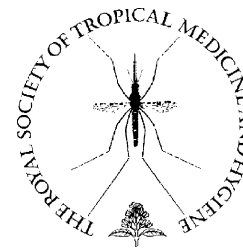




available at www.sciencedirect.com



journal homepage: www.elsevierhealth.com/journals/trst



Distribution of survival times of deliberate *Plasmodium falciparum* infections in tertiary syphilis patients

W. Sama^{a,*}, K. Dietz^b, T. Smith^a

^a Department of Public Health and Epidemiology, Swiss Tropical Institute, Socinstrasse 57, Postfach CH-4002, Basel, Switzerland

^b Institut für Medizinische Biometrie, Westbahnhofstrasse 55, D-72070 Tübingen, Germany

Received 12 July 2005; received in revised form 3 November 2005; accepted 3 November 2005

Available online 31 January 2006

KEYWORDS

Plasmodium falciparum;
Survival time;
Distribution;
Quantile–quantile plots;
Neurosyphilis;
Malariatherapy

Summary Survival time data of *Plasmodium falciparum* infections from deliberate infection of human subjects with *P. falciparum* between 1940 and 1963 as a treatment for neurosyphilis in the USA (Georgia) have been used to test the fits of five commonly used parametric distributions for survival times using quantile–quantile plots. Our results suggest that the best fit is obtained from the Gompertz or Weibull distributions. This result has important implications for mathematical modelling of malaria, which has for the past century exclusively assumed that the duration of malaria infections has an exponential distribution. It is desirable to know the correct distribution because its shape profoundly influences the length of monitoring needed in an intervention programme for eliminating or reducing malaria.

© 2005 Royal Society of Tropical Medicine and Hygiene. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Mathematical modelling of malaria has flourished since the days of Ross (1911), who was the first to model the dynamics of malaria transmission. Ross assumed that as far as infection with malaria is concerned, an individual host could be in only one of two states: either susceptible or infected, and that susceptibles have some constant probability per unit time (the infection rate, λ) of becoming infected, and infected individuals have some constant probability per unit time (the recovery rate or the clearance rate, μ) of recovering.

The events of infection and recovery were assumed to be Poisson processes, in that they occur randomly in time within the population. Hence, it follows from the well known property of the Poisson distribution (Cox and Miller, 1965) that the intervals of time between successive arrivals (and also between arrival and departure) follow an exponential distribution with constant rate λ (and μ , respectively). It is likely that recovery rates are dependent on a number of time-related factors, for instance the age of the infection and the development of host immunity. Nevertheless, this assumption of a constant recovery rate is one of the more prevalent assumptions in epidemic theory as a whole (for example, see: Anderson, 1982; Anderson and May, 1991; Bailey, 1957, 1975, 1982). There is extensive literature on mathematical models of malaria with some fundamental departures

* Corresponding author. Tel.: +41 61 284 82 82;
fax: +41 61 284 81 05.

E-mail address: wilson.sama@unibas.ch (W. Sama).

from the view of malaria epidemiology as conceived by Ross (see, for instance, work by Dietz (1988) and the references therein). However, even in models where it is assumed that the duration of an infection depends on the age of the host (an assumption that attempts to incorporate acquired immunity), it is still assumed that the duration at a given age is exponentially distributed (Smith and Vounatsou, 2003).

Clements and Paterson (1981) used a Gompertz distribution for the duration of a malaria vector's life time, i.e. the duration of the infection in the vector, but we are not aware of any published models that have used alternatives to the exponential distribution for the duration of malaria infection in humans, or where this assumption has been tested. There are two major difficulties in collecting empirical data from malaria-endemic areas for use in testing this assumption. When there is an obligation to treat all the infections discovered, this precludes monitoring and limits the possibilities for estimating the duration in non-immune individuals. Individuals in endemic areas are subjected to repeated reinfection and the common diagnostic tool (light microscopy) in the field is not able to differentiate between infections derived at different points in time. PCR-based methods have now demonstrated that individuals indeed harbour multiple infections and such methods are being used to classify malaria infections into different types (Felger et al., 1993, 1999). However, given that even PCR is not 100% sensitive, the continual appearance and disappearance of specific types in consecutive blood samples in longitudinal studies makes it difficult to characterise such types as new or persisting infections.

The duration of infection is most easily estimated from the rate of clearance of parasites from the blood following a single infection. Owing to difficulties in obtaining such data in the field, some studies (Eichner et al., 2001; Molineaux et al., 2001; Paget-McNicol et al., 2002; Recker et al., 2004) have used data (malariatherapy data) from *Plasmodium falciparum* infections deliberately induced as a treatment for neurosyphilis in the mid 20th century in the USA, at a time when there were no antibiotics for the treatment of neurosyphilis (Collins and Jeffery, 1999; Jeffery and Eyles, 1955). Although these data were collected from non-immune individuals in a non-endemic area, they represent an excellent source of data for testing the assumption of an exponential distribution for the survival time.

We studied the fits of five alternative distributions commonly used for survival data as an approximation to the lifetime of malaria infections within the host using malariatherapy data and address the question of the applicability of the results to endemic areas. In the following, we use the word hazard to mean the clearance rate (or recovery rate) of infections.

2. Methods

2.1. Data

Malariatherapy data were collected in the USA (Milledgeville Hospital, Georgia and National Institute of Health Laboratories, Columbia, South Carolina) during 1940–1963, at a time when malariatherapy was a recommended treatment for neurosyphilis (Collins and Jeffery, 1999). Different strains of

P. falciparum were inoculated with either sporozoites (generally through mosquito bite) or infected blood. Microscopic blood examinations were performed almost daily.

Of a total of 334 patients in our database, 157 were from Georgia and 177 were from South Carolina. The average duration of infections was 135.2 days (standard error (SE) 8.8 days) in Georgia patients and 75.4 days (SE 4.2 days) in South Carolina patients, irrespective of whether they received treatment or the time at which treatment was given. Ninety-nine of 157 Georgia patients received treatment, with a total of 540 days when treatment was given, whilst 116 of 177 South Carolina patients received treatment, with a total of 1030 days when treatment was given. This suggests that infections in patients from the Georgia hospital persisted for longer than those in South Carolina and this appears to reflect more treatment in the latter hospital. Hence, it is more likely that the Georgia infections were more similar to untreated natural infections in a typical malaria-endemic setting in Africa. The period for monitoring after the last positive slide also varied, hence so does the confidence that an infection was spontaneously cleared. For this analysis, we consider only patients from the Georgia hospital who did not receive any antimalarial treatment on their last day of positivity of asexual parasitaemia and were followed up for a qualifying period of at least 60 days after their last positive slide. This consists of 54 patients; 29 of them received the Santee-Cooper strain, 23 received the El Limon strain and 2 received the McLendon strain. No substantial differences in the mean (arithmetic) duration of infections were observed by varying the qualifying period. Among the 54 patients, 23 of them received some subcurative treatment before their last positive slide.

2.2. Distributional assumptions

We consider five different distributions commonly used to model survival data: exponential, log-normal, gamma, Weibull and Gompertz. We focus mainly on the fit of these distributions as an approximation to the lifetime of malaria infections within the host using data from malariatherapy patients. For more on the basic properties of the above distributions, for instance hazard rates, quantiles, see the following publications (Evans et al., 2000; Johnson et al., 1995; Klein and Moeschberger, 1997; Wilk and Gnanadesikan, 1968).

A powerful tool for exploring distributional fit to data is by using the graphical technique referred to as the quantile–quantile plot or probability plot (Chambers et al., 1983). The basic idea behind this plot is the following. Suppose that y_1 to y_n are the observed data and that $y_{(1)}$ to $y_{(n)}$ are the values of the data sorted from smallest to largest, so that y_i is the p_i empirical quantile for $p_i = (i - 0.5)/n$. (The $y_{(i)}$ are commonly called the order statistics.) Also suppose $F(y)$ is the cumulative distribution function of the theoretical distribution in question. Now the p quantile of F , where $0 < p < 1$, is a number that we will call $Q_t(p)$, which satisfies $F(Q_t(p)) = p$. In the theoretical quantile–quantile plot, $Q_e(p_i)$ (the p_i empirical quantile, which is equivalent to $y_{(i)}$) is plotted against $Q_t(p_i)$. If the theoretical distribution is a close approximation to the empirical distribution, then the quantiles of the data will closely match the theoretical quantiles

and the points on the plot will fall near a straight line (the null or reference configuration for the plot). However, random fluctuations in any particular data set will cause the points to slightly drift away from the line. Any large or systematic departures from the line should be judged as indicating lack of fit of the distribution to the data. For more discussion of such plots, see Chambers et al. (1983).

To obtain the theoretical quantiles, the parameters of the distributions considered must be estimated. We estimate these parameters by maximum likelihood methods implemented in Fortran 95 (Compaq Visual Fortran Version 6.6; Compaq Computer Corporation, Houston, TX, USA) using the quasi-Newton algorithm (Gill and Murray, 1976) from the NAG FORTRAN library (NAG Fortran Manual, Mark 19; NAG Ltd, Oxford, UK). Confidence intervals were calculated by inverting the observed information matrix (Davison, 2003). A goodness-of-fit test is performed using the Kolmogorov–Smirnov (K-S) test, and the Akaike information criterion (AIC) is used to compare the fits from different distributions.

3. Results

There was no substantial difference between the 23 patients who received subcurative treatment before their last positive slide (average duration of 210.6 days) and the 31 patients who did not (average duration of 212.3 days). The minimum survival time of the *P. falciparum* infections observed for the 54 patients considered was 14 days, whilst the maximum was 417 days. The mean survival time was 211.6 days and the median was 215.5 days. The distribution of the complete data is concentrated around the median, with moderately long tails both to the left and right (Figure 1). The Wilcoxon rank-sum test indicates that the two samples (those who received the Santee-Cooper strain and those who received the El Limon strain) come from the same distribution ($z = -0.193$, $P = 0.85$).

The five parametric distributions considered are shown in Table 1 together with the maximum likelihood estimates, the mean lifetimes of the infections, the AIC values for comparison of the fits from the different distributions, and the K-S goodness-of-fit test.

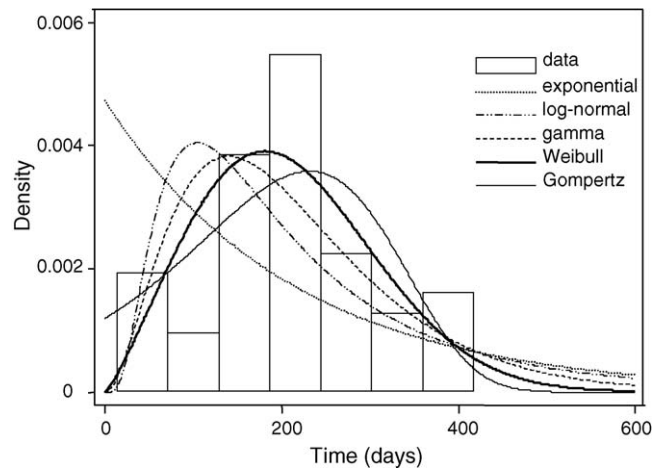


Figure 1 A histogram of the observed data and the probability density functions of exponential, log-normal, gamma, Weibull and Gompertz distributions. The parameters used for the distributions are maximum likelihood estimates from the fits of these distributions to the malariatherapy data.

The quantile–quantile plots suggest that the Gompertz distribution provides the best fit to the data. This is closely followed by the Weibull and gamma distributions. The worst fits are obtained from the exponential and log-normal distributions (Figures 1 and 2). This ranking is also confirmed by the AIC values, whilst the K-S test suggests that the Weibull fit is slightly better than the Gompertz (Table 1). The fits of the distributions depicted in Figure 1 seem to favour more the results from the K-S test. The tails of the probability density functions of the Gompertz and Weibull distributions are shorter than the remaining three distributions considered (Figure 1). Plots of the hazard rates indicate that, for the Gompertz, Weibull and gamma distributions, the hazard increases with the age of the infection, with the rate of increase highest for the Gompertz and lowest for the gamma distribution. The hazard rates for the log-normal peaks and decreases again with increased age of infection (Figure 3).

Table 1 Estimates of parameters and expected lifetimes (with 95% CI in parentheses), the Kolmogorov–Smirnov (K-S) statistics (with P -values in parentheses) and the Akaike information criterion (AIC) for some common parametric distributions for survival times, using malariatherapy data

Probability density function, $f(x)$	Hazard rate, $H(x)$	Estimate of parameters	Estimate of mean lifetime	K-S	AIC
Exponential $\frac{1}{\sigma} e^{-x/\sigma}$	$\frac{1}{\sigma}$	$\sigma = 211.6$ (155.1, 268.0)	211.6 (155.1, 268.0)	0.2924 (<0.001)	688.3
Weibull $\frac{c}{\sigma} \left(\frac{x}{\sigma}\right)^{c-1} e^{-(x/\sigma)^c}$	$\frac{c}{\sigma^c} x^{c-1}$	$c = 2.2$ (1.7, 2.7) $\sigma = 236.6$ (207.0, 266.2)	209.5 (183.0, 236.1)	0.1001 (0.651)	654.4
Gamma $\frac{1}{\sigma \Gamma(\alpha)} \left(\frac{x}{\sigma}\right)^{\alpha-1} e^{-(x/\sigma)}$	$\frac{f(x)}{1-F(x)}$	$\alpha = 3.0$ (1.9, 4.0) $\sigma = 71.6$ (44.0, 99.2)	211.6 (178.7, 244.4)	0.1531 (0.159)	663.8
Log-normal $\frac{1}{\sigma \sqrt{2\pi} x} e^{-((\log x - \zeta)^2 / 2\sigma^2)}$	$\frac{f(x)}{1-F(x)}$	$\zeta = 5.2$ (5.0, 5.4) $\sigma = 0.73$ (0.59, 0.86)	230.4 (179.1, 281.6)	0.1945 (0.034)	681.7
Gompertz $\theta e^{\alpha x} \exp\left(\frac{\theta}{\alpha}(1 - e^{\alpha x})\right)$	$\theta e^{\alpha x}$	$\alpha = 0.0085$ (0.0058, 0.0114) $\theta = 0.0012$ (0.0006, 0.0021)	210.7 (184.2, 237.3)	0.1030 (0.615)	648.8

$F(x)$ is the cumulative distribution function.

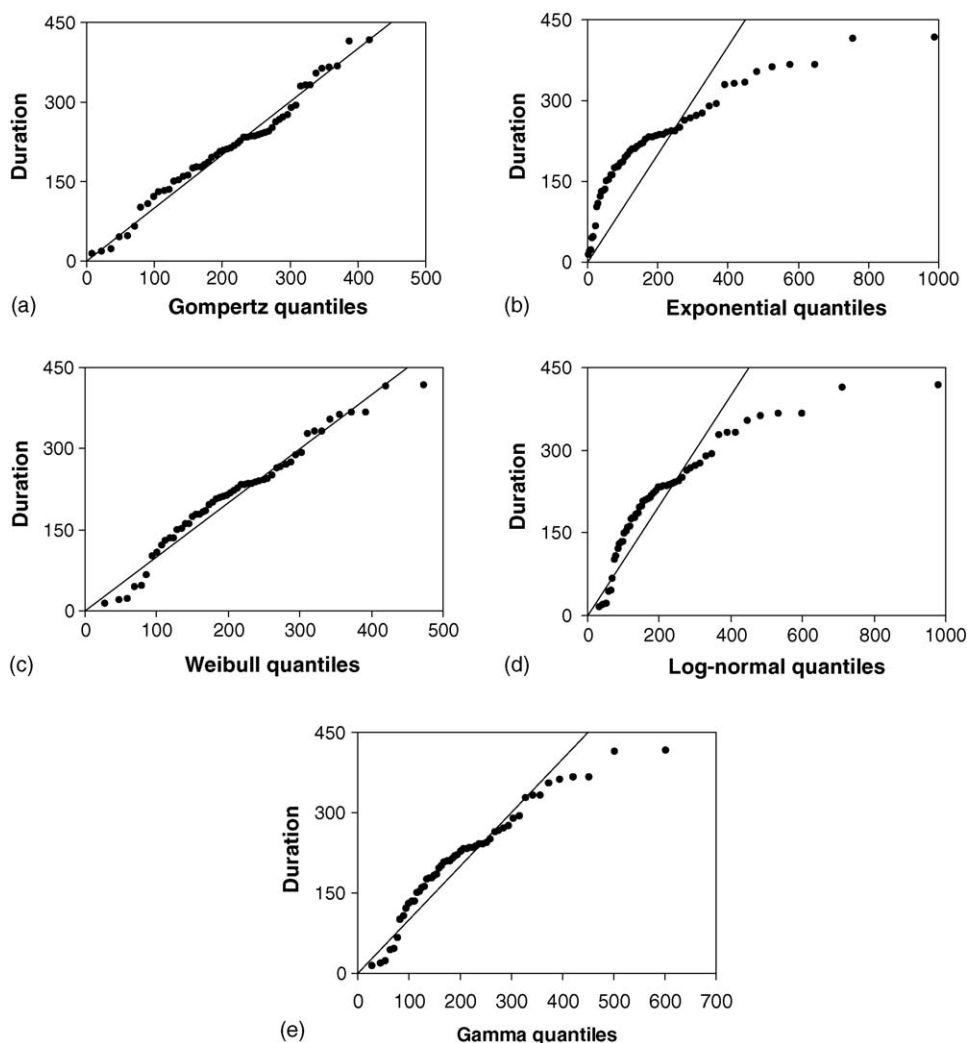


Figure 2 (a–e) Quantile–quantile plots of the malariatherapy data. The straight line represents the reference line.

4. Discussion

We have used survival time data of *P. falciparum* infections from deliberate infection of human subjects with *P. falciparum* between 1940 and 1963 as a treatment for tertiary syphilis in the USA to test the fits of five commonly used parametric distributions for survival times using quantile–quantile plots.

Each quantile–quantile plot compares the empirical distribution of one variable with a theoretical distribution; all other information, in particular the relationship of this variable to others, is ignored. The malariatherapy data thus provide a good source of data to fit the above distributions using this method since the individuals are residing in non-endemic areas where favourable conditions (for example past exposure or acquired immunity) to modify the duration of infection are not available. This, however, raises the question of the applicability of the results to endemic areas. It has been reported that infections in the naïve host mimic the levels of parasitaemia reached by *P. falciparum* infections induced to treat neurosyphilis in the USA (Collins and Jeffery, 1999). This result can therefore be a useful

approximation for naïve individuals in endemic areas. Modifications can then be made to account for past exposure or acquired immunity in older individuals. For instance, if the Weibull distribution is used, then one possible modification to account for acquired immunity will be to fix the shape parameter and to make the scale parameter a suitable function of age.

Our results suggest that the best fit is obtained from the Gompertz or Weibull distributions. A moderate fit is obtained using the gamma distribution. Good fits are not obtained using the log-normal and exponential distributions.

These results have important implications for mathematical modelling of malaria, which has for the past century exclusively assumed that the duration of malaria infections has an exponential distribution. It is important to know the correct distribution because its shape profoundly influences the length of monitoring needed in an intervention programme for eliminating or reducing malaria. The density functions of the Gompertz and Weibull distributions (Figure 1) have shorter tails, indicating that, on average, after a certain period of time since infection the prevalence of residual infections in the population is less than

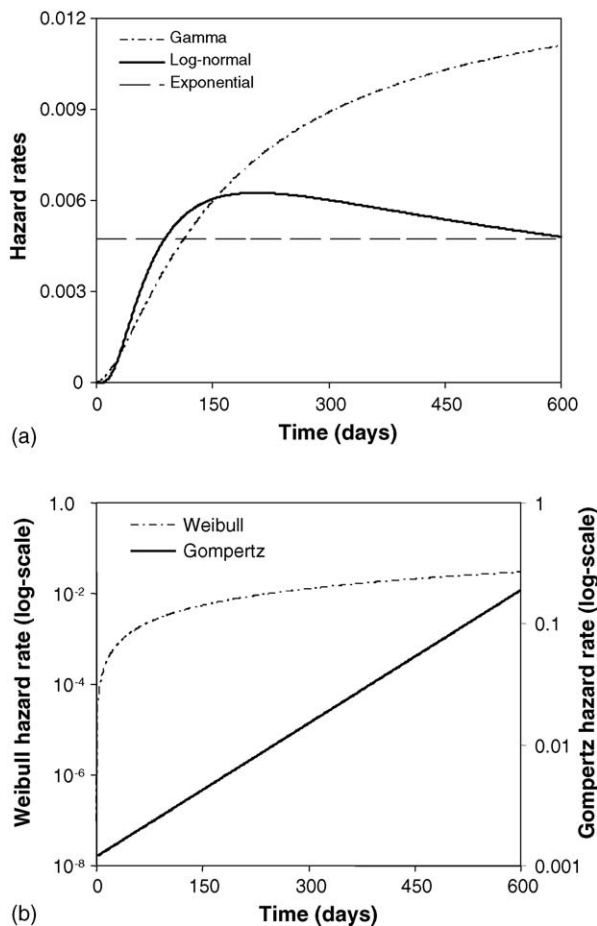


Figure 3 Plots of the hazard rates for (a) the gamma, log-normal and exponential distributions and (b) the Weibull and Gompertz distributions using the malariatheapy data.

suggested by the exponential distribution, and hence that the monitoring period can be shorter if the Gompertz or Weibull distribution is considered. The implications of the dynamics of what will happen if transmission is interrupted can be discussed by looking at the graph of the hazard rates (Figure 3). For instance, if transmission is interrupted at a time of the year when people harbour more new infections (for instance the rainy season) than old ones, then the Gompertz and Weibull distributions suggest that a longer period of monitoring is required after the intervention, whilst for the exponential distribution the hazard does not depend on the age of the infection. The period of monitoring required may also have implications for the optimal period of intervention depending on the kind of intervention.

Collins and Jeffery (1999) report that in patients not receiving antimalarial drug therapy during the primary course of infection, the length of infection for patients infected with the McLendon strain was considerably shorter than in those infected with the El Limon or Santee-Cooper strains. The subset of data used in our analysis mainly comprises infections with one or other of two strains only (El Limon and Santee-Cooper), which had similar durations. It is quite possible that other strains will have different distributions of durations, leading to more variation in the field than was recorded in the malariatheapy patients.

The theory of malaria eradication proposed in the 1950s (Macdonald, 1957) is partly based on the length of time for which eradication campaigns should be carried out to ensure that there is no residual transmission. Most of this theory has assumed that the length of time for which the duration of untreated infections will die away naturally has an exponential distribution. This theory must be reviewed in light of the new evidence from this analysis. It is, however, worth noting that the estimated average duration of 210.7 days obtained in this analysis using the Gompertz distribution agrees very well with Macdonald's (1950) estimate of 200 days (based on the exponential distribution) that was used for the planning of a malaria eradication programme (Macdonald and Göckel, 1964).

The availability of computer-intensive statistical methods makes it feasible to fit distributions with more complicated mathematical forms. We therefore propose the Gompertz and Weibull distributions as alternative distributions to approximate the lifetimes of *P. falciparum* infections.

Conflicts of interest statement

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

Acknowledgements

Wilson Sama is in receipt of a stipend from the Stipendiumkommission of the Amt für Ausbildungsbeiträge of the Canton of Basel and the Swiss National Science Foundation.

References

- Anderson, R.M., 1982. The Population Dynamics of Infectious Diseases. Theory and Applications. Chapman and Hall, New York.
- Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.
- Bailey, N.T.J., 1957. The Mathematical Theory of Epidemics. Charles Griffin and Co. Ltd, London.
- Bailey, N.T.J., 1975. The Mathematical Theory of Infectious Diseases and its Application, second ed. Charles Griffin and Co. Ltd, London.
- Bailey, N.T.J., 1982. The Biomathematics of Malaria. Charles Griffin, London.
- Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., 1983. Graphical Methods for Data Analysis. Chapman and Hall, New York.
- Clements, A.N., Paterson, G.D., 1981. The analysis of mortality and survival rates in wild populations of mosquitoes. *J. Appl. Ecol.* 18, 373–399.
- Collins, W.E., Jeffery, G.M., 1999. A retrospective examination of sporozoite- and trophozoite-induced infections with *Plasmodium falciparum* in patients previously infected with heterologous species of *Plasmodium*: effect on development of parasitologic and clinical immunity. *Am. J. Trop. Med. Hyg.* 61 (Suppl. 1), 36–43.
- Cox, D.R., Miller, H.D., 1965. The Theory of Stochastic Processes. Methuen and Co. Ltd, London.
- Davison, A.C., 2003. Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.
- 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.

- Eichner, M., Diebner, H.H., Molineaux, L., Collins, W.E., Jeffery, G.M., Dietz, K., 2001. Genesis, sequestration and survival of *Plasmodium falciparum* gametocytes: parameter estimates from fitting a model to malariatherapy data. *Trans. R. Soc. Trop. Med. Hyg.* 95, 497–501.
- Evans, M., Hastings, N., Peacock, B., 2000. *Statistical Distributions*, third ed. John Wiley and Sons Inc., New York.
- Felger, I., Tavul, L., Beck, H.P., 1993. *Plasmodium falciparum*: a rapid technique for genotyping the merozoite surface protein 2. *Exp. Parasitol.* 77, 372–375.
- Felger, I., Irion, A., Steiger, S., Beck, H.P., 1999. Genotypes of merozoite surface protein 2 of *Plasmodium falciparum* in Tanzania. *Trans. R. Soc. Trop. Med. Hyg.* 93 (Suppl. 1), 3–9.
- Gill, P.E., Murray, W., 1976. Minimization subject to bounds on the variables. National Physical Library, NPL Report NAC 72.
- Jeffery, G.M., Eyles, D.E., 1955. Infectivity to mosquitoes of *Plasmodium falciparum* as related to gametocyte density and duration of infection. *Am. J. Trop. Med. Hyg.* 4, 781–789.
- Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. *Continuous Univariate Distributions*, second ed. John Wiley and Sons Inc., New York.
- Klein, J.P., Moeschberger, M.L., 1997. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Verlag, New York.
- Macdonald, G., 1950. The analysis of malaria parasite rates in infants. *Trop. Dis. Bull.* 47, 915–938.
- Macdonald, G., 1957. *The Epidemiology and Control of Malaria*. Oxford University Press, London.
- Macdonald, G., Göckel, G.W., 1964. The malaria parasite rate and interruption of transmission. *Bull. World Health Organ.* 31, 365–377.
- Molineaux, L., Diebner, H.H., Eichner, M., Collins, W.E., Jeffery, G.M., Dietz, K., 2001. *Plasmodium falciparum* parasitaemia described by a new mathematical model. *Parasitology* 122, 379–391.
- Paget-McNicol, S., Gatton, M., Hastings, I., Saul, A., 2002. The *Plasmodium falciparum* var gene switching rate, switching mechanism and patterns of parasite recrudescence described by mathematical modelling. *Parasitology* 124, 225–235.
- Recker, M., Nee, S., Bull, P.C., Kinyanjui, S., Marsh, K., Newbold, C., Gupta, S., 2004. Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature* 429, 555–558.
- Ross, R., 1911. *The Prevention of Malaria*, second ed. John Murray, London.
- 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.
- Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for analysis of data. *Biometrika* 55, 1–17.