survival models used in this analysis are given. The exponential is the only single-parameter distribution, and its properties depend entirely on the mean duration of an infection. All other distributions make use of two parameters. We refer to them as scale and shape, respectively, instead of using distribution-specific names. For each survival model, a restricted range of possible "shapes" exists, with restrictions being different among the distributions. Abbreviations: for the gamma function $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$, for the lower incomplete gamma function $\gamma(s,x)=\int_0^x t^{s-1}e^{-t}dt$, and for the error function $erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

Survival distribution	Scale	Shape	Mean	Variance	PDF	CDF
Exponential	$1/\mu>0$		$1/\mu$	$1/\mu^2$	$\mu e^{-\mu x}$	$1-e^{-\mu x}$
Weibull	$\lambda>0$	$k>0$	$\lambda\Gamma(1+\frac{1}{k})$	$\lambda^2\Gamma(1+\frac{2}{k})-\mu^2$	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$	$1-e^{-(x/\lambda)^k}$
Lognormal	μ	$\sigma>0$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	$\frac{1}{2}+\frac{1}{2}\,erf\left[\frac{\ln x-\mu}{\sigma\sqrt{2}}\right]$
Gamma	$\theta>0$	$k>0$	$k\theta$	$k\theta^2$	$x^{k-1}\frac{exp(-x/\theta)}{\Gamma(k)\theta^k}$	$\frac{\gamma(k,x/\theta)}{\Gamma(k)}$

et al. (2006b) showed no age-dependence of the duration of infection, and because the present analysis is intended to be a proof of concept, we chose to parameterize the survival models with simple constant values, rather than e.g. modeling them as functions of host age. The age dependence of detectability was modeled as a logit-linear function,

$$l(a)=\ln\left(\frac{q(a)}{1-q(a)}\right)=q_0+q_1(a-\bar{a}),$$

where a is the age of a host (in 2-month units), and $\bar{a}$ is the average age in the dataset.[1] The detectability of infections in a host of age a can then be obtained by using the inverse logit function:

$$q(a)=\frac{1}{e^{-l(a)}+1}. \quad (1)$$

Model equations

Let $(n_{k,1},n_{k,2},..,n_{k,63})$ denote the realizations of 63 Poisson random variables with means $(\omega_{k,1},\omega_{k,2},..,\omega_{k,63})$, where $\omega_{k,i}$ is the expected frequency of observed pattern i in individual k. In order to derive the $\omega_{k,i}$, we firstly derive the expected frequencies of the 21 (hypothetical) true patterns, $\tau_{k,i}$ representing true infection status. True patterns are also indexed using the binary number they encode, but only patterns comprising a single uninterrupted subsequence of "ones" are considered.

Individual k is of age b_k at the time of the baseline survey, t_b, implying that it was born at time t_b-b_k. As some parameters of infection dynamics can be functions of time (seasonality), and others functions of the age of the host, we will use t as variable of integration, and refer to the age of an individual at time t as $a_k(t)=t-t_b+b_k$. We denote the length of inter-survey intervals with ε, and assume equally spaced surveys.

Consider for the acquisition of infections a Poisson process with intensity $\lambda(t)$, the force of infection. Since true patterns consist of uninterrupted sequences of "ones", every true pattern can be defined by the times of the first and the last survey where the infection is present, $t_{1,i}$ and $t_{2,i}$, respectively. These imply, given the study design, a time interval where its causative infection may be acquired and a time interval where it must be cleared. Those intervals can be obtained for each true pattern from the age of the individual at baseline, b_k, and the number of surveys, s, in the study.

To illustrate this, we use examples of true patterns representative of the different possibilities. Our aim is to obtain the interval of possible infection time points, $[\alpha,\beta]$, and the interval within which an infection must be cleared if pattern i is generated, $[\gamma,\delta]$. Consider, for example, true pattern 110000. An infection generating this sequence can be acquired between birth of the host and the first survey. Thus, $\alpha=t_b-b_k$ and $\beta=t_b$. The situation is different if the time of first presence of the infection, $t_{1,i}$, is after baseline, as in pattern 001111. An infection which leads to this true pattern can only be acquired between the 2nd and the 3rd survey, therefore $\alpha=t_{1,i}-\epsilon$ and $\beta=t_{1,i}$. To summarize this, we write

$$\alpha_{k,i}=\begin{cases}t_b-b_k, & \text{if } t_{1,i}=t_b\\ t_{1,i}-\epsilon, & \text{otherwise}\end{cases}$$

and

$$\beta_i=t_{1,i}.$$

An infection acquired at time t will, given $\alpha\le t\le\beta$, generate true pattern i with a nonzero probability. We call this probability $p_{k,i}(t)$. If given $p_{k,i}(t)$, we can obtain $\tau_{k,i}$, the expected frequency of true pattern i in host k, as

$$\tau_{k,i}=\int_{\alpha_{k,i}}^{\beta_i}\lambda(t)p_{k,i}(t)dt. \quad (2)$$

This probability depends on the distance to the surveys in time, and the properties of the survival distribution used for modeling clearance of infections. The properties of the survival distribution may in turn depend on the age of the host at time t, when the infection is acquired.

The probability that an infection acquired at time t generates pattern i, $p_{k,i}(t)$, is equal to the probability that the infection is cleared in the interval $[\gamma_i,\delta_i]$. Therefore

$$\begin{aligned}p_{k,i}(t)&=\int_{\gamma_i}^{\delta_i}f(u-t)du\\&=S(\gamma_i-t)-S(\delta_i-t),\end{aligned}$$

where f and S are the probability density function and survivor function, respectively, of the survival distribution used to model clearance of infections.

To obtain γ_i, the start of the clearance interval, we consider again patterns 001111 and 110000 as examples, and conclude, trivially, that an infection cannot be cleared before $t_{2,i}$, the time of the last survey it is present conditional on producing pattern i.

$$\gamma_i=t_{2,i}.$$

The time point until an infection must be cleared in order to generate true pattern i depends on whether the last survey the infection is present coincides with the last survey of the study, or not. If so, it can be cleared anytime after the last survey, if not, the infection has to be cleared before the survey which follows the one at $t_{2,i}$, so

$$\delta_i=\begin{cases}\infty, & \text{if } t_{2,i}=t_b+s\epsilon\\ t_{2,i}+\epsilon, & \text{otherwise}\end{cases}.$$

[1] For the sake of comparing our results with those from Sama et al. (2006b), $\bar{a}$ was set to 120.72 (in units of 2 months, corresponding to the survey interval). This is the average age in the partial dataset used by Sama et al. (2006b).

The complete expression for the number τ_i of true patterns of type i to be expected in host k is then

$$\tau_{k,i} = \int_{\alpha_{k,i}}^{\beta_i} \lambda(t) \int_{\gamma_i}^{\delta_i} f(u-t)du \; dt,$$

or, in terms of the survivor function S,

$$\tau_{k,i} = \int_{\alpha_{k,i}}^{\beta_i} \lambda(t)[S(\gamma_i - t) - S(\delta_i - t)]dt. \quad (3)$$

A more formal but equivalent approach, explaining the presented heuristics, is outlined in Appendix A.

The expected frequencies $\omega_{k,i}$ of observed patterns in individual k can be obtained using the probability $P_{i,j}$ that true pattern i gives rise to observed pattern j, as follows:

$$\omega_{k,j} = \sum_i P_{i,j}(q_k)\tau_{k,i},$$

where q_k is the detectability of infections within host k at the time of the study. To calculate $P_{i,j}$ we denote the individual digits of either binary sequence by $d_{n,i} \in \{0,1\}$, and $d_{n,j} \in \{0,1\}$. Then the probability that true pattern i gives rise to observed pattern j is calculated as

$$P_{i,j} = \prod_{n=1}^{s} o\left(d_{n,i}, d_{n,j}\right),$$

where s is the number of surveys, and $o(d_{n,i}, d_{n,j})$ is the probability that true presence or absence of a particular genotype at position n results in a positive or negative outcome of detection, assuming perfect specificity:

$$o\left(d_{n,i}, d_{n,j}\right) = \begin{cases} 1, & \text{if } d_{n,i} = 0 \text{ and } d_{n,j} = 0 \\ 0, & \text{if } d_{n,i} = 0 \text{ and } d_{n,j} = 1 \\ 1-q, & \text{if } d_{n,i} = 1 \text{ and } d_{n,j} = 0 \\ q, & \text{if } d_{n,i} = 1 \text{ and } d_{n,j} = 1 \end{cases},$$

where q is the host-specific detectability, possibly modeled as a function of the age of the host. Considering all observed patterns j and all hosts k, and assuming a Poisson distribution of the actual data $n_{k,j}$ with expectations $\omega_{k,j}$ we obtain the overall likelihood

$$L_{Data} = \prod_k \prod_j \frac{e^{-\omega_{k,j}} \omega_{k,j}^{n_{k,j}}}{n_{k,j}!}.$$

Since the terms involving $n_{k,j}$ are independent of the statistical model fitted, they can be omitted from the likelihood computations without altering the ranking of models. The statistical models can then be compared using Akaike's information criterion (AIC).[2]

Model implementation and parameter estimation

All models were implemented using the Java™ programming language. Maximum-likelihood estimates of parameters were obtained by minimization of AIC values using the "UncMin" algorithm by Schnabel et al. (1985). A Java version of this algorithm was obtained from http://www1.fpl.fs.fed.us/optimization.html. Numerical integration was performed using a Romberg integration algorithm with modified stopping criterion,[3] from the Apache Commons Math Library (The Apache Software Foundation, 2010).

A major challenge was to reduce the required computation time such that models could be fitted within acceptable time by a single-processor computer. Apart from choosing a gradient-based optimization algorithm, this could be achieved by making some of the numerical integrations redundant through discretization of host ages. To this end the following assumption was introduced: the expected frequencies of any true pattern i in two different hosts are assumed to be equal, if $t_{1,i}$ is not at baseline and if the two hosts are in the same age group throughout the time interval where pattern i can be acquired. The reason for not pooling the patterns where $t_{1,i}$ coincides with the baseline survey, is the following: since host age has two distinct meanings in the context of our model, namely age of the host as a measure for immunity, and age of the host as time of exposure, one could not simply group hosts by age without altering the results. As an example, it may seem reasonable to have an age group ranging from 3 to 5 years, as immunity would – by hypothesis – not change very much within this age range. But, a host of age 5 will have had 2 years more time to acquire infections and may – depending on the shape of the survival distribution of infections – have a higher multiplicity of infection (MOI), and different pattern frequencies. However, if only one of the two collinear time variables is discretized, namely host age as measure of immunity, said error is not introduced, while some integrals become redundant and only need to be calculated once.

Results

Simulated data

For the purpose of validating our method, simulated datasets were produced using Monte-Carlo simulation. Number and ages of the hosts in the simulated datasets were identical to the Ghanaian dataset, and a constant, homogeneous FOI of 18 infections per person and year was assumed, as approximately measured on average by Sama et al. (2006b). The number of infections a person experienced between birth and the last survey round was sampled from a Poisson distribution with mean λa, where λ is the force of infection, and a is the age of the human host at the last survey round. Actual infection time-points were then sampled from a uniform distribution within said interval.

Subsequently, a duration was assigned to each infection, using one of four survival distributions with parameter values as measured from malariatherapy data (Sama et al., 2006a). A Bernoulli random variable with mean $q = 0.5$ was then used to determine for each survey round and clone whether detection was successful or not. All parameters of infection dynamics could be recovered well from the simulated data, as shown in Table 2 and Fig. 1.

Estimates from the Ghanaian dataset

All of the 216 study participants included in the statistical analysis tested positive for *P. falciparum* on at least one survey. Parasite prevalence in the dataset was 48% by microscopy, 75% by PCR, and the mean MOI was 4.5 per person (these measures are not corrected for imperfect detection). A total of 103 different msp2 genotypes were found, with the most frequent genotype representing 10.2% of all fragments detected.

The four different models for infection survival showed the following order of goodness of fit, as measured by AIC: the Weibull model fits the data best (AIC: 8029.1), followed by the gamma (AIC: 8029.4), lognormal (AIC: 8045.1), and exponential model (AIC: 8127.4). Parameter estimates are given in Table 3, and the correlation matrix of the Weibull model in Table 4. All non-exponential distributions show an increased clearance in the early stages of an infection, i.e. they are positively skewed (to the left). The estimated mean durations, which can be calculated from the scale and shape parameters (Table 3) and the distribution-specific expressions for the mean (Table 1), are as follows (in days): 139.9 (Weibull), 54.9 (gamma), 205.3 (lognormal),

[2] AIC was calculated as $2n - 2l$, with the number of parameters n and the log-likelihood l.

[3] Absolute instead of relative precision was used as stopping criterion. This substantially reduced the computation time needed, presumably because the relative change in integral values per iteration may become smaller than machine precision for true patterns with very low expected frequency.

Table 2
Parameter estimates from simulated datasets. The data were produced by stochastic simulation using survival models and parameter values from Sama et al. (2006a), with constant values of FOI and detectability. Every row corresponds to a survival model tested on a simulated dataset. Columns correspond to the different parameters of infection dynamics, with estimated parameter values shown to the left of the true values. The FOI is given in infections acquired per year, otherwise the time unit is per 2 months (corresponding to the survey interval).

	FOI		Scale		Shape		Detectability	
	$\hat{\lambda}$	λ	$\hat{s}_1$	s_1	$\hat{s}_2$	s_2	$\hat{q}$	q
Exponential	19.05	18	3.22	3.53			0.48	0.5
Weibull	17.95	18	3.93	3.94	2.38	2.2	0.49	0.5
Lognormal	17.90	18	1.18	1.11	0.53	0.53	0.49	0.5
Gamma	17.25	18	0.72	1.19	5.23	3	0.49	0.5

and 219.7 (exponential). There is a substantial difference between the FOI estimates of the gamma and Weibull models, which are very similar in terms of goodness of fit (Fig. 3). Measurements of detectability are in good quantitative agreement (Fig. 4). All models measure a detectability of between 40% and 50% in young ages, which decreases to just above 10% in the old ages.

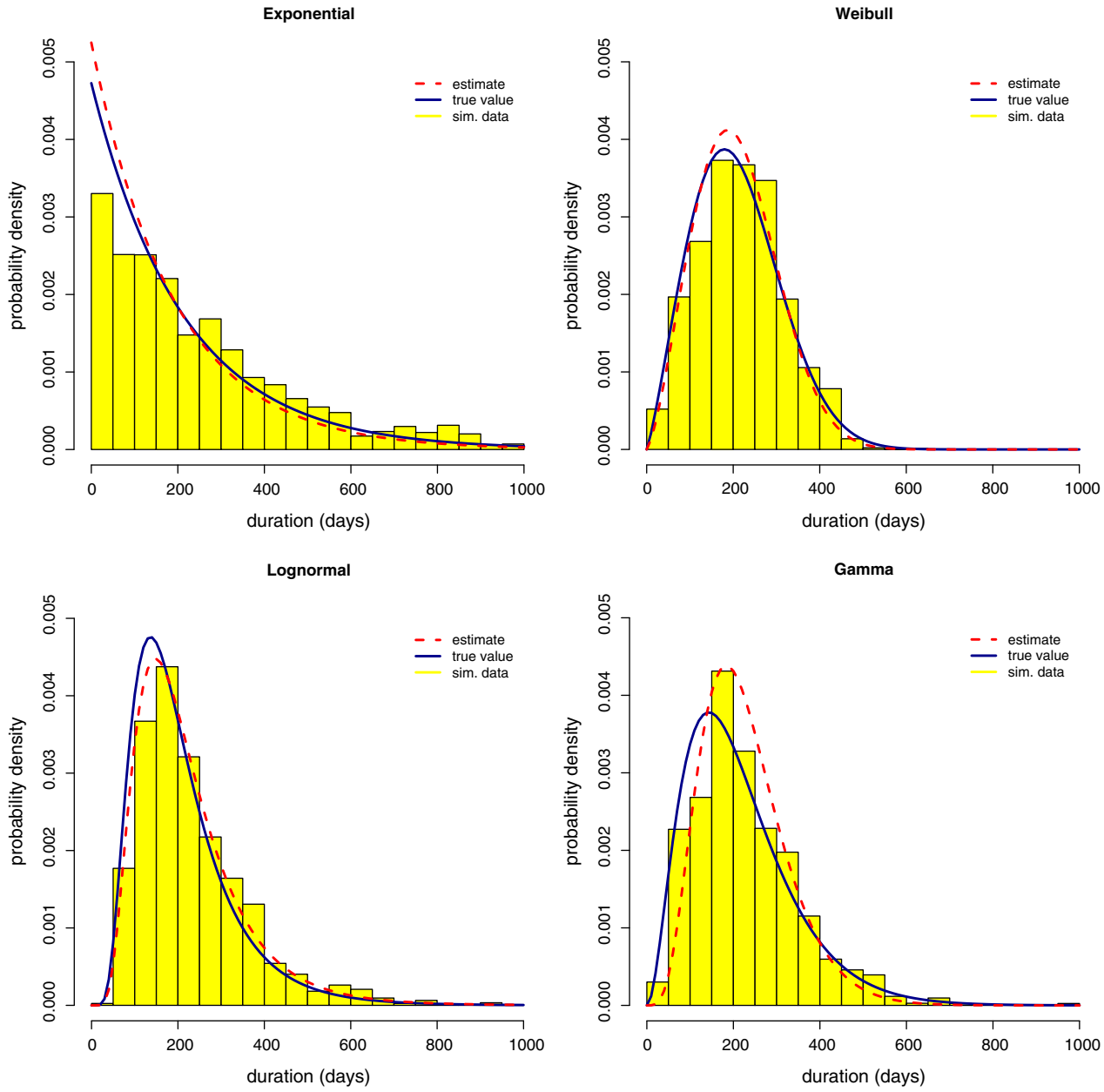

Fig. 1. Validation using simulated data. Exponential, Weibull, lognormal and gamma models were fitted to simulated datasets. The histograms indicate the actual distribution of the duration of those infections which were present at some point during the simulated study. Solid blue lines indicate the PDF of the distribution that the durations were sampled from — the estimates from malariatherapy data given by Sama et al. (2006b). The dashed lines indicate the PDFs of the distributions as recovered by our statistical model.

Table 3
Parameter estimates from Ghanaian dataset. The results of fitting four statistical models to the Ghanaian dataset are shown. These models use different survival distributions for modeling of the clearance of infections. Goodness of fit is indicated by AIC, and parameter values are given plus/minus the corresponding standard error. The Weibull fits the data best, as indicated by its lowest value of AIC, but is closely followed by the gamma model. Seasonality in the force of infection was accounted for by estimating a separate parameter for every two-month season, assuming the same seasonality pattern had repeated in the past. Scale and shape parameters are specific to each distribution, as explained in Table 1. The detectability can be calculated for a host of given age using Eq. (1).

Survival model	Force of infection (by season)						Survival		Detectability		AIC
	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	s_1	s_2	q_1	q_2	
Weibull	62.4 ± 0.04	35.9 ± 0.04	32.0 ± 0.09	10.6 ± 0.10	21.1 ± 0.10	49.1 ± 0.10	1.51 ± 7E−3	0.59 ± 6E−3	−0.86 ± 0.03	−0.00533 ± 2E−4	8029.1
Gamma	160.2 ± 0.02	95.8 ± 0.02	91.0 ± 0.02	8.5 ± 0.02	58.4 ± 0.02	123.2 ± 0.02	8.63 ± 0.02	0.11 ± 0.02	−0.87 ± 0.03	−0.00530 ± 3E−4	8029.4
Lognormal	40.4 ± 0.12	25.7 ± 0.13	21.4 ± 0.13	7.5 ± 0.11	15.0 ± 0.14	38.6 ± 0.12	1.66 ± 0.02	1.20 ± 0.02	−0.89 ± 0.03	−0.00541 ± 3E−4	8045.1
Exponential	43.9 ± 0.1	20.0 ± 0.1	18.9 ± 0.1	4.5 ± 0.1	14.9 ± 0.1	40.7 ± 0.1	3.66 ± 0.02		−0.98 ± 0.03	−0.00507 ± 2E−4	8127.4

Discussion

By extending the method of Sama et al. (2005, 2006b), such that parametric survival distributions can be used for modeling of infection survival, it has become possible to obtain more detailed information on clearance of *P. falciparum* infections in naturally exposed populations. A validation with simulated datasets suggests that parameters are identifiable using our approach and the present study design.

Distribution estimates

All estimated distributions are positively skewed (Fig. 2), with exception of the exponential distribution, which does not have the freedom to measure more than the mean duration. Positive skewness means that most infections are cleared rather soon after inoculation. This is very different from the estimates of Sama et al. (2006a) who find that most infections in malariatherapy data last for an amount of time similar to the mean duration, and that very short or very long infections are rare. It also becomes apparent that in the Weibull and gamma models, which fit the data best, this effect is more pronounced than in the lognormal model. We attribute this to the fact that the PDF of a lognormal distribution is constrained to be unimodal and therefore does not have the freedom to estimate such extreme early clearance of infections, as would be required in order to attain a better fit to the data. Thus, a consistent picture emerges, where the finding of early clearance of infections is independent of the clearance model. This positive skew of distributions is consistently associated with a lower estimated mean duration, which suggests that assuming an exponential distribution may lead to an overestimation of the mean infection duration. Across the models, shorter estimates of the mean duration are in turn associated with higher estimates of FOI. This presumably reflects the simple notion that a given number of detected infections can be explained by either higher FOI and shorter duration, or the other way round. Detectability estimates seem to largely agree across the different survival models, with only the exponential model estimating a slightly lower overall detectability (Fig. 4).

Table 4
Correlation matrix. Correlation matrix for the parameters of the best-fitting Weibull model. An expected overall negative correlation between the average FOI and the mean duration of infection may be obscured, as these are functions of several parameters. However, a clear positive correlation between the shape parameter s_2 and the detectability of an individual of average age, q_1, indicates that different interpretations of the data are possible concerning clearance of infections and detectability.

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	s_1	s_2	q_1	q_2	
1	−0.04	0.12	0.18	0.11	0.17	0.01	0.01	−0.01	0.00	λ_1
	1	0.11	−0.06	0.10	−0.06	0.03	0.02	0.00	0.00	λ_2
		1	0.30	0.89	0.33	−0.15	0.01	0.00	−0.01	λ_3
			1	0.32	0.90	0.07	0.06	−0.02	−0.01	λ_4
				1	0.36	−0.14	0.01	0.00	−0.01	λ_5
					1	0.06	0.06	−0.02	−0.01	λ_6
						1	0.05	−0.14	−0.02	s_1
							1	0.48	0.09	s_2
								1	0.12	q_1
									1	q_2

Validation

The successful recovery of parameters from the simulated datasets suggests that our statistical approach in conjunction with the study design in principle allows for correct identification of the distributions of infection duration. As a caveat, the simulated data is certainly idealized compared to a real dataset: there is no seasonal or spatial heterogeneity in the FOI, and no inter-individual variation in parameters concerning clearance and detection of infections.

Simulated datasets can also be used to establish optimal study designs for measuring a certain quantity of interest. In this case, we could confirm that the study design is suitable for identification of the distribution of infection durations — given it is similar to the published measurements from malariatherapy data (Sama et al., 2006a). The high discrepancy of the two best fitting models (Weibull and gamma) concerning duration — and FOI estimates from the Ghanaian dataset may be connected to the study design: the differences between the two models fall largely into an interval of infection durations which is shorter than the survey interval of 60 days. The bulk of infections can therefore not be observed very well with the current study design, and shorter survey intervals seem necessary to obtain more accurate measurements. According to Bretscher et al. (2010), there is a lower limit of one week for survey intervals. This has to do with the fact that our method assumes statistical independence of detections at different survey rounds. Due to the complicated temporal behavior of *P. falciparum* detectability at short time scales, we suggest a survey interval of 2 weeks or more for future, similar studies.

The difference to malariatherapy data

Estimates from the Ghanaian dataset suggested the presence of a large proportion of short infections, which is in disagreement with previous measurements from malariatherapy data. Hypotheses on the cause of this difference can broadly be divided into two groups: hypotheses attributing the differences to the distinct immune status of the malariatherapy and Ghanaian study populations, respectively, and other hypotheses. Here, we will shortly line out a range of possible explanations.

Acquired immunity could make it harder for the parasite to permanently colonize a host which was exposed before. Since the FOI in the Ghanaian dataset is very high, the population can be considered immune on average, while the malariatherapy patients are immunologically naïve. This hypothesis requires the effects of immunity to last and thus to accumulate over time. Because the age of a human host can be used as a proxy for exposure, an increase of such long-lasting immunity effects with age is expected, which should lead to more short infections in older individuals. Whether this is indeed the case could be tested by comparing host age groups using the present analysis method.

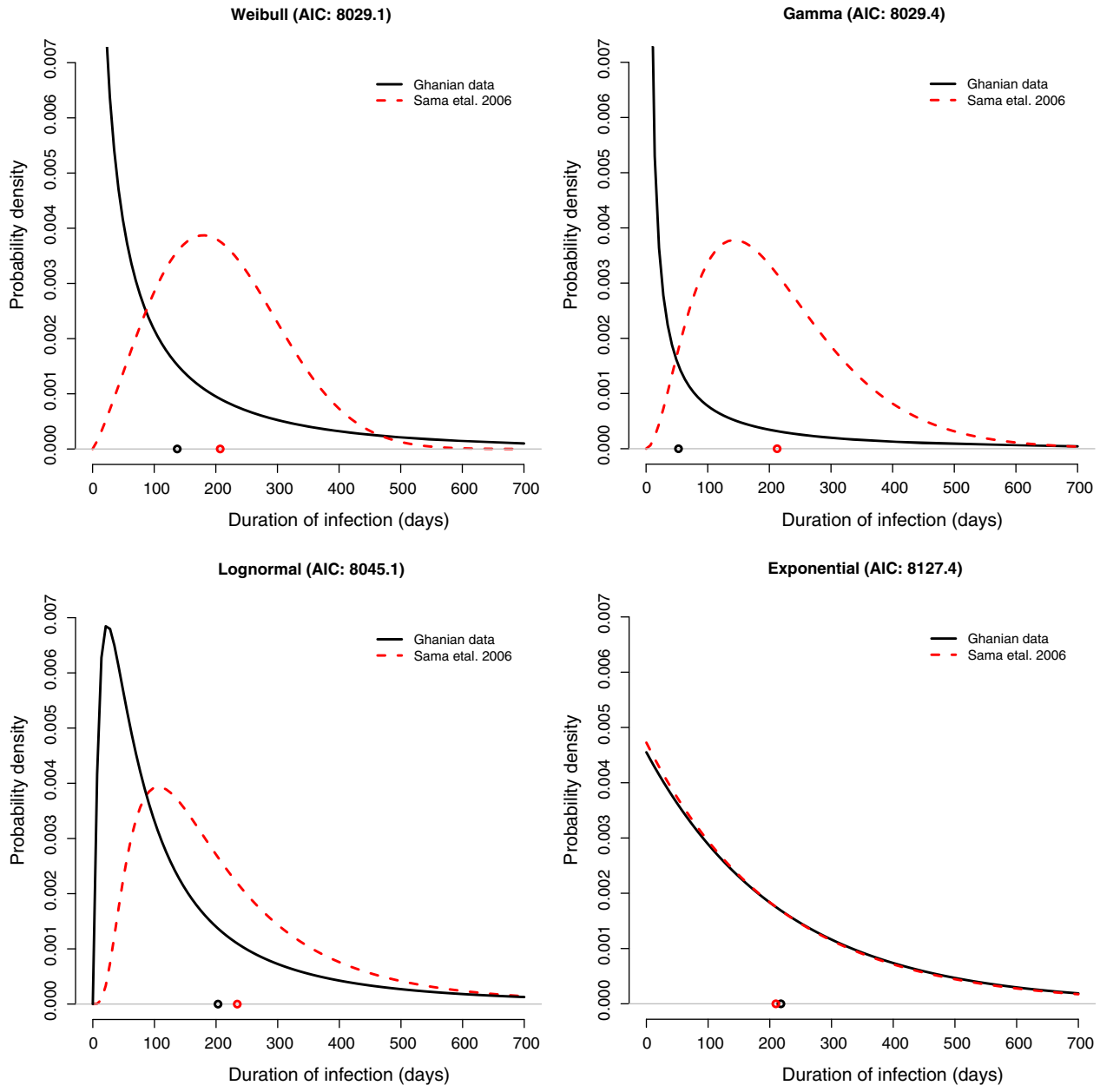

Fig. 2. Comparison of results with malariatherapy data. The PDFs of the distributions of infection duration as measured from the Ghanaian dataset (solid lines) are shown together with the estimates from malariatherapy data by Sama et al. (2006a) (dashed red lines), in order of decreasing goodness of fit. Circles on the abscissa indicate the corresponding mean durations. The Weibull survival model fitted the data best, followed by the gamma, lognormal, and exponential models. With the exception of the gamma model, estimated mean durations differ only slightly between malariatherapy data and the data from the naturally exposed population. All non-exponential distributions estimated from the Ghanaian dataset are positively skewed.

Short-lived effects of immunity, on the other hand, may include the interaction of concurrent infections within one host: the host populations from the two datasets also differ with respect to their MOI, as there are only single infections in the malariatherapy patients. An effect of interactions between concurrent infections, mediated by short-term effects of immunity, might be confirmed by observing a change of the distribution of infection durations with MOI.

Various factors not related to immunity also have the potential to explain the observed distribution of infection durations. Among these are the following: heterogeneous (unreported) treatment in the population could clear all infections early in some individuals, who mostly treat their infections, and let infections persist in another subpopulation, which rarely treats their infections. Averaged over the study population, this should convey a picture which is consistent with the results of this analysis. This explanation appears, however, unlikely when comparing the number of treatments sold by local health centers to estimates of the expected number of episodes in the area. An alternative explanation attributes the difference of distribution estimates to genetic differences between malariatherapy strains and wild type strains in the Navrongo area. It seems plausible that doctors treating syphilis patients with *P. falciparum* would not favor strains which are cleared after a very short time, requiring a re-infection of the patient. Yet, natural selection may well be doing the same, as a shorter infection duration reduces the R_0 of a strain. Other possibilities include genetic differences between the Ghanaian population and the malariatherapy patients, perhaps with respect to mutations protective against malaria, or an interaction of syphilis with malaria in the patients. In addition, differences in the infective dose or the route of

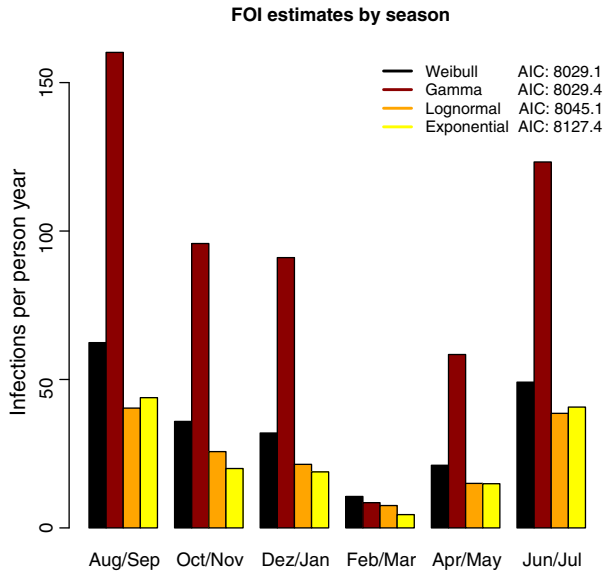

Fig. 3. Estimates of the force of infection. Different seasonal patterns of FOI were measured by the four models. Each group of bars compares the estimates of all the statistical models for a given season, and differences represent the uncertainty in measurements of the FOI with respect to assumptions about clearance of infection. Within one season, estimates are arranged from left to right in order of decreasing goodness of fit of the corresponding survival model. The gamma model estimated the highest FOI, which is consistent with it also estimating the shortest average duration of infection (see Fig. 2). The overall pattern of seasonality in transmission is consistent across the models.

infection may play a role, as some of the malariatherapy patients were infected using sporozoites, either through mosquito bites or via subcutaneous injection, and others through infected blood (Sama et al., 2006a; Collins and Jeffery, 1999).

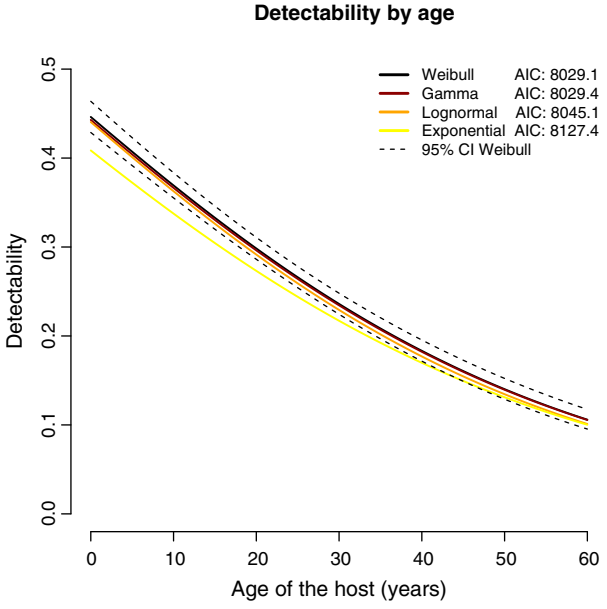

Fig. 4. Estimates of detectability. The obtained estimates of detectability differ only slightly across different models of infection clearance. A detectability below 50% is estimated consistently, and all models agree on a decrease of detectability with host age. The logit-linear relationship of q with age, which is assumed here, does not allow for e.g. a peak in the youngest children. This may lead to inaccurate estimates for these age groups, but since the proportion of infants in the dataset is small, this is unlikely to influence estimates of other parameters.

Limitations of the method

The application of our statistical method to data requires, for now, the assumption that re-infection with the same genetic marker is a rare event. This assumption has been discussed before (Gatton and Cheng, 2008). However, it gains special significance in cohort studies of long duration: if an immigration-death model is used, what matters is not only the probability to find in a host more than one infecting clone with identical marker genotype at any one time, as considered in Gatton and Cheng (2008), but rather the probability that an individual experiences more than one infection with the same genotype within the study period. The latter must depend on marker diversity, the force of infection as well as on the study duration. For practical purposes, the validity of the assumption can be tested for a given dataset by successively removing the most frequent marker allele from the analysis and observing a possible change in parameter estimates.

Conclusions

The estimated distribution of *P. falciparum* infection durations in exposed individuals in Northern Ghana is different from the distribution in malariatherapy infections (Sama et al., 2006a). This difference is mainly in the shape of the distributions: in the Ghanaian population, many infections are cleared at an early stage and others remain for a long time, while in the malariatherapy data infections are most often cleared close to their expected age at clearance (the mean duration of infection). The measured mean duration is shorter for the more flexible survival models compared to the exponential distribution. At this point it is not possible to decide among a multitude of possible hypotheses as to what causes the different distributions of infection durations in the two datasets. We have demonstrated that it is possible to gain information about the distribution of durations from longitudinal genotyping data, together with other parameters of infection dynamics. Our method represents – for the part concerning clearance of infections – an extension of existing methods of survival analysis, with the additional complication that the actual time-points of truncation and censoring are different for every infection, unknown and stochastic. This uncertainty is overcome by inferring simultaneous estimates of FOI and detectability. The software used to carry out the analyses can be obtained as a platform-independent Java™ executable on http://www.swisstph.ch/resources/software.html. There might be situations where assuming an exponential decay of infections can be a good assumption in order to reduce the number of parameters in the statistical model. Such a situation may occur if the total duration of a study is too short to contain sufficient information on the higher moments of the distribution of infection durations.

Authors' contributions

MTB performed the statistical analysis and drafted the manuscript. NM assisted in solving computational issues. NC contributed to the mathematics presented in the appendix. IF supervised the molecular genetic studies. SO coordinated the data collection. TS carried out the study design, assisted in the statistical analysis, and helped drafting the manuscript. All authors read and approved the final manuscript.

Conflicts of interest statement

The authors have no conflicts of interest concerning the work reported in this paper.

Acknowledgments

This work was supported by Bill & Melinda Gates grants nos. 39777 and 39777.01, Swiss National Science Foundation project no. 320030–125316, and MACEPA. The computations were performed at the Basel

Computational Biology Center (www.bc2.ch/center) of the Biozentrum, University of Basel and SIB Swiss Institute of Bioinformatics. The support of volunteers at www.malariacontrol.net, contributing computation power via the internet, was indispensable for an earlier, individual-based version of the analysis method. We extend our sincere thanks to the study participants and the staff of the Navrongo Health Research Centre involved in the field work. We thank Prof. Klaus Dietz for helpful comments.

Appendix A. Relationship to existing methods

The relationship between the method described above and previously published methods, most exhaustively explained in Sama et al. (2005), is not immediately apparent. We illustrate the mathematical relationship of the presented heuristics with standard approaches of modeling immigration-death processes briefly, and show that the two approaches lead to equivalent expressions.

Rather than calculating the expected frequencies of true pattern types, we consider for the purpose of this illustration the simpler problem of calculating the number of infections present at any time point x, using t as variable of integration. Acquisition of infections at a rate $\lambda(t)$ is assumed to occur within the time interval $[0,x]$. Survival of infections is modeled using parametric survival distributions. These appear in form of the hazard $h(a)$, which for every infection depends on its current age a. The hazard is defined as

$$h(a) = -\frac{S'(a)}{S(a)}, \tag{.1}$$

where the survivor function $S(a)$ is the fraction of infections surviving at least until age a. Its negative derivative is the PDF of the corresponding parametric survival distribution. The hazard is therefore the rate at which surviving infections of age a are being cleared.

A.1. Exponential survival of infections

We consider first the special case of exponential survival, where the age-independent hazard is often called clearance rate and denoted by μ. In analogy to Eq. (2) we write the number of infections $n(x)$ at time x as

$$n(x) = \int_0^x \lambda(t) p_x(t) dt.$$

The probability $p_x(t)$ that an infection acquired at time t will still be present at time x is simply equal to $S(x-t)$. The survivor function of the exponential distribution has the form $S(a) = e^{-\mu a}$, which, assuming a constant force of infection λ, leads to

$$n(x) = \lambda \int_0^x e^{-\mu(x-t)} dt,$$

for the number of infections $n(x)$ present at time x. The value of this integral is

$$n(x) = \left[\frac{\lambda}{\mu} e^{-\mu(x-t)}\right]_0^x = \frac{\lambda}{\mu}(1-e^{-\mu x}).$$

This is a familiar result and the solution of the differential equation

$$\frac{dn(x)}{dt} = \lambda - \mu n(x), \tag{.2}$$

with $n(0) = 0$, which constitutes a simple model for superinfection and is explained in Dietz (1988). In fact, it was this model of superinfection in connection with the CDF of the exponential distribution which allowed Sama et al. (2005) to work out all expected true pattern frequencies. Our approach to calculating these frequencies is therefore equivalent in the case of exponential survival of infections.

A.2. Non-exponential survival of infections

In the general case, a model for the age structure of the parasite population within a host is required. Such a model is given by the McKendrick–von Foerster equation (Hethcote, 2000), a partial differential equation (PDE) of the form

$$\frac{\partial u(a,x)}{\partial x} + \frac{\partial u(a,x)}{\partial a} = -h(a)u(a,x), \tag{.3}$$

with boundary conditions $u(0,x) = \lambda(x)$ and $u(a,0) = 0$. The function $u(a,x)$ denotes the age-density[4] of infections with a certain age a after time x, and $\lambda(x)$ is the force of infection, the rate at which infections enter the population with an age of 0. Given $u(a,x)$, the total number of infections present after time x is

$$n(x) = \int_0^x u(a,x) da, \tag{.4}$$

the integral of $u(a,x)$ over all existing ages. Eq. (.3) can be solved using the "method of lines", which yields

$$u(a,x) = \lambda(x-a) S(a). \tag{.5}$$

By inserting the solution for u into Eq. (.4) we obtain the cumulative number of infections of all ages present at a time point x as

$$n(x) = \int_0^x \lambda(x-a) S(a) da. \tag{.6}$$

By substitution of the integration variable as $t = x - a$ and reversing integration we obtain

$$n(x) = \int_0^x \lambda(t) S(x-t) dt. \tag{.7}$$

This expression can also be obtained from Eq. (3), as the special case when $\alpha = 0$, $\beta = \gamma = x$ and $\delta \to \infty$[5]:

$$n(x) = \int_0^x \lambda(t) S(x-t) dt = \int_\alpha^\beta \lambda(t) \left[S(\gamma - t) - \underbrace{\lim_{\delta \to \infty} S(\delta - t)}_{0} \right] dt.$$

The approach described in this paper therefore represents an extension of the approach by Sama et al. (2005), making it possible to use non-exponential survival distributions in models of superinfection.

References

Binka, F., Morris, S., Ross, D., Arthur, P., Aryeetey, M., 1994. Patterns of malaria morbidity and mortality in children in northern Ghana. Trans. R. Soc. Trop. Med. Hyg. 88, 381–385.
Bretscher, M.T., Valsangiacomo, F., Owusu-Agyei, S., Penny, M.A., Felger, I., Smith, T., 2010. Detectability of *Plasmodium falciparum* clones. Malar. J. 9 (1), 234 PMID: 20718959.
Collins, W.E., Jeffery, G.M., 1999. A retrospective examination of the patterns of recrudescence in patients infected with *Plasmodium falciparum*. Am. J. Trop. Med. Hyg. 61, 44–48.
Dietz, K., 1988. Mathematical models for transmission and control of malaria. In: Wernsdorfer, W.H., McGregor, I. (Eds.), Malaria, Principles and Practice of Malariology. Churchill Livingstone, Edinburgh, pp. 1091–1133.
Falk, N., Maire, N., Sama, W., Owusu-Agyei, S., Smith, T., Beck, H., Felger, I., 2006. Comparison of PCR-RFLP and genescan-based genotyping for analyzing infection dynamics of *Plasmodium falciparum*. Am. J. Trop. Med. Hyg. 74, 944–950.
Felger, I., Irion, A., Steiger, S., Beck, H., 1999. Genotypes of merozoite surface protein 2 of *Plasmodium falciparum* in tanzania. Trans. R. Soc. Trop. Med. Hyg. 93, 3–9.

[4] Density not in the sense of a probability density. Rather, in analogy to the density of mass in physics, $u\,da$ denotes the number of infections within the age range $[a, a+da]$.

[5] Corresponding to the number of infections present at time x which were acquired between $\alpha = 0$ and $\beta = x$, and are cleared anytime between $\gamma = x$ and $\delta \to \infty$.

Gatton, M.L., Cheng, Q., 2008. Can estimates of antimalarial efficacy from field studies be improved? Trends Parasitol. 24, 68–73.
Hethcote, H., 2000. The mathematics of infectious diseases. SIAM Rev. 42, 599–653.
McKenzie, F.E., Smith, D.L., O'Meara, W.P., Riley, E.M., 2008. Strain theory of malaria: the first 50 years. Adv. Parasitol. 66, 1–46.
Nagelkerke, N.J., Chunge, R.N., Kinoti, S.N., 1990. Estimation of parasitic infection dynamics when detectability is imperfect. Stat. Med. 9 (10), 1211–1219.
Owusu-Agyei, S., Smith, T., Beck, H., Amenga-Etego, L., Felger, I., 2002. Molecular epidemiology of *Plasmodium falciparum* infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana. Trop. Med. Int. Health 7, 421–428.
Sama, W., Dietz, K., Smith, T., 2006a. Distribution of survival times of deliberate *Plasmodium falciparum* infections in tertiary syphilis patients. Trans. R. Soc. Trop. Med. Hyg. 100 (9), 811–816.
Sama, W., Owusu-Agyei, S., Felger, I., Dietz, K., Smith, T., 2006b. Age and seasonal variation in the transition rates and detectability of *Plasmodium falciparum* malaria. Parasitology 132, 13–21.
Sama, W., Owusu-Agyei, S., Felger, I., Vounatsou, P., Smith, T., 2005. An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect. Stat. Med. 24, 3269–3288.
Schnabel, R.B., Koonatz, J.E., Weiss, B.E., 1985. A modular system of algorithms for unconstrained minimization. ACM Trans. Math. Software 11, 419–440.
Smith, T., Felger, I., Fraser-Hurt, N., Beck, H., 1999. Effect of insecticide-treated bed nets on the dynamics of multiple *Plasmodium falciparum* infections. Trans. R. Soc. Trop. Med. Hyg. 93 (Suppl 1), 53–57.
Smith, T., Vounatsou, P., 2003. Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. Stat. Med. 22, 1709–1724.
Szmitko, P.E., Kohn, M.L., Simor, A.E., 2009. *Plasmodium falciparum* malaria occurring 8 years after leaving an endemic area. Diagn. Microbiol. Infect. Dis. 63, 105–107.
The Apache Software Foundation, 2010. Apache Commons Math Library, Release 2.1. http://commons.apache.org/math2010.