Statistical Analysis of Within-host Dynamics of *Plasmodium falciparum* Infections

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Wilson Bigina Sama-Titanji
aus Bamenda (Kamerun)

Basel, 2006.

Statistical Analysis of Within-host Dynamics of *Plasmodium falciparum* Infections

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Wilson Bigina Sama-Titanji
aus Bamenda (Kamerun)

Basel, 2006.

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät der Universität Basel auf Antrag von Prof. Dr. Klaus Dietz, Dr. I. Felger, Prof. Dr. H. Becher, Prof. Dr. T.A. Smith.

Basel, den 04 April 2006

Prof. Dr. Hans-Jakob Wirz
Dekan der Philosophisch-
Naturwissenschaftlichen Fakultät

To my loving wife Marion
&
In memory of my beloved mother

Contents

List of Figures

List of Tables

Acknowledgements

It would have been impossible to realise this work without the profound expertise of my supervisor Prof. Dr. Tom Smith. Special thanks are due to Prof. T. Smith for introducing me to this area of research. His patience, scientific counsel and enthusiasm, friendship, encouragement and sense of humour, humble nature and simplicity were my best sources of support. I also received tremendous scientific support from Prof. Dr. Klaus Dietz, Dr. Penelope Vounatsou, which helped to greatly improve the work reported here. I am highly indebted to Prof. K. Dietz for his constant and timely response to my numerous emails and phone calls and for ever being ready for discussion and also Dr. Louis Molineaux for the fruitful discussions we had during their visits to the STI and our meetings at Tübingen.

I wish to express my sincere thanks to the Director of the Swiss Tropical Institute (STI), Prof. Dr. Marcel Tanner, and the Head of Department of Public Health and Epidemiology, Prof. Dr. Mitchell Weiss, for establishing the framework and infrastructure for my research at the Institute.

A substantial part of the data used for the analysis described in this work was generated from research collaboration between the Navrongo Health Research Center in Ghana and the Swiss Tropical Institute in Basel. I wish to thank all the members who worked in both teams especially the principal investigators from both teams: Dr. Seth Owusu-Agyei, Director of Kintampo Health Research Center in Ghana, and Dr. Ingrid Felger of the STI for making the data available and for their scientific counselling and fruitful discussions. I am also thankful to Beatrice Glinz, André Tiaden, Nicole Falk, and Martin Maire for the huge contribution they made in genotyping.

Special thanks also go to Prof. Dr. Heiko Becher from Ruprecht-Karls Universität Heidelberg, who accepted to act as a co-referee in the role of an external expert.

I wish to thank all the members of staff, senior scientists and fellow student colleagues of the STI especially Dr. Armin Gemperli, Abdallah Abouihia, Nicholas Maire, Amanda Ross and Laura Gosoniu for their ever-ready assistance in explaining some statistical concepts and programming software, to Dr. Gerry Killeen for his critical scientific comments, and also to Daniel Anderegg for his assistance in scientific writing. My sincere gratitude to Christine Walliser, Eliane Ghilardi,

Cornelia Naumann, Louise Miedaner, and Margrit Slaoui, for their pleasant manner in sorting out many administrative issues as well as many other practical matters during the course of this work. A special thanks to Christine Walliser for her ever caring attitude and encouragement, and to Nicolas Maire and Manuel Hetzel for translating the summary of this thesis from English to German.

Many thanks to fellow students of the STI for their support, fruitful discussions and the nice times we spent together and also for their friendship and hospitality, in particular Tobias Erlanger, Manuel Hetzel, Monica Daigl, Dr. Collins Ahorlu, Christian Nsanzabana, Dr. Mike Hobbins, Stefanie Granado, Dr. Benjamin Koudou, Clemence Esse, Dr. Giovanna Raso, Elisabetta Peduzzi, Cinthia Acka, Dr. Lucy Ochola, Dr. Sohini Banerjee, Marlies Craig, Dr. Abraham Hodgson, Dr. Shr-Jie Wang, Claudia Sauerborn, Dr. Charles Mayobana, Bonaventure Savadogo, Nafomon Sogoba, Honorati Masanja, Dr. Monica Wymann, Musa Mabaso, Barbara Matthys, Bianca Plüss, Gaby Gehler, Honorati Masanja, Brama Kone, Markus Hilty, Stefan Dongus, Yvonne Geissbuehler, Goujing Yang, Oliver Briet, Peter Steinmann, Rea Tschopp, Josh Yukich.

For the excellent maintenance of computing resources my thanks go to Martin Baumann, Simon Roelly and Dr. Urs Hodel. I have special admiration to Martin Baumann for his dynamism and efficiency in handling multiple tasks. I also wish to express deep appreciation to the STI library team especially Heidi Immler, Mehtap Tosun, Annina Isler, and Fabienne Fust.

Thanks to Prof. Vincent Titanji, Dr. Gideon Ngwa and Dr. Eric Achidi of the University of Buea (Cameroon), and my friends, Abuh Rolland and John Ngonjo in Cameroon for their constant moral support and encouragement.

Basel is now a second home to me thanks to the wonderful friends I have encountered. I wish to express my sincere gratitude to the Makia's family (Divine, Claudia and the kids), Dr. and Mrs Rudin (Peter and Erika), the Pickering's family (Michael, Christine, Sarah, Deborah, Daniel), Evert Bikker, Albert Dreyfuss, Monica Daigl, Tobias Erlanger, Flavia Pizzagalli, Lucia Schönenberger, Sebastien Leuzinger, Anna Koryakina, Sama Junior Doh, Willy Tabot, Christian Nsanzabana, Tanja Grandinetti, Annick Staub, Irena Salc, Divine Yemba, Flavia Trepp, Celestine and Gisela Ebie, Ayuk Moses, the Galabe's family (Charles, Celestine, Sambobga and

Nahletem), Mirella Mahlstein, Babson Ajibade, Joan Gelpcke, Peter Elangwe and to all the members of the Association of Cameroonians in Switzerland for the wonderful times we have spent together and for their constant concern, moral support, and encouragement.

Much needed assistance and encouragement came from members of the entire Titanji family. I am especially grateful to my parents, Francis Sama Titanji, Prof. and Mrs. Titanji (Vincent and Beatrice), and my siblings Judith Foyabo, Grace Nasang, Ernest Duga, Kehmia Nuboyin, Loema Bidjemia, Boghuma Kabisen, Legima Nulla for their advice, encouragement and moral support.

Special thanks to my dear wife and best friend, Marion Enowmba Tabe, whose encouragement has been a tremendous source of support to all activities in my life.

I wish to express my sincere gratitude and thanks to the Stipendiumkommission of the Amt für Ausbildungsbeiträge of the Canton of Basel for sponsoring my Ph.D. studies for 3 years, and the Swiss National Science Foundation for continuing the sponsorship for 1 year. I also wish to thank the Swiss Tropical Institute (STI) for sponsoring my Masters in applied statistics studies at the University of Neuchâtel. The knowledge acquired in this program contributed tremendously to my understanding of the subject matter.

My sincere thanks to the Dissertationenfonds of the University of Basel, and the Swiss Tropical Institute for sponsoring the printing of this thesis.

Above all, I thank the Almighty God who provides me with the day to day strength needed to pursue the difficult and turbulent route of academics.

Summary

Plasmodium falciparum malaria remains one of the world's most important infectious diseases, with at least 300 million people affected worldwide and between 1 and 1.5 million malaria related deaths annually. The eradication program of WHO which was launched in 1955 was motivated by mathematical transmission models. However the evaluation of recent advances in malaria control (using insecticide-impregnated bednets and new therapeutic regimes such as artemesinin derivatives, combination therapy) has largely neglected the effects on transmission, and malaria transmission models have failed to capitalise on enormous advances in computing and molecular parasitology.

Two important factors in models of malaria transmission are the extent of superinfection and the length of time for which clones of malaria parasites persist in the partially immune host. These determine to a large extent the likely effects of vaccines, of impregnated bed nets, and of residual spraying with insecticides on malaria transmission. The effects of acquired immunity on these quantities are also important, both in understanding transmission and the likely parasitological effects of vaccination. However the estimation of these quantities is difficult because malaria infections are often not detectable in the blood.

As with many other laboratory tests used to detect infectious agents, methods for detecting malaria parasites generally have imperfect sensitivity, especially for light infections. Statistical modeling should take into account the occurrence of false negatives, otherwise naïve estimates will provide misleading information on the transition dynamics of the infection. The deterministic models in literature that made some allowance for imperfect detectability had no good way of estimating its extent

because only light microscopy was available at that time for assessment of malaria parasitaemia in the field.

Advances in molecular typing techniques (for instance the PCR and GeneScan) and computer-intensive statistical methods make it feasible to estimate these quantities from field data. The goal of the present study was therefore to address the following questions:

- What is the duration of untreated malaria infections in endemic areas? How does this vary depending on the age and exposure of the human host?
- How is the incidence of malaria superinfection in endemic areas related to age and exposure?
- What is the detectability of the PCR

We addressed these questions using the following approaches:

Statistical analysis of data from a panel survey comprising 6 two-monthly samples from an age-stratified cohort of 300 individuals in the Kassena-Nankana District (KND) of Northern Ghana (an area holoendemic for *P. falciparum*). The *msp*-2 locus of the parasite was used as a marker locus to track individual parasite clones. PCR-RFLP typing of this locus and GeneScan using a subset of 69 individuals from this cohort provided the genotyping data used for the analysis. We developed and fitted an immigration-death model to this data. The model was fitted using both Maximum Likehood methods (using the maximization algorithm in the NAG Routines implemented via the software Fortran 90) and using Bayesian inference (via MCMC simulation employing the Metropolis algorithm in the software WinBUGS 1.4).

We also analysed data obtained for patterns of infection determined by light microscopy in the Garki project, an intensively monitored experiment in malaria eradication in

Northern Nigeria, carried out in 1971-1977. Similarly, we analysed parasitological data from two other eradication projects in West Papua from 1953-1955 and from the Pare-Taveta scheme in East Africa during 1955-1966. We developed and fitted exponential decay models to these data using WinBUGS 1.4.

In many malaria transmission models, the survival time of infections within the host is assumed to follow an exponential distribution. The last source of data used for our analysis is malariatherapy data from Georgia (U.S.A.) collected during 1940-1963. This data was used to test this commonly used assumption in the literature. We fitted using Maximum likelihood methods, four alternative statistical distributions commonly used for survival data and compare the fits using standard statistical tests.

The main results of our findings were as follows:

Allowing for the fact that many infected people have multiple parasite clones, it was estimated that untreated *Plasmodium falciparum* infections in asymptomatic individuals residing in Navrongo will last for approximately 600 days. This result has implications for evaluating the effect of intervention programs in endemic settings. We conclude that a waiting time of about 2 years is needed to draw conclusions about the effectiveness of intervention programs such as insecticide spraying, treated bednets, and mosquito source reduction.

Using data from PCR-RFLP analysis, we estimated that the rate at which individuals acquire new infections in the Navrongo site is on average 16 per year, while data from the GeneScan technique gave an estimate of 19 new infections per year.

Though it is often reported that children acquire infections more often than adults, we did not find any relationship between the infection rate and age. We could not draw any firm conclusions from the results from our methods regarding the relationship between past exposure and the duration of infection. However some of our results indicate a tendency for the duration of infection to decrease with age, suggesting that as immunity increases, there is a higher tendency to clear infections faster.

The GeneScan technique for analyzing infection dynamics has a better performance than the PCR-RFLP. Using GeneScan, a total of 119 alleles were detected, while using the PCR-RFLP, only 70 alleles were detected using samples from the same 69 individuals.

Using PCR genotyping data of blood samples from the 69 individuals, it was estimated that only 47% of the alleles present in a host is detected in a finger-prick blood sample.

The best fit for the distribution of the survival time was obtained from two distributions namely: the Weibull and the Gompertz distribution as opposed to the exponential distribution which has been the most commonly used distribution. This suggests that duration of *Plasmodium falciparum* may also depend on the age of the infection. The results obtained here indicate that the older the infections, the faster it will be cleared. These results also have important implications for models of malaria transmission and for planning intervention programs. For instance if an intervention program is carried out at a time of the year when people harbour a lot of new infections, it will require a longer waiting period to evaluate the effect of this program. However the data used to obtain this result was obtained from naïve individuals in non-endemic settings. We did not test this result on data from endemic areas. It is therefore recommended that these alternative

distributions should be tested using data from endemic areas and the fits compared with that from the exponential distribution.

Zusammenfassung

Plasmodium falciparum Malaria ist nach wie vor eine der wichtigsten Infektionskrankheiten. Weltweit sind mindestens 300 Millionen Menschen betroffen und jedes Jahr sterben zwischen 1 und 1.5 Millionen Menschen an der Krankheit. Das 1955 initiierte Ausrottungsprogramm der WHO war mit mathematischen Übertragungsmodellen motiviert worden. Bei der Evaluierung jüngster Fortschritte in der Malaria-Kontrolle (mittels Insektizid-behandelter Moskitonetze und neuer medikamentöser Behandlungen, wie z.B. Artemisinin-Derivate) wurde jedoch deren Einfluss auf die Malaria-Übertragung grösstenteils vernachlässigt. Auch haben mathematische Modelle der Malaria-Übertragung bisher nicht von den enormen Fortschritten in der Computer-Technik und der molekularen Parasitologie profitieren können.

Die Häufigkeit von Superinfektionen (die gleichzeitige Gegenwart mehrerer Infektionen in einem Wirt) und die Dauer einer Infektion in einem semi-immunen Wirt sind wichtige Faktoren in Übertragungs-Modellen. Sie beeinflussen die Voraussage des Effekts von Malaria-Impfstoffen, Insektizid-behandelten Moskitonetzen und der Moskito-Kontrolle durch Besprühen von Wänden mit Insektiziden auf die Malaria-Übertragung in grossem Masse. Der Einfluss der Semi-Immunität auf diese Faktoren ist ebenfalls wichtig für das Verständnis der Malaria-Übertragung und des Effekts von Impfkampagnen auf die Parasitendichte. Die Schätzung dieser Grössen wird jedoch durch den Umstand erschwert, dass Malariainfektionen im Blut oft nicht nachweisbar sind.

Wie viele andere Labortests zum Nachweis von Infektionserregern sind auch die Methoden zum Nachweis von Malariaparasiten nicht vollständig sensitiv. Dies trifft vor

allem auf Infektionen mit niederen Parasitendichten zu. Statistische Modelle, welche das Auftreten falsch negativer Ergebnisse nicht berücksichtigen, führen zu falschen Aussagen über die Übertragungsynamik einer Infektion. Einige publizierte deterministische Modelle berücksichtigen die unvollständige Sensitivität der Diagnose. Diese Modelle können aber auf keine verlässlichen Daten zurückgreifen um die Sensitivität der Diagnose zu schätzen, da zum Zeitpunkt der Publikation nur die Lichtmikroskopie zur Diagnose zur Verfügung stand.

Fortschritte im Bereich der molekularen Genotypisierung (zum Bespiel PCR und GeneScan) und rechenintensive statistische Methoden ermöglichen heute die Schätzung dieser Grössen anhand von Daten aus epidemiologischen Studien. Die vorliegende Arbeit befasst sich darum mit den folgenden Fragen:

- Wie lange dauern unbehandelte Malaria-Infektionen in Malaria-endemischen Gebieten? Welchen Einfluss haben Alter und Exposistion des menschlischen Wirts darauf?
- In welcher Beziehung steht die Inzidenz von Superinfektion in Malaria-endemischen Gebieten zu Alter und Exposition des Wirts?
- Wie gross ist die Sensitivität der PCR?

Wir nutzten die folgende Ansätze zum Studium dieser Fragen:

Eine statistische Analyse einer longitudinalen Studie im Kassena-Nankana District (KND) im Norden Ghanas (ein Gebiet, in welchem *P. falciparum* holoendemisch ist), welche Blut-Proben von 300 Individuen aus verschiedenen Altersklassen umfasste. Die Daten wurden in 6 Erhebungen im Abstand von 2 Monaten gesammelt. Die verschiedenen Parasiten-Klone wurden mit Hilfe des *msp*-2 Marker-Lokus unterschieden.

Die Genotypisierung erfolgte mittels PCR-RFLP und GeneScan in einer Auswahl von 69 Individuen dieser Kohorte. Wir entwickelten ein Immigration-Death Modell und schätzen die verwendeten Parameter anhand der Felddaten mittels Maximum-Likelihood Methoden (implementiert in FORTRAN 90 mit Hilfe der NAG Programm-Bibliothek) sowie Bayesian Inference (MCMC Simulation mit Hilfe des Metropolis Algorithmus in WinBUGS 1.4)

Des weiteren analysierten wir Daten, welche im Rahmen des Garki Projekts zur Studie von Infektions-Mustern und mit Hilfe von Lichtmikroskopen zur Diagnose erhoben wurden. Das Garki Projekt zur Ausrottung der Malaria wurde von 1971-1977 im Norden Nigerias durchgeführt. Gleichzeitig nutzten wir Daten von zwei Studien in West Papua (1953-1955) und Pare-Taveta, Ostafrika (1955-1966). Wir entwickelten und optimierten auf exponentiellem Zerfall basierende Modelle mit Hilfe von WinBUGS 1.4.

Viele Modelle der Malaria-Übertragung nehmen an, dass die Dauer individueller Infektionen im menschlichen Wirt einer Exponential-Verteilung folgt. Wir verwendeten einen Datensatz aus einer Malariatherapie-Studie in Georgia (USA) von 1940-1963, um diese häufig gemachte Annahme zu validieren. Wir optimierten vier verschiedene Modelle, denen verschiedene häufig gebrauchte Verteilungen für Survival-Daten zugrunde lagen, mit Hilfe von Maximum-Likelihood Methoden und verglichen die Resultate mit etablierten statistischen Tests.

Der folgende Abschnitt fasst die wichtigsten Resultate zusammen.

Unter Berücksichtigung der Tatsache, dass viele infizierte Personen gleichzeitig mehrere Parasiten-Klone beherbergen, wurde die Dauer einer unbehandelten *Plasmodium falciparum* Malariainfektion in Navrongo auf ca. 600 Tage geschätzt. Dieses Resultat hat

Auswirkungen auf die Auswertung des Effekts von Interventionsprogrammen in endemischen Regionen. Wir folgern daraus, dass eine Wartezeit von ca. zwei Jahren notwendig ist, bevor eine abschliessende Beurteilung einer Kampagne (z.B. Insektizide, behandelte Moskitonetze, Moskito-Brutplatz-Reduktion Kontrolle möglich ist.

Anhand der PCR-RFLP Daten schätzten wir die Rate, mit welcher Individuen in Navrongo neu infiziert warden, auf 16 Infektionen pro Jahr. Die Schätzung mittels Daten aus der GeneScan Analyse ergab einen Wert von 19 neuen Infektionen pro Jahr.

Wir fanden keinen Zusammenhang zwischen Alter und Infektionsrate. Dies steht im Widerspruch zu vielen publizierten Studien, die über höhere Infektionsraten bei Kindern als bei Erwachsenen berichten. Des weiteren fanden wir keinen eindeutigen Zusammenhang zwischen Infektionsdauer und vorangegangener Exposition. Gewisse Resultate deuten jedoch auf einen Abnahme der Infektionsdauer mit zunehmendem Alter hin. Das liesse auf eine schnellere Beseitigung der Parasiten bei Individuen mit einem höheren Mass an erworbener Immunität schliessen.

Die Zuverlässigkeit der GeneScan Methode bei der Analyse der Infektionsdynamik war grösser als die der PCR-RFLP Methode. Bei der Analyse von 69 Blut-Proben wurden mit GeneScan insgesamt 119, mit PCR-RFLP nur 70 Allele gefunden.

Aufgrund der PCR-Genotypisierung der Blutproben von 69 Individuen schätzten wir, dass nur 47% aller in einem Wirt vorhandenen Allele in einer mittels Fingerstich abgenommenen Blutprobe nachgewiesen werden können.

Die Infektionszeiten konnten am besten mit einer der folgenden beiden Verteilungen angenähert werden: Die Weibull- und die Gompertz-Verteilung, im Gegensatz zur in

diesem Kontext meist verwendeten Exponential-Verteilung. Dies legt einen Zusammenhang zwischen der bisherigen und der verbleibenden Dauer einer Infektion nahe. Je älter eine Infektion ist, desto schneller wird sich der Wirt von ihr befreien. Diese Resultate haben auch wichtige Konsequenzen für Modelle der Malaria-Übertragung und für die Planung von Kampagnen zur Malaria-Bekämpfung. So muss beispielsweise die minimale Wartezeit vor der Evaluation einer Kampagne der Jahreszeit während der Kampagne angepasst werden. Wenn die Kampagne zu einer Zeit mit vielen Infektionen pro Wirt stattfand, muss die Wartezeit verlängert werden. Allerdings stammen die Daten, die diesem Resultat zugrunde liegen, von Malaria nicht-immunen Individuen aus nicht-endemischen Gebieten. Sie wurden nicht mit Daten aus Malaria-endemischen Gebieten validiert. Wir empfehlen deshalb, die auf alternativen Verteilungen beruhenden Modelle auf Daten aus Malaria-endemischen Gebieten anzuwenden und die Resultate mit den Voraussagen der auf Exponential-Verteilungen beruhenden Modellen zu vergleichen.

CHAPTER 1

Introduction: Biology and Epidemiology of Malaria

1.1 Malaria life cycle

Malaria is a vector-borne infectious disease caused by protozoan parasites of the genus *Plasmodium.* There are four malaria parasite species in humans, namely *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale*.

Plasmodium parasites undergo many stages of development, and their complete life cycle occurs in both humans and mosquitoes. The parasites are transmitted to humans by female mosquitoes of the genus *Anopheles*. About 60 of the 390 species of *Anopheles* mosquito transmit the malaria parasite. Of these, only a dozen species are important in the transmission of malaria worldwide. Usually just one or two species play a role in malaria transmission in a particular region where the disease is prevalent.

The life cycle of the parasite is depicted in figure 1. Malaria transmission begins when a female mosquito bites a human already infected with the malaria parasite. The mosquito ingests blood containing immature male and female gametes (sex cells) of the malaria parasite. Inside the mosquito's stomach, the gametes quickly mature. A male gamete fuses with a female gamete to produce a cell known as a zygote. The zygote enters the wall of the mosquito's gut and develops into an oocyst. The oocyst multiplies to produce

thousands of cells known as sporozoites. The sporozoites leave the wall of the gut and migrate to the mosquito's salivary glands. The mosquito phase of the malaria parasite's life cycle is normally completed in 10 to 14 days. This development process occurs more slowly in areas with cooler temperatures. Sporozoite development of *Plasmodium falciparum* is slowed particularly by low temperatures, preventing transmission of this parasite in temperate climates except during summer.

When the infected mosquito bites another human, sporozoites in the mosquito's saliva transfer to the blood of the human. Sporozoites travel in the blood to the liver. In liver cells over the course of one to two weeks, the sporozoites divide repeatedly to form 30,000 to 40,000 merozoites. The merozoites leave the liver to enter the bloodstream, where they invade red blood cells. Inside these blood cells, the merozoites multiply rapidly until they force the red cells to burst, releasing into the bloodstream a new generation of merozoites that go on to infect other red blood cells. Some merozoites divide to form gametocytes, immature male and female gametes. If another mosquito bites the human and ingests these gametocytes, the life cycle of the malaria parasite begins again.

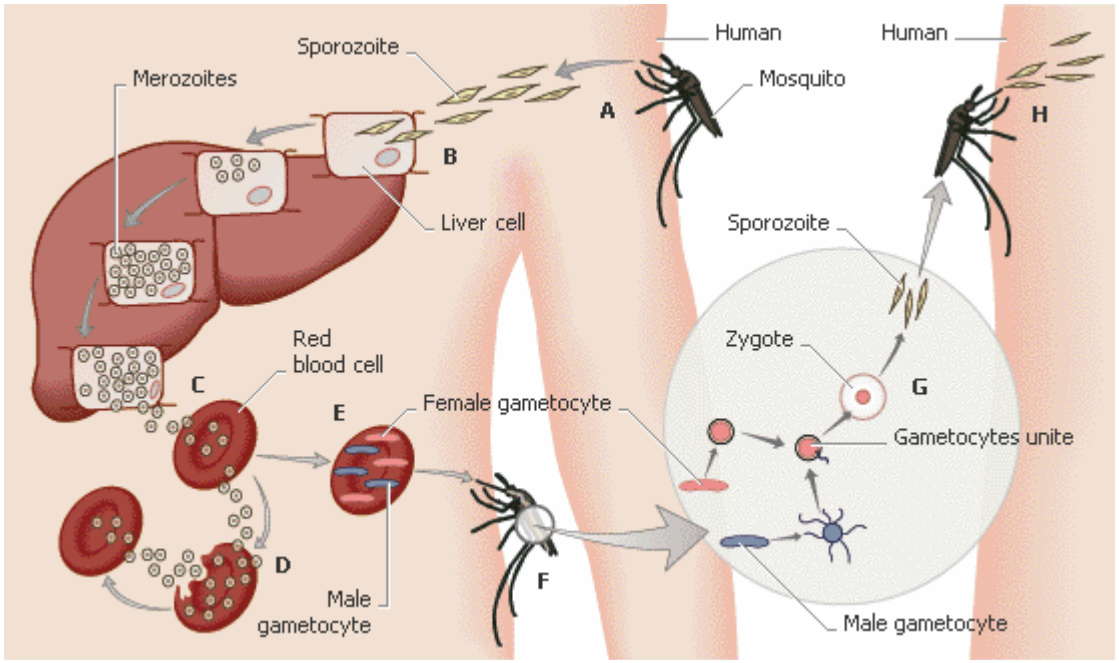


Figure 1. Life cycle of malaria parasite. (Source: http://encarta.msn.com).
(A) Mosquito infected with the malaria parasite bites human, passing cells called sporozoites into the human's bloodstream. (B) Sporozoites travel to the liver. Each sporozoite undergoes asexual reproduction, in which its nucleus splits to form two new cells, called merozoites. (C) Merozoites enter the bloodstream and infect red blood cells. (D) In red blood cells, merozoites grow and divide to produce more merozoites, eventually causing the red blood cells to rupture. Some of the newly released merozoites go on to infect other red blood cells. (E) Some merozoites develop into sex cells known as male and female gametocytes. (F) Another mosquito bites the infected human, ingesting the gametocytes. (G) In the mosquito's stomach, the gametocytes mature. Male and female gametocytes undergo sexual reproduction, uniting to form a zygote. The zygote multiplies to form sporozoites, which travel to the mosquito's salivary glands. (H) If this mosquito bites another human, the cycle begins again.

1.2 Distribution of the malaria parasite

Malaria mostly occurs today in tropical and subtropical countries, particularly sub-Saharan Africa and Southeast Asia. According to the World Health Organization, malaria is prevalent in over 100 countries. Each year more than 300 million cases of malaria are diagnosed, and more than 1.5 million die of the disease (WHO, 1999a, 1999b). *Plasmodium falciparum* is the most common species in tropical areas and is transmitted primarily during the rainy season. This species is the most dangerous, accounting for half of all clinical cases of malaria and 90 percent of deaths from the disease. *Plasmodium*

vivax is the most widely distributed parasite, existing in temperate as well as tropical climates. *Plasmodium malariae* can also be found in temperate and tropical climates but is less common than *Plasmodium vivax*. *Plasmodium ovale* is a relatively rare parasite, restricted to tropical climates (Gilles and Warrell, 1993).

1.3 Symptoms

The main symptoms that characterizes malaria are intermittent fever outbreaks that develops when merozoites invade and destroy red blood cells. Upon initial infection with the malaria parasite, the episodes of fever frequently last 12 hours and usually leave an individual exhausted and bedridden. Repeated infections with the malaria parasite can lead to severe anemia, a decrease in the concentration of red blood cells in the bloodstream.

The pattern of intermittent fever and other symptoms in malaria varies depending on which species of *Plasmodium* is responsible for the infection. Infections caused by *Plasmodium falciparum*, *Plasmodium vivax*, and *Plasmodium ovale* typically produce fever approximately every 48 hours. Infections caused by *Plasmodium malariae* produce fever every 72 hours.

Infections caused by *Plasmodium falciparum* are marked by their severity and high fatality rate. This type of malaria can also cause severe headaches, convulsions, and delirium. The infection sometimes develops into cerebral malaria, in which red blood cells infected with parasites attach to tiny blood vessels in the brain, causing inflammation and blocking the flow of blood and oxygen. In *Plasmodium vivax* and *Plasmodium ovale* infections, some merozoites can remain dormant in the liver for three

months to five years. These merozoites periodically enter the bloodstream, triggering malaria relapses (Gilles and Warrell, 1993).

1.4 Treatment and Diagnosis

Malaria is difficult to diagnose based on symptoms alone. This is because the intermittent fever and other symptoms can be quite variable and could be caused by other illnesses. A diagnosis of malaria is usually made by examining a sample of the patient's blood under the microscope to detect malaria parasites in red blood cells. Parasites can be difficult to detect in the early stages of malaria, in cases of chronic infections, or in *Plasmodium falciparum* infections because often in these cases, not many parasites are present. Recent advances have made it possible to detect proteins or genetic material of *Plasmodium* parasites in a patient's blood.

There are three main groups of antimalarial drugs, namely (i) aryl aminoachohols compounds (for example Chloroquine, Mephloquine), (ii) antifols (e.g. pyrimethamine) and (iii) artemisinin compounds (artemether, artesunate) (Ridley, 2002). With the exception of the artemisinins, *P. falciparum* has developed resistance to all existing drug classes (Simon *et al*., 1988; White, 1992, 1999a; Trape, 2001). To prevent or delay the emergence and spread of resistance, combination therapy, employing two compounds with unrelated mechanisms of action, is increasingly promoted (Peters, 1990; White 1999b; Hastings *et al*., 2002)

1.5 Immunity

After repeated infections, people who live in regions where malaria is prevalent develop a limited immunity to the disease. This partial protection does not prevent people from developing malaria again, but does protect them against the most serious effects of the infection.

Most of the deaths and severe illnesses caused by malaria occur in infants, children, and pregnant women (Bloland *et al*., 1996; Breman, 2001). Infants and children are vulnerable because they have had fewer infections and have not yet built up immunity to the parasite. Pregnant women are more susceptible to malaria because the immune system is somewhat suppressed during pregnancy. In addition, the malaria parasite uses a specific molecule to attach to the tiny blood vessels of the placenta, the tissue that nourishes the fetus and links it to the mother. After exposure to this molecule during her first pregnancy, a woman's immune system learns to recognize and fight against the molecule. This phenomenon makes a woman particularly vulnerable to malaria during her first pregnancy, and somewhat less susceptible during subsequent pregnancies (Mcgregor, 1984; Steketee *et al*., 2001). The malaria infection of the mother is a major reason for abortion and stillbirth and reduces the survival chances of a newborn (Bouvier *et al*., 1997).

1.6 Prevention and Control

Malaria can be prevented by two strategies: eliminating existing infections that serve as a source of transmission, or eliminating people's exposure to mosquitoes. Eliminating the source of infection requires aggressive treatment of people who have malaria to cure

these infections, as well as continuous surveillance to diagnose and treat new cases promptly. This approach is not practical in the developing nations of Africa and Southeast Asia, where malaria is prevalent and governments cannot afford expensive surveillance and treatment programs.

Eliminating exposure to mosquitoes, the second strategy, can be accomplished by permanently destroying bodies of stagnant water where mosquitoes lay their eggs; treating such habitats with insecticides to kill mosquito larvae; fogging or spraying insecticides to kill adult mosquitoes; or using mosquito netting or protective clothing to prevent contact with mosquitoes.

In 1950 the World Health Organization adopted an indoor spraying program with the goal of eradicating malaria worldwide within eight years (WHO, 1991; Gilles and Warrell, 1993). However, budget considerations limited preliminary research, and the program did not take into account the complex differences in the patterns of malaria transmission in different parts of the world. The eradication program was very successful in some countries, particularly in Europe, North America and North Africa, but in other countries, it did not lead to a significant or sustained reduction of malaria cases.

By 1969 it had become clear that eradicating malaria altogether was out of reach, and WHO shifted its focus to malaria control (WHO, 1995). In many countries, the primary means of preventing malaria is the use of insecticide-treated bed nets (Lengeler, 2004). Recent research has shown that these nets are one of the most effective malaria prevention strategies available, but even their modest cost is beyond the means of many families in developing nations.

The resurgence of malaria and the widespread problems of drug and insecticide resistance have focused increasing attention on the need for a malaria vaccine. Developing such a vaccine has been difficult because the malaria parasite has hundreds of different strategies for evading the human immune system. Progress has also been slow because the malaria parasite is difficult to raise in the laboratory and study, since it must live inside the cells of another organism. Despite these hurdles, scientists have developed several possible vaccines that are now being tested in humans.

1.7 Infection dynamics in the human population

Mathematical models have played a role in understanding epidemiology and targeting interventions since the days of Ross (1911), who was the first to model the dynamics of malaria transmission. Modeling malaria transmission though presents a number of challenges additional to those of modeling pathogens for which there is absolute refractoriness to infection. The phenomenon of superinfection, which arises as a consequence of this lack of refractoriness, was introduced into malaria modeling by Macdonald (1950b, 1957). A further element of Macdonald`s theory was his equation for the basic reproduction number Macdonald(1952), which is inversely proportional to the recovery rate from infection, r, or equivalently, proportional to the average duration of infection ($1/r$). Using data from the Garki project (Molineaux and Gramiccia, 1980), a number of different mathematical models were subsequently used to estimate both r and the force of infection, h, (also important as a determinant of the impact of a preventative intervention). These include the mathematical model proposed by Dietz *et al.* (1974) which Nedelman (1984, 1985) analysed in detail.

Much of this literature, focusing on the appropriate mathematical formulation of the queueing theory implicit in Macdonald's writings was summarised by Molineaux *et al.* (1988). In the absence of molecular typing data, however, there were only limited possibilities for testing the different models. Few recent studies have attempted to estimate *r* in natural populations, and there is also a dearth of information on the effect of naturally acquired immunity, or super-infection on either *h* or *r*. It has been suggested that vaccines should aim to emulate natural immunity (Alonso *et al*., 1995), and most malaria vaccine development is based on the hope that the vaccine will affect one or both of *h* and *r,* but there are hardly any data on the relationship between immunological status and these parameters.

In studying the dynamics of malaria transmission in the human population, one needs to understand what an individual infection is, but must also consider as well what happens when many such infections occur together in the human population. A basic challenge is the measurement of the rates of acquisition of infections and of their duration in endemic populations. Another challenge is how to analyse superinfection in endemic areas, and how to analyse parasite dynamics while allowing for infections which temporarily have densities below the limits of detection. Both these problems have become tractable with the use of molecular typing techniques and computer intensive statistical techniques.

1.7.1 Estimation of the force of infection and recovery rate

The acquisition of new infections is measured by the force of infection, *h,* equal to the number of new infections per person at risk in unit time, and in many situations, *h* is the measure of choice for measuring malaria transmission. For instance, in areas of very high endemicity, where the parasite rate (percentage infected) in children may well not be very

informative because it approaches saturation. *h* is also an important determinant of the incidence of disease and comparisons of the force of infection are needed in trials of preventative interventions. Baseline values of *h* are especially valuable when such trials are being planned, since the sample size and the duration of the observation period should be planned taking account of the frequency of new infections.

Clearance of infections is measured by the recovery rate, *r,* equal to the proportion of infections lost in unit time, or equivalently by the mean duration of infection, which is $1/r$. In addition to its importance as a determinant of prevalence, *r* plays a major role in transmission models, as a factor affecting the probability that humans transmit malaria to mosquitoes. Interventions that increase *r* therefore reduce the overall level of transmission and *r* largely determines the time-scale of any effects of an intervention on transmission.

There are considerable problems involved simply in estimating these quantities. If a cohort is studied in the field, then the simplest estimate of *h* is obtained by summing the number of distinct new episodes of patent infection, and dividing this by the total time at risk. The time at risk is equal to the number of days free of parasites during which an infection might have been acquired. If there are several episodes, corresponding to only one true inoculation, then this method will generally lead to a substantial overestimate of *h* since this might be considered as distinct new episodes. Conversely, a simple estimate of *r* is obtained by summing the number of apparent clearance events, (again, this will equal to the number of episodes) and dividing this by the total time at risk (i.e. the number of days of parasitaemia). Just as *h* is overestimated by counting all the different periods of patent infection as separate episodes, so is *r* correspondingly overestimated.

Any error in the determination of *h* or *r* is likely to lead to a compensating error in the determination of the other quantity.

The simplest way of avoiding this problem is to estimate *h* in situations where established infections are not present, and to obtain *r* from people who are protected from new inoculations. These though are atypical scenarios and it is not clear to what extent rates estimated for them should be generalised. It is desirable to be able to estimate *h* and *r* at the same time in people whose infections are not treated, and who are naturally exposed to new inoculations. This problem can only be adequately addressed with the use of molecular typing data, and with models which allow for problems of detectability

Infant Conversion Rate

When transmission rates are very high, children under one year of age can be followed from birth until they become infected, in order to provide an estimate of *h* as originally proposed by Macdonald (1950b). This estimate of *h* is called the infant conversion rate

The infant conversion rate represents a very sensitive measure at levels of transmission intensity high enough for estimation to be reasonably precise. However Macdonald's model does not allow for recovery from infection, and consequently predicts that the prevalence will eventually reach 100%. This is not usually what is observed and Macdonald (1950b) originally explained this by allowing for recoveries in such cases.

Estimation of the force of infection after clearing parasites

A standard method recommended by WHO (1997) for determining *h* is to use a safe and effective schizonticidal drug to clear parasitaemia from a representative cohort of people. The subjects are then bled at intervals, perhaps once a week to once a month (depending

on the level of malaria transmission) and blood films collected for examination for asexual parasites. In this way the distribution of times to re-infection can be estimated and analysed in the same way as infant conversion data to estimate *h*. As a means of estimating *h* in natural settings, this procedure faces a number of difficulties. The presence of pre-existing infections might well affect *h* so the incidence estimated when the infections have been cleared, are not necessarily close to what it would have been in the untreated host population. For the method to work, the drug or combination used must be one against which there is negligible resistance, but should not persist in the circulation for a long period. Sulphadoxine-Pyrimethamine (Fansidar®) has been widely used for this purpose (Alonso *et al*., 1994; Beier *et al*., 1994; Msuya and Curtis, 1991; Stich *et al*., 1994), but has a long half life in the bloodstream, so drug persisting in the circulation directly affects *h*. The treatment of partially-immune people with anti-malarial drugs in the absence of clinical signs or symptoms is not generally recommended and this becomes more difficult to justify if it is necessary to use a drug combination with a high incidence of side effects.

Estimation of the force of infection from serological data

A less invasive strategy for estimating *h* is to use serological data. This strategy proved very useful in studies of the dynamics of viral diseases of childhood (Anderson and May, 1991), where the pathogen generally provokes a well-defined immune response which persists for the lifetime of the host. People of different ages are sampled in cross-sectional surveys, their immune responses determined and sero-positivity is then a reliable measure of cumulative prevalence. Such serological data can be treated as though they came from a birth cohort, and the model proposed by Macdonald (1950b) cab be

fitted to such data using an estimate of the seroconversion rate for *h*. Draper *et al.* (1972) used Indirect Flourescent Antibody (IFA) tests to estimate *h* for *P.falciparum* in this way, but they were forced to omit children under one year of life from their surveys, because maternal antibodies would complicate the picture. This method therefore cannot be used at very high levels of transmission, where many individuals are infected during the first year of life. At lower levels of transmission this problem does not arise because older individuals can be studied. Indeed such serological estimates of *h* are particularly appropriate, where transmission levels are low enough for cumulative prevalence in adults to be a useful measure as in Brazil (Burattini *et al*., 1993).

Estimates of r in the absence of new infections

Corresponding to the strategy of estimating *h* by following cohorts of uninfected people until they become infected, *r* can be estimated by following naturally infected individuals at frequent intervals and recording when the infection disappears. However, this seemingly simple problem is plagued with a series of interrelated difficulties:

(a) When there is an obligation to treat all the infections discovered this precludes follow-up to estimate *r*. This limits the possibilities for estimating *r* in non-immune individuals or during eradication programs.

(b) Parasitaemia cannot be monitored continuously, so a method is needed to estimate the length of time for which the infection persisted after the last positive determination.

(c) Infected people can have a parasite density below the limit of detection (sub-patent parasitaemia) and may subsequently become patent, so a single negative blood sample cannot be taken to imply that the infection has been cleared.

(d) If new infections can occur during the period of follow-up, then these must be allowed for in the estimation of *r*. These difficulties have meant that there have been few attempts to estimate *r* and that the estimates that we have are based on analyses of even fewer datasets.

Since situations in which there is no possibility of new infections are atypical, methods are needed for simultaneously estimating both *r* and *h* where both new infections and recovery are possible. A relatively simple model is the two-compartment model originally proposed by Ross (1916) who obtained values for *r* which were dismissed as unbelievably low by subsequent malariologists (e.g. Macdonald (1950a)), even long before the very dynamic patterns of parasite typing data had been observed. Very different estimates of *h* and *r* are obtained by fitting Ross's model to transition data, as did (Bekessy *et al*., 1976) to the data of the Garki project, and Gazin *et al* (1988) to data from Burkina Faso. More recently it has been applied to populations in Papua New Guinea (Genton *et al*., 1995) and Tanzania (Smith *et al*., 1999a). However Ross's model is not an adequate model for superinfection, and makes no allowance for sub-patent infections.

1.7.2 Superinfection

Evidence for Superinfection

Ross's model assumes that once a person is infected, new inoculations have no epidemiological impact. This can be interpreted as implying that there is absolute resistance against superinfection. This corresponds to the original concept of premunition (Sergent *et al*., 1924). It is also the appropriate formulation on the assumption that the

superinfections are 'wasted', as they would be if all parasites were equivalent. However, there is now extensive evidence that not only does super-infection occur, but that it is an important phenomenon in the epidemiology of malaria.

More recently, PCR-based methodology has made it possible to demonstrate many different parasite genotypes in the same hosts, with up to 8 distinct genotypes at the *msp*-2 locus detectable in single blood samples from endemic areas (Beck *et al*., 1997). Detailed PCR analysis of cloned parasites from two patients have led Druilhe *et al.* (1998) to suggest that even these estimates grossly underestimate the number of clones which can co-infect an individual host. Moreover, analysis of repeated samples from the same host clearly indicate that often only a proportion of the parasite clones present in the host will be detected in any one blood sample (Daubersies *et al*., 1996; Farnert *et al*., 1997).

Estimation of h and r using molecular typing data

Long before this evidence for multiple infections was available, Macdonald had dismissed the idea that existing infections stimulate protection against superinfection. He postulated that "The existence of infection is no barrier to superinfection, so that two or more broods of organisms may flourish side by side, unaltered by the others" Macdonald (1950a).

Macdonald's model and the developments of this (Dietz *et al*., 1974; Fine, 1975; Molineaux *et al*., 1988) were originally evaluated in relation to datasets where the actual number of superinfections could not be assessed. The availability of PCR techniques makes it possible not only to assess the number of co-infections directly, but also to model the dynamics of individual parasite clones. Thus this model provides a basis for

studying the dynamics of infections where the parasites have been typed (Smith *et al*., 1999c). It is likely that it provides a better representation of reality and more direct estimates of both *h* and *r* than models fitted to microscopy data only. This is because the typing provide direct evidence of the extent and persistence of superinfections. However this improvement could be illusory because even using PCR techniques, not all parasite types present are detected in any single blood sample, so the failure to detect a particular parasite type in a blood sample does not mean that it is absent from the host.

1.7.3 Detectability and parasite dynamics

Estimation of detectability of individual parasite clones

Problems of detectability arise in studies of malaria parasite dynamics in at least three distinct ways. Firstly, low overall parasite densities, or the simultaneous sequestration of the bulk of the parasite load can mean that a sample appears negative by microscopy, although the host is actually parasitised. The extent of this problem can be assessed by comparing positivity by microscopy with that determined using PCR (Owusu-Agyei *et al.,* 2002).

A second problem of detectability arises because the same host might be infected with more than one parasite clone of the same genotype. The extent of the bias introduced in this way in estimates of multiplicity can be estimated from the frequencies with which different genotypes occur together (Hill and Babiker, 1995). If a sufficiently high resolution typing system is used, this phenomenon usually only has small effects on the data analysis (Felger *et al*., 1999a).

The third, distinct, way in which parasites may fail to be detected, is when PCR itself

does not detect parasites which are present in the host. This can easily arise in studies using parasite genotyping because the erythrocytic cycle of all the parasites in a given clone can be synchronised, so that even if the host is consistently positive by PCR, specific genotypes appear to come and go.

Note that the two quantities, Q, the sensitivity of microscopy, and q, the detectability of individual parasite genotypes, are different. q is the proportion of the genotypes present in the host which can be identified by PCR in any one blood sample.

1.8 Objectives of this study.

Two important factors in models of malaria transmission are the extent of superinfection and the length of time for which clones of malaria parasites persist in the partially immune host. These determine to a large extent the likely effects of vaccines, of impregnated bed nets, and of residual spraying with insecticides on malaria transmission. The effects of acquired immunity on these quantities are also important, both in understanding transmission and the likely parasitological effects of vaccination. However the estimation of these quantities is difficult because malaria infections are often not detectable in the blood.

The availability of molecular typing data and computer-intensive statistical methods makes it feasible to estimate these quantities from field data. The main questions addressed in the present study are following:

- What is the duration of untreated malaria infections in endemic areas? How does this vary depending on the age and exposure of the human host?

- How is the incidence of malaria superinfection in endemic areas related to age and exposure?
- What is the detectability of individual parasite genotypes and how does it vary with age

The specific objectives are to

- Assess existing methods for estimating the duration of malaria infections using data obtained from optical microscopy. This is discussed in Chapter 2.
- Extend existing models of superinfection by allowing for imperfect detection. This model is fitted to genotyping data obtained by from the PCR-RFLP technique. This is the topic of Chapter 3.
- Develop models to assess the dependence of age and exposure to infection duration, force of infection, and detectability. In addition, to assess the seasonal variation of infection rates (force of infection). This is described in Chapter 4.
- Comparison of PCR-RFLP and Genescan based genotyping methods. Comparison of results of models developed in Chapter 3. This is presented in Chapter 5.
- Determining the distribution for the survival time of *Plasmodium falciparum* infections. This is discussed in Chapter 6.

CHAPTER 2

Estimating the duration of *Plasmodium falciparum* infection from trials of indoor residual spraying

Wilson Sama, Gerry Killeen, & Tom Smith.
Swiss Tropical Institute, Basel, Switzerland.

This paper has been published in the
American Journal of Tropical Medicine and Hygiene 2004; **70(6):**625-634

2.1 Abstract

We reviewed the use of simple mathematical models to estimate the duration of *Plasmodium falciparum* infection after transmission has been interrupted. We then fit an exponential decay model to repeated cross-sectional survey data collected from three historical trials of indoor residual spraying against malaria: one from two contiguous districts in Tanzania-Kenya carried out in 1954, the others in West Papua (1953), and the Garki project in northern Nigeria (1972-3). A cross-sectional analysis of these datasets gave overall estimates of 602 days (95%CI 581 – 625) for the infection duration in Pare Taveta, 734 days (95% CI: 645 – 849) in West Papua and 1329 days (95% CI 1193 – 1499) for Garki. These estimates are much greater than the most widely quoted figures for the duration of untreated *P. falciparum* infections and although these may be exaggerated because some re-infections occurred despite intensive vector control, prevalence was still dropping when all these projects ended. Longitudinal survival analysis of the Garki data gave much shorter estimates of duration (186 days, 95% CI: 181 – 191), but effects of imperfect detection of parasites by microscopy severely bias these estimates. Estimates of infection duration for different age groups showed considerable variation but no general age trend. There was also no clear relationship between malaria endemicity and infection duration. Analyses of successive sampling from the same individuals with parasite typing are needed to obtain more reliable estimates of infection duration in endemic areas. Periods of several years may be required to evaluate long-term effects of interventions on malaria prevalence.

2.2 Introduction

The duration of untreated malaria infections is an important determinant of the level of transmission in endemic areas and determines the time-scale of the effects on prevalence of reductions in malaria transmission. Its reciprocal, the clearance rate of infections (r) is a parameter in many mathematical models of transmission and immunity. Several models, for instance (Dietz *et al.*, 1974; Aron and May, 1982; Aron, 1988), assume that an important effect of natural immunity is to increase r. Despite this importance, few studies have attempted to estimate the duration of *falciparum* infections in natural populations, and there is also a dearth of information on the effects of naturally acquired immunity (or even of age) on infection duration.

Most detailed studies of duration of *falciparum* parasitaemia, refer to malaria infections deliberately used for treatment of syphilis (therapeutic malaria) and report average infection durations of 200-300 days (James *et al.*, 1932, 1936; Eyles and Young, 1951; Jeffrey and Eyles, 1954; Ciuca *et al.*, 1955; Molineaux *et al.*, 2001). Data such as these convinced most malariologists that untreated infections would generally persist for periods of this order, though occasionally *P. falciparum* infections are reported in returned tourists and immigrants from endemic areas whose last exposure was much further in the past.

Most of these data deal with induced malaria in non-immune subjects, and while these may be applicable in areas of low endemicity subject to epidemics, this does not necessarily reflect the duration in endemic areas where people are repeatedly re-infected. The most widely quoted figure, of 200 days, for the total duration of infection is that derived by Macdonald (1950b), who analysed weekly parasitaemia data recorded for a small group of Puerto Ricans (Earle *et al.*, 1938). However reassessment of the original dataset 25 years later led to the conclusion that

infections with *P. falciparum* might still be patent some 30 months after the original infection and possibly longer (Earle, 1962).

It is difficult to see how, in the absence of parasite typing data, duration of infection could be reliably estimated from field studies in areas with ongoing re-infection. However the decay in the parasite rate when transmission is interrupted can be used to estimate the average duration of infection. A seminal paper in this field was that of Macdonald and Göckel (1964). Using cross-sectional data from a number of attempts at eradication, they fitted a simple model of constant clearance rate to the parasite prevalence, P:

$$\frac{dP}{dt} = -rP \qquad (1)$$

with solution:

$$r = \frac{1}{t}\log\left(\frac{P_0}{P}\right) \qquad (2)$$

where P_0 is the prevalence at time $t = 0$, immediately prior to the interruption of transmission, and claimed that the results were broadly consistent with a duration of 200 days. Here log refers to the natural logarithm and P to the prevalence at time t.

In this paper we make use of this model to estimate the total duration of *Plasmodium falciparum* infection after transmission has been interrupted, but improve on the basic model by allowing for recruitment of new individuals and for changes in age. Modeling the natural duration of infection is considerably complicated when anti-malarial treatment is available. We therefore fitted the models to *P. falciparum* prevalence data from three historical datasets from malaria research projects that preceded the introduction of primary health care providing anti-malarial treatments: the Pare-Taveta scheme (East Africa High Commision, 1960), a pilot project in West Papua (Metselaar, 1957), and the Garki project (Molineaux and Gramiccia (1980). We also test whether the duration of infection depends on the age of the host in these studies.

2.3 Methods

2.3.1 Data sources.

(i) ***The Pare-Taveta Malaria Scheme, Tanzania – Kenya***

In order to find out whether malaria transmission could be interrupted by the adoption of a certain technique of residual spraying, a large-scale trial was conducted in the Taveta sub-district of Kenya and the Pare district of Tanzania (East Africa High Commision, 1960). The first round of residual spraying with Dieldrin and DDT was begun in July 1955 and five further spraying cycles were carried out at approximately 8-month intervals, there being an interval of at least two months between the end of one spray round and the beginning of the next. The ecological contrasts in the different parts of the whole study area led to the division of the study area in five distinct zones: the South Pare swamp villages, the S. Pare roadside villages, the S. Pare mesoendemic area, the Taveta forest, and the North Pare hyperendemic area. Repeated cross-sectional surveys were carried out at different times for each of the different zones, with irregular survey periodicity within each study zone. The total number of surveys carried out was not the same (between 7 and 9) for all the study sites within the study area. The population included in the treated area was about 5300. The number of people examined at each survey was given as intervals, probably indicating the minimum and maximum number within each age group examined during the entire sequence of surveys for each study zone (see Table 2). In this paper we use the midpoints of these intervals to approximate the number of samples examined.

(ii) ***Malaria research in Netherlands New Guinea (West Papua)***

A similar experiment was carried out in West Papua. The experimental area (known as the “Lake Area”) was situated south-west of Hollandia (now called Jayapura) in the basin of the Sentani lake. It consisted of two different parts namely the meso-endemic and the holo-endemic parts.

Spleens were examined and thick Giemsa-stained blood films made from a representative sample of the population in the Lake Area before spraying. The results of the two complete surveys made respectively one (1955) and two years (1956) after the first application of insecticide (DDT) in the Lake Area are summarised by age groups in the original report (Metselaar, 1957). Unlike the Pare Taveta scheme, the number of samples examined for *P. falciparum* was clearly stated in the original report. However in Table 2, we simply show the minimum and maximum number of people in the different age groups sampled in the different surveys. Though the time interval between the surveys was also irregular, the surveys were carried out at the same time in the meso-endemic and holo-endemic parts of the study area.

(iii) ***The Garki Project***

The data used here were collected from the Garki project, an intensively monitored trial in malaria control in Northern Nigeria, carried out in 1969-1976. In contrast to the Pare-Taveta and West Papua datasets, in Garki the individuals were identified and parasitological status could therefore be analysed in the same individuals longitudinally. From April 1972 to October 1973, villages in 3 concentric areas were treated with one of 3 control strategies (A1, A2 and B), described in detail by Molineaux and Gramiccia (1980). Since the objective of this paper is to estimate the duration of *P. falciparum* infections in untreated individuals, we analyse only data collected from the 6 sentinel villages in the intervention area where there was no mass drug administration (area B). Residual indoor spraying with insecticide, propuxur, for three or four rounds, at intervals of about two months, was applied to this area both before and during each of two main transmission seasons (1972, 1973). Eight surveys of the entire population of these villages were carried out prior to the intervention, and a further 8 surveys during the intervention. We consider the data from survey 8 as comprising a baseline for our analysis, and analyse

changes in parasitological status during surveys 9-16. Table 2 shows the number of people in each age group that were examined at baseline. There were very little changes in these numbers in the subsequent surveys.

Table 2. Age distribution of the number of samples examined at different surveys in the three study areas.

(a) Pare-Taveta

Age groups (years)	Number Examined				
	South Pare swamp villages	Taveta forest	S. Pare roadside villages	S. Pare mesoendemic area	N. Pare hyperendemic area
< 1	8–85	15–34	38–68	14–39	43–82
< 2	15–63	16–40	21–66	7–56	50–68
2 – 4	17–136	50–68	71–177	43–101	131–208
5 – 9	36–262	64–105	144–328	69–165	224–349
10 – 14	37–181	39–69	153–270	69–134	202–304
15 – 19	48–121	27–42	41–115	18–49	44–121
20 – 39	41–250	41–92	77–362	26–91	149–265
40+	44–265	36–58	62–199	15–43	57–132

(b) West Papua

Age-groups	Number Examined	
	Holoendemic part	Mesoendemic part
0–2 months	22–42	24–51
3–5 months	17–34	20–35
6–8 months	15–53	23–31
9–11 months	11–21	19–46
1 year	74–122	65–141
2 years	61–95	75–113
3–5 years	179–208	171–318
6–8 years	174–209	271–311
9–11 years	74–157	120–249
12–14 years	76–90	84–106
15–24 years	132–234	95–296
25–34 years	147–257	82–293
35–44 years	97–221	93–224
> 44 years	77–168	66–154

(c) Garki

	Age groups (years)							
	< 1	< 2	2–4	5–9	10–14	15–19	20–39	40+
Number Examined	96	71	182	389	147	114	873	491

Note: The number of people examined for the Pare-Taveta and West Papua studies are given as intervals. For instance, 22-42 for the West Papua study indicates that out of the four surveys carried out in this area, 22 and 42 subjects of the corresponding age group were examined at two of these surveys while at the remaining two surveys, the number of subjects lies in the interval 22-42. For Garki, only the numbers at baseline are given. There was not much variation in these numbers in the 8 subsequent surveys.

2.3.2 Models

The basic model is a differential equation model describing the rate of change of the proportion positive with time. The underlying assumptions are (*i*) that transmission was completely interrupted after the comprehensive application of residual insecticides, (*ii*) that there was no drug treatment so loss of infections were due to spontaneous dying out of parasites or immune response of the host, and (*iii*) that the proportion positive falls in an exponential manner with time. Examining figure 2.1 for example shows little evidence for transmission after the beginning of the intensive vector control measures (which started around the third survey), and the general picture we get from the graphs is prevalence continually dropping after the last pre-spraying surveys, so our assumption is quite reasonable.

Figure 2.1

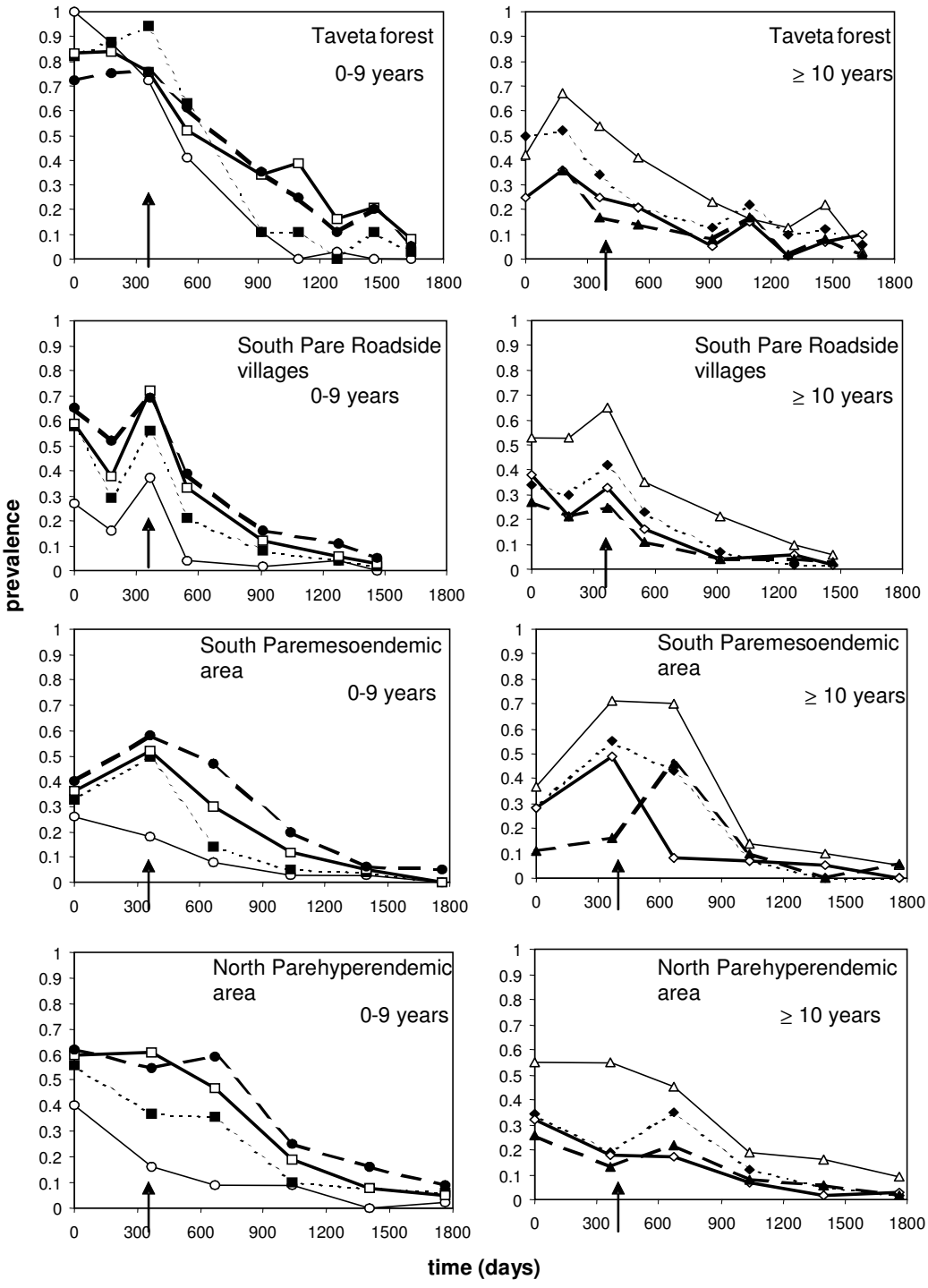

Taveta forest
0-9 years
Taveta forest
≥ 10 years
South Pare Roadside villages
0-9 years
South Pare Roadside villages
≥ 10 years
South Paremesoendemic area
0-9 years
South Paremesoendemic area
≥ 10 years
North Parehyperendemic area
0-9 years
North Parehyperendemic area
≥ 10 years
prevalence
time (days)

Figure 2.1 Observed changes in prevalence of *P. falcparum* parasitaemia with time for different age groups within the Pare-Taveta project. The figure refers to four of the five sites in the project. The corresponding data for the fifth site are shown in figure 2.5. The arrows ↑ indicates the onset of spraying.

0-9 years: 0–11 months (○), 12–23 months (■), 2–4 years (□), 5–9 (●)

≥ 10 years: 10–14 (Δ), 15–19 (♦), 20–39(◊), ≥ 40 (▲)

We assume that there are k age groups at baseline with each individual being assigned into one of age groups i, $1 \le i \le k$. Let $P_{i,t}$ denote the proportion positive at time t among those in age group i at time t_0 (i.e. at baseline). We assume that this proportion is adequately described by the equation

$$\frac{dP_{i,t}}{dt} = -r_i P_{i,t} \qquad (3)$$

where r_i is the clearance rate at time t for individuals in age group i at baseline. If P_{i,t_0} is the proportion positive at baseline then it follows that:

$$P_{i,t} = P_{i,t_0} e^{-r_i (t-t_0)} \qquad (4)$$

The total duration of infection was estimated using two different methods of analysis, namely: repeated cross-sectional and longitudinal survival analysis. Both methods were used to analyse the data from the Garki (area B) study while only the repeated cross-sectional analysis method was applied to the other two datasets.

(a) ***analysis of infection duration from repeated cross-sectional data***

(i) Garki data

In the Garki data, the exact age of each individual was known and the individuals were identified and followed up longitudinally. However, for the purpose of comparison with the other two datasets we start by analysing this dataset as if they were collected from unlinked cross-sectional surveys, and assign individuals to different age groups using the age groupings in the Pare-Taveta study. We fitted model (4) with separate estimates for P_{i,t_0} and a common estimate for r_i (i.e. $r_i = r$ for all the age groups). In order to allow for effects of age on infection duration, we also fitted model (4) with separate estimates of P_{i,t_0} and of r_i for each age group and to test for a linear trend in the effect of age on r we fitted the model with r_i in (4) substituted with the following age dependent term:

$$r_i = r_0 + r_1 a_{i,0} \qquad (5)$$

where $a_{i,0}$ is the mid-age of age group i at baseline.

To allow for random variation, we assume a binomial error function for the parasite prevalence, i.e.

$$X_{i,t} \sim Bin(n_i(t), P_{i,t}), \qquad (6)$$

where $X_{i,t}$ is number of positive samples in age group i at time t and $n_i(t)$ is the total number of samples examined in age group i at time t.

(ii) Pare-Taveta and West Papua data

We take into consideration the fact that the population sampled in each age group varies for each survey by attempting to capture in our model the proportions of the population within each age group at a later time that were in the different age groups at baseline. For this we make the additional assumption that the age distributions are approximately uniform. That is, the number of individuals within each age group is proportional to the width of the age group. We make use of this assumption in deriving equation 7 below.

First, we give a brief explanation, in non-mathematical terms, of the scenario that the equations developed below attempt to capture. Consider, for example, 3 distinct cohorts (or age groups) at baseline, i.e., at time t_0 (see fig. 2.2).

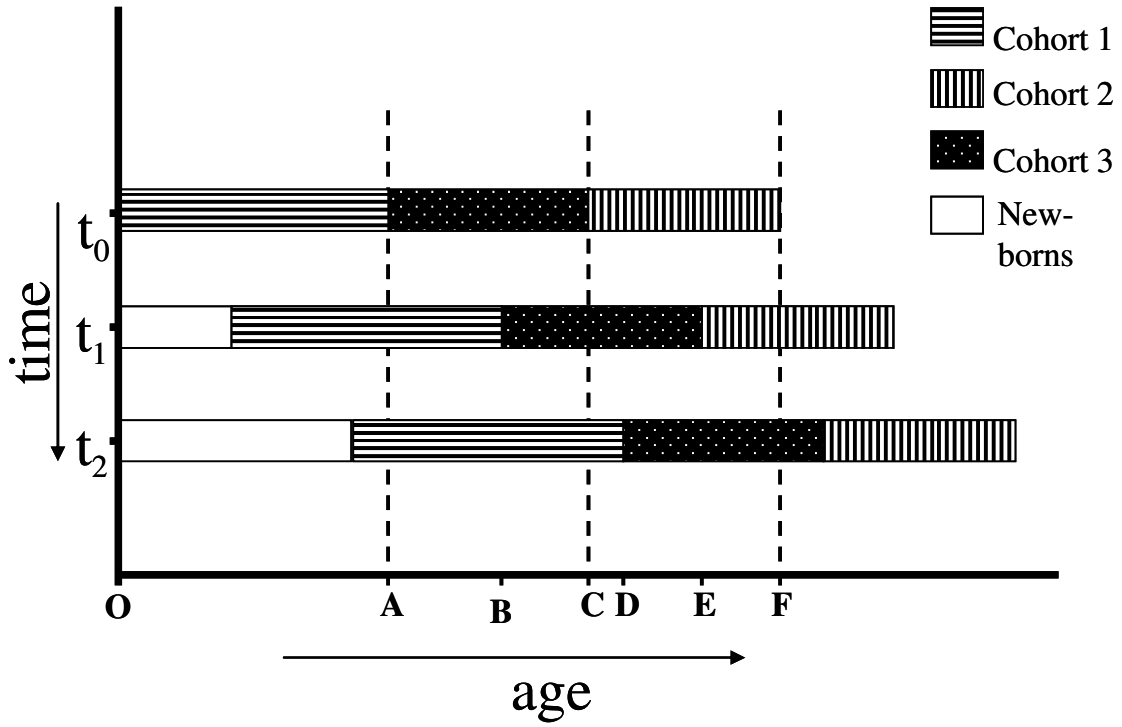


Figure 2.2. Schematic representation of the movement of cohorts from one age group to another between survey rounds at different times (t).

We assume that the prevalence in each age group falls in an exponential manner with time (given by equation 4), with the possibility of a different rate of fall for the different age groups. Since different cohorts were studied at different surveys, we wish to estimate for example, at a later time (say t_1, fig. 2.2) the proportion of a newly recruited cohort 2 who were in cohort 1 at baseline (t_0) (given by equation 7). Looking at figure 2.2 again this proportion is given by the segment AB. We estimate this proportion and take it back to its original group at baseline. It is

clear that when the time interval between the baseline and any subsequent survey is long enough, then some individuals within a specific cohort at baseline will move to two or more subsequent cohorts. From the example given by fig. 2.2, we find that at time-point t_2, a certain proportion (AC) of cohort 1 has moved to cohort 2 while another proportion (CD) has moved to cohort 3. We sum up these proportions, take them back to their original age groups (i.e. the age group they belonged to at baseline) and apply the corresponding exponential fall that was assumed for this group to start with. This is summarised by equation 8.

Let L_i and U_i denote respectively the lower and upper limits of age category *i*. Then the probability that any individual, *x*, will belong to age group *j* at time *t* given that he was in age group *i* at time t_0, (or equivalently the proportion of age group *j* at time *t* that were in age group *i* at time t_0) is given by:

$$\Pr(x \in j, t \mid x \in i, t_0) = \max\left[\frac{\min(U_i + t - t_0, U_j) - \max(L_i + t - t_0, L_j)}{U_j - L_j}, 0\right] \qquad (7)$$

and, it follows from this and equation (4) that the expected value of the prevalence in age group *j* at time *t*, $E(P_{j,t})$, is:

$$E(P_{j,t}) = \sum_i \Pr(x \in j, t \mid x \in i, t_0) P_{i,t} \qquad (8)$$

which is a function of the unknown parameters r_i, P_{i,t_0}. We report results obtained by assuming a binomial error function for the parasite prevalence in both areas, i.e.,

$$P_{j,t} \sim Bin(n_i(t), E(P_{j,t})), \qquad (9)$$

The estimates for the *r*'s were obtained in a similar manner as described in section (a) above. In addition, analyses were carried out separately for different sites within the study areas to test for effects of malaria endemicity on infection duration.

(b) ***Longitudinal survival analysis of infection duration***

Secondly, we fitted an exponential model to the survival times of the infections in the Garki dataset. We studied only changes in the parasitological status of individuals who were present and positive at baseline (8th survey) until the survey when they were either negative or absent. We assumed that the observations for each individual were independent of each other (i.e. an individual positive at two consecutive surveys was treated as two separate counts and so on) and we estimated the proportion (P) infected at time t+1, (I_{t+1}), conditional on being infected at time t, (I_t), as follows:

$$\Pr(I_{t+1} \mid I_t) = e^{-rt} \qquad (10)$$

We also took into account random variation by assuming a binomial error function as follows

$$X_{+,i} \sim Bin(n_{+,i}, P) \qquad (11)$$

where $X_{+,i}$ equals the total number positive at all surveys j ($8 \le j < i$) and positive at survey i, while $n_{+,i}$ equals the total number positive at all surveys j ($8 \le j < i$) and present at survey i, ($9 \le i \le 16$).

The model parameters were estimated using WinBUGS version 1.3 (Spiegelhalter *et al*., 2000), and assuming gamma priors for the r's. The quoted results are based on samples of 29,500 values from the posterior densities, following a burn-in of 30,000 iterations.

2.4 Results

The model for repeated cross-sectional data gave good fits to the data (see for example figure 2.3). Overall estimates for the total duration of infection across all age groups of 602 days were (95%CI 581 – 625) for Pare-Taveta, 734 days (95% CI: 645 – 849) for West Papua and 1329 days (95% CI 1193 – 1499) for Garki (area B).

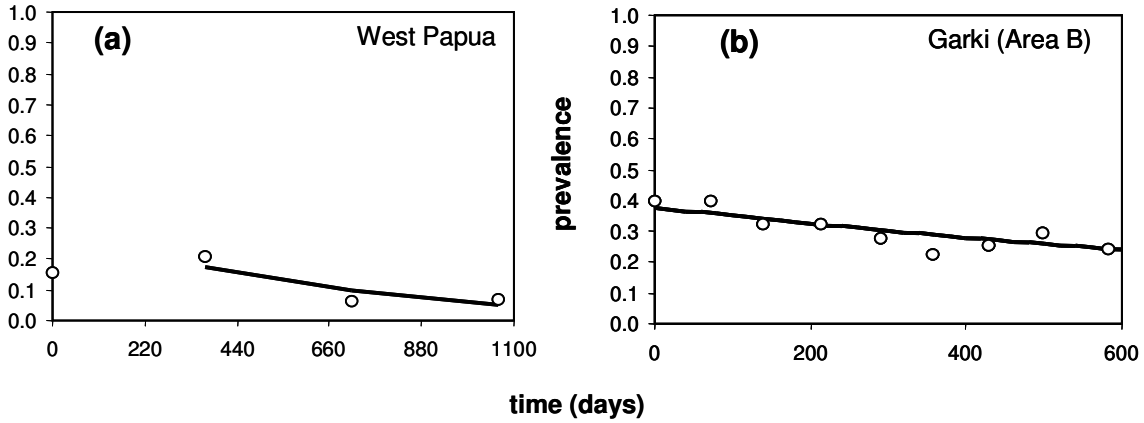


Figure 2.3. Observed and predicted prevalence curves for the pooled data in the West Papua study and the Garki (area B) study using the repeated cross sectional analysis. Open circles indicate observed values while full lines indicate predicted values. **(a)** predicted prevalence curve for Garki obtained by using equation 4. **(b)** predicted prevalence curve for West Papua obtained by using equation 8.

Prevalence appeared to fall faster in the younger age groups than the older ones (Figure 2.1), however this was due to the fact that newborns recruited at the subsequent surveys came in at a time when the insecticide spraying already had a massive effect in reducing transmission. Extension of the models to allow variation between age groups in the estimates of the duration of infection in each of the study areas did not indicate any clear age trend. The age pattern for the estimates in the West Papua was especially noisy, probably due to relatively small size of this dataset (only four surveys were conducted in this trial). However the clearance rate increased modestly with age in all 3 areas (r_1 =0.0125 per year (95% CI 0.0075-0.01776) for West Papua; r_1 =0.0043 (0.0021-0.0066) for Pare-Taveta, r_1 =0.0074 (0.0047-0.0103) for Garki).

Figure 2.4 shows the fit of the model to the data and also the predicted prevalence at each time point in the cohort initially present at baseline. This is shown (fig. 2.4) explicitly for the eight different age groups in one of the sites (Pare swamp) in the Pare-Taveta study. There is a clear

difference in the rate of fall in prevalence in the younger age groups (especially, the first age group, 0-11 months) because new-borns were directly recruited into these age groups. The difference between these two curves become less visible as the age groups become older because at subsequent surveys the newborns were not old enough to attain these age groups and hence affect the observed prevalence within them. The model fit to the Garki data was also good (see fig. 2.5).

Figure 2.4

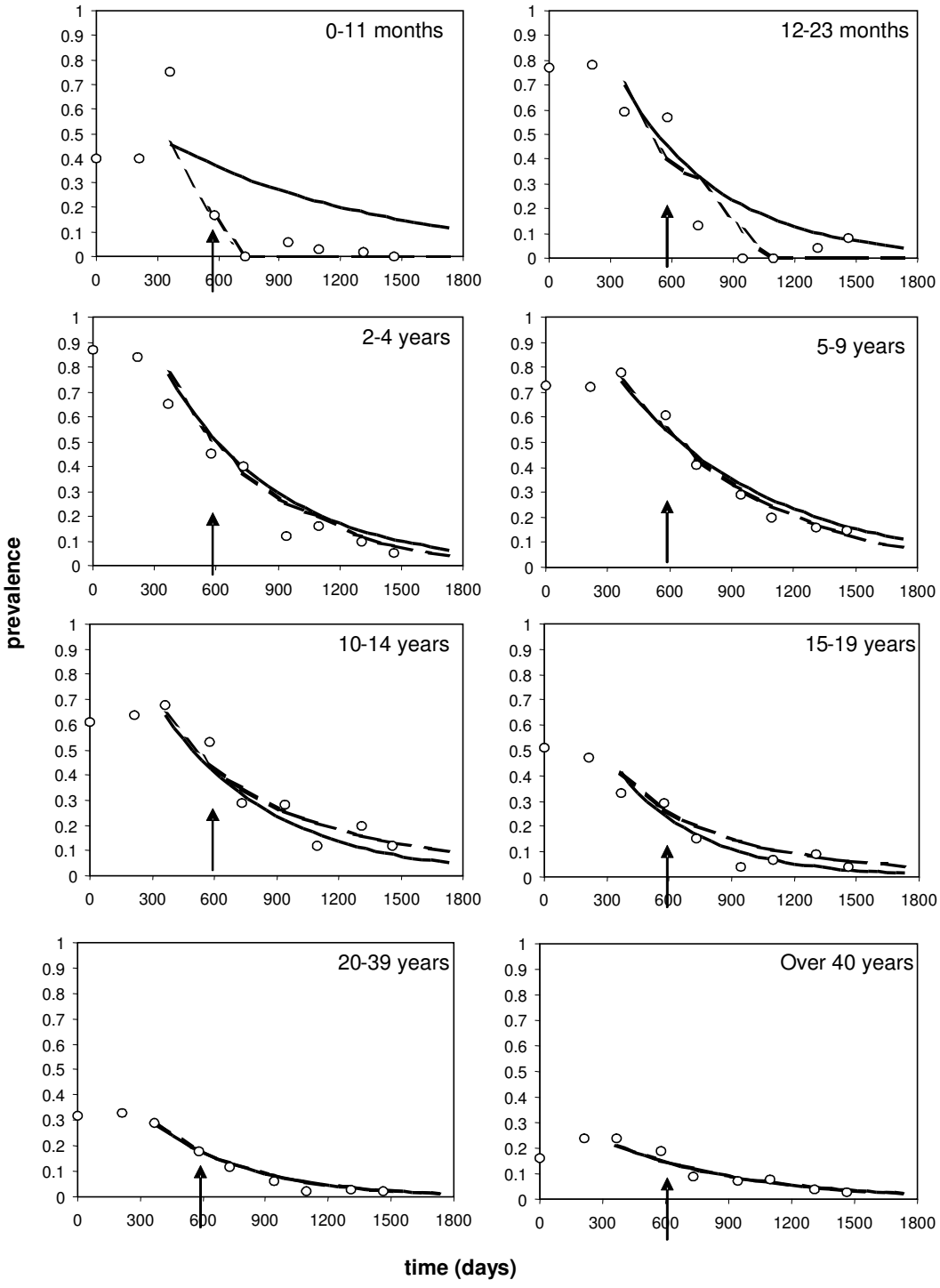

0-11 months
12-23 months
2-4 years
5-9 years
10-14 years
15-19 years
20-39 years
Over 40 years
prevalence
time (days)

Figure 2.4 Observed and predicted prevalence curves for different age groups in South Pare swamp villages (one of the study sites in the Pare-Taveta study): predicted curve is based on repeated cross-sectional modeling analysis. All the data collected from the surveys before the first round of residual spraying are pooled and considered as baseline data in this analysis. The arrow ↑ indicates the onset of spraying.

Open circles indicate observed prevalence

Full line indicates predicted prevalence and this is obtained using equation 4: this predicts what would have happened if only those originally present at baseline had been followed in a longitudinal manner.

Broken lines indicate predicted prevalence using equation 8. This is the model fit to the data and it takes into the fact that individuals are moving from one age category to another over time.

Figure 2.5

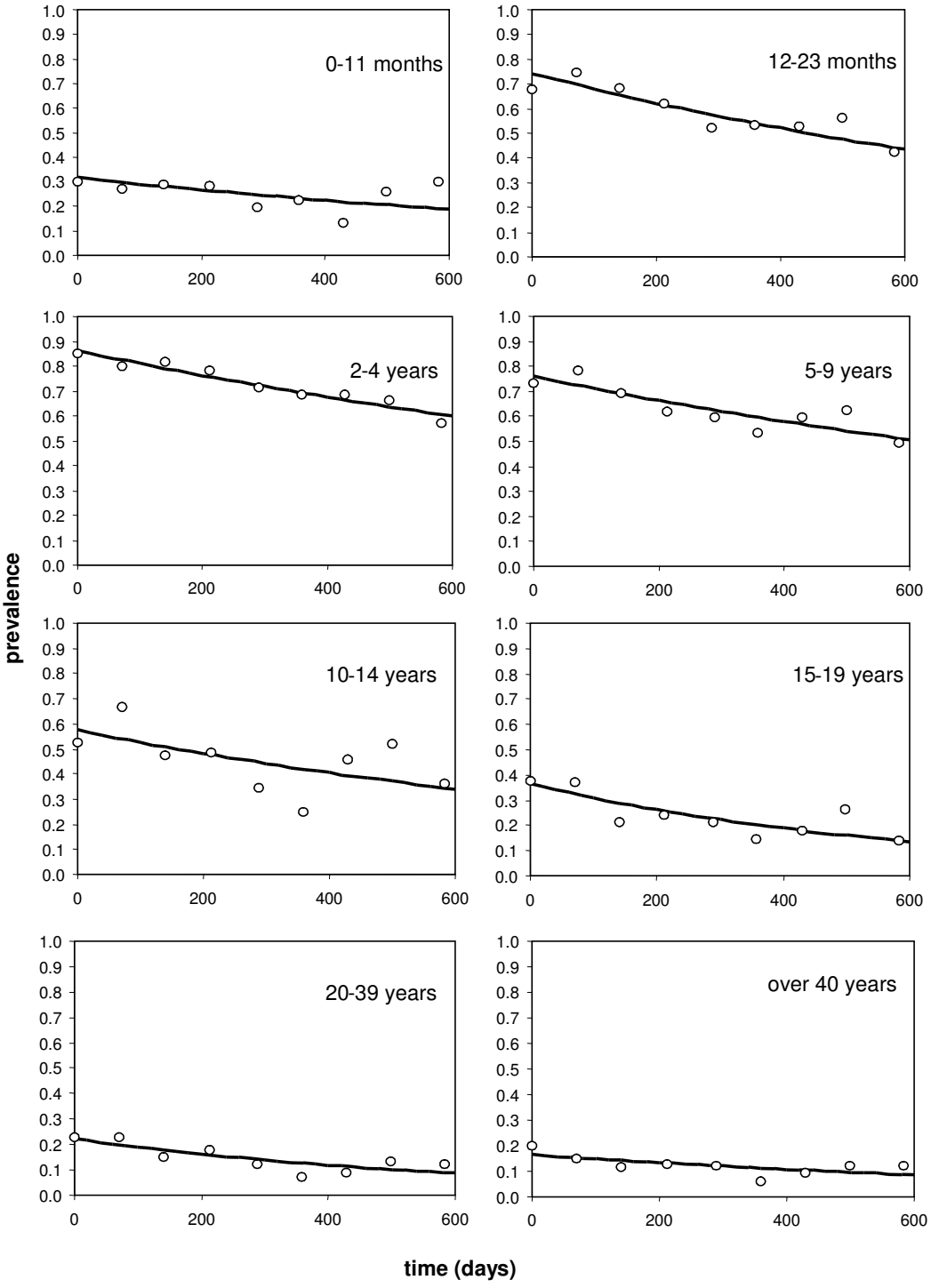

0-11 months
12-23 months
2-4 years
5-9 years
10-14 years
15-19 years
20-39 years
over 40 years
prevalence
time (days)

Figure 2.5 Observed and predicted prevalence curves for 8 different age groups in the Garki (area B). Open circles indicate the observed values while the full line indicate the predicted values. Prediction is based on the repeated cross sectional analysis (using equation 4).

We also obtained separate estimates for the total duration of infection for the different zones in the West Papua and Pare-Taveta studies. The duration of infection for the pooled data in the holo-endemic part of the study area in West Papua was estimated to be 699 days (95%CI 604 – 824) and that for the pooled data in the meso-endemic part was 820 days (95%CI 639 – 1131). The estimates for the pooled data in each of the five different sites in the Pare-Taveta study are shown in fig. 2.6 while the predicted age prevalence curves for the five sites are shown on fig. 2.7. Fig 2.7 suggest that Taveta forest was a highly endemic area while the South Pare mesoendemic area was the least endemic area. A comparison of figs. 2.6 and 2.7 shows that there is no general relationship between infection duration and malaria endemicity. This was confirmed by performing a Spearman's rank correlation analysis, which gave a correlation coefficient, *R,* of 0.30 ($\boldsymbol{P = 0.62}$).

In contrast to the analyses of repeated cross-sectional data, the longitudinal survival analysis of the Garki data gave an overall estimate of 186 days (95% CI: 181 – 191), which is far lower than all the other estimates quoted above but closer to most of the values in the literature.

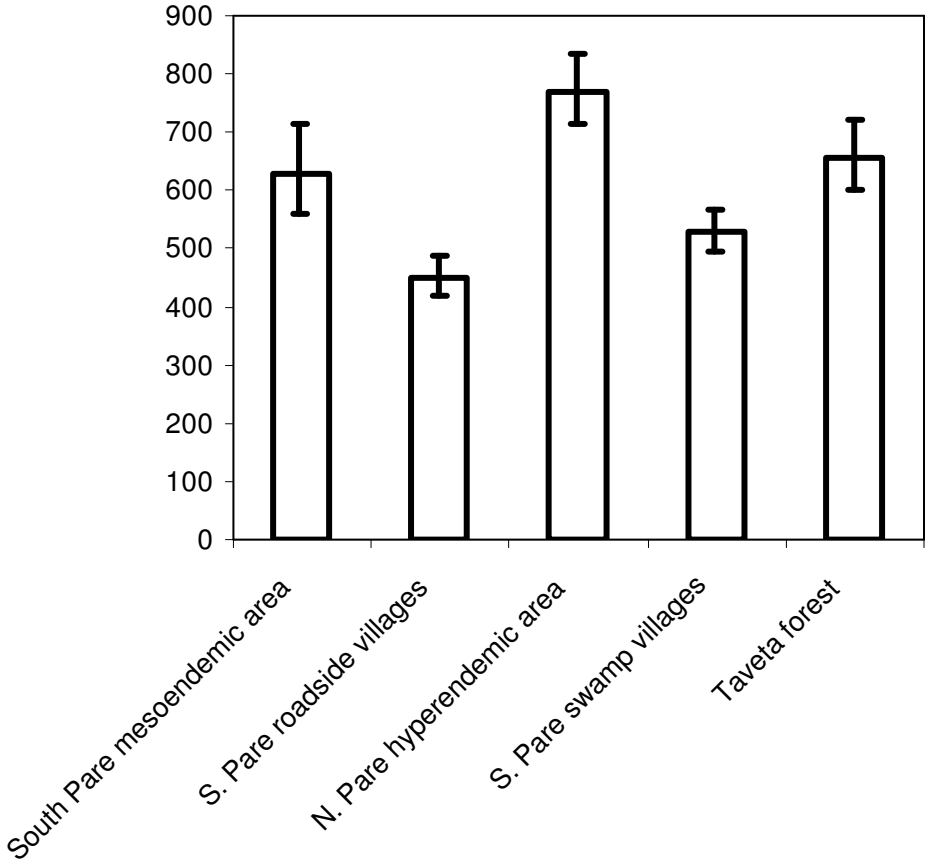


Figure 2.6. Duration of infection for the pooled data in each of the different sites in the Pare-Taveta study: Taveta forest, South Pare swamp villages, North Pare hyper-endemic area, South Pare roadside villages, South Pare meso-endemic area.

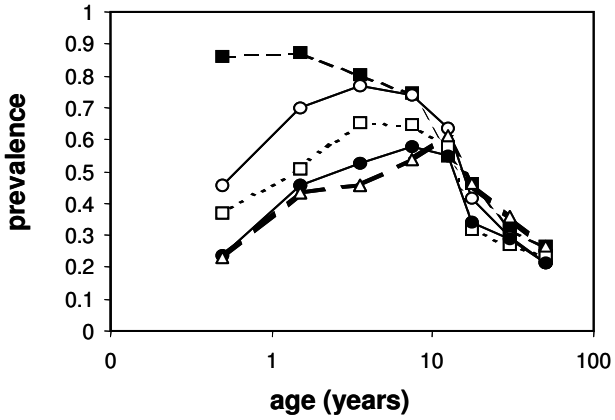


Figure 2.7 Estimated age prevalence curves at baseline using the model given by equation 8 for the different sites in the Pare-Taveta study: Taveta forest (■), South Pare swamp villages (○), North Pare hyper-endemic area (□), South Pare roadside villages (●), South Pare meso-endemic area (Δ).

2.5 Discussion

We have revisited three historical datasets from trials of indoor residual spraying, two of them comprising unlinked repeated cross-sectional surveys, and the third, a longitudinal cohort study, and used an exponential decay model to estimate the duration of infection ($1/r$) for untreated malaria infections in endemic areas.

The overall estimates for the duration of infection from the analysis of the repeated cross-sectional data are similar in the three areas, but are much higher than the most widely quoted values, derived two generations ago from analyses by Macdonald (1950b), Macdonald and Göckel (1964). There was no obvious relationship between infection duration and the endemicity of malaria (figs. 2.6 and 2.7). Our results suggest a waiting time of at least 2 to 3 years is needed before evaluating the effect of interventions that suppress transmission without actively clearing parasites such as insecticide treated nets, indoor residual spraying and mosquito source reduction.

Among the sites for which Macdonald & Göckel (1964) presented curves for the decline in parasite rates when transmission was interrupted, the Pare Taveta and West Papua sites showed the lowest rates of decline. According to Macdonald & Göckel (1964) this was because transmission was not completely interrupted in these sites. However, although there were some infections in children born after the start of spraying, the comprehensive application of insecticide ensured that these were very few in number (figs 2.1 & 2.4). This indicates that post-intervention new infections were rare and cannot account for the persistently high prevalence (fig. 2.1).

A number of studies in endemic areas have reported that infections are of relatively short duration in very young children. For example Walton (1947) found that the average duration of

infection in infants in Freetown, Sierra Leone, was little over 3 months, at a time when there was relatively little transmission there. More recently it was estimated that the duration of infections with parasites belonging to the msp2 FC27 allelic family increased with age using data collected from Tanzanian children aged 6-30 months (Smith *et al*., 1999c; Smith and Vounatsou, 2003). An increase in duration with age during the first two years of life has also been reported in a study with Ghanaian children (Franks *et al*., 2001).

We chose the Pare Taveta, West Papua, and Garki data for re-analysis because all three of them allowed us to analyse the duration of infection by age. The Pare-Taveta and West Papua data found faster initial decrease in prevalence in the youngest age group than in the older ones. However Macdonald & Göckel (1964) suggested that infection duration appeared shorter in young children only because the data had been analysed inappropriately, and that a cohort analysis should have been carried out. This is because straightforward analysis of the decrease in parasite prevalence with time after transmission is interrupted fails to allow for the recruitment of new uninfected individuals (new-borns). Our new analysis allows for this effect, and indeed we find that after this adjustment there is no indication in these datasets that parasites are cleared faster in the youngest children. Indeed, there seems to be a slow decrease in duration with age.

While we agree with Macdonald & Göckel (1964) about the biases due to recruitment of unexposed newborns, all our estimates of clearance rates, except those from the cohort analysis of the Garki data, remain much lower than those reported based on other datasets quoted by Macdonald & Göckel (1964). Part of the reason for this may be that few of the datasets they used are from surveys in endemic communities in the absence of mass treatment. For instance the data they refer to from the eradication of *Anopheles gambiae* from Brazil (Soper and Wilson, 1943) appear to be incidence figures for clinical cases.

The explanation of why the estimates of duration from unlinked data are lower than those from the cohort analysis is that the latter systematically underestimates duration by treating temporarily sub-patent infections as though they had been cleared (Bekessy *et al*., 1976). This can be illustrated by a typical profile of parasite density during follow-up of a malariatherapy patient (fig. 2.8). The patient was inoculated at time A and parasites were cleared at some time, H, after the last day, G, on which parasites were detectable. The true duration of infection is thus H-A. However, longitudinal studies that do not allow for imperfect detectibility, such as that of Macdonald (1950b) and our own longitudinal Garki analysis, give estimates of the duration of parasitaemic episodes (either D-C, or F-E). Estimates of durations from longitudinal data can be even shorter if sampling is frequent because of sequestration during the 48-hour cycle of the parasite. Using longitudinal microscopy data from blood smears collected at very frequent intervals among inhabitants of a single village in Papua New Guinea an estimate of 3-27 days was obtained for the duration of parasitaemic episodes of asymptomatic *P. falciparum* infections (Bruce *et al*., 2000a).

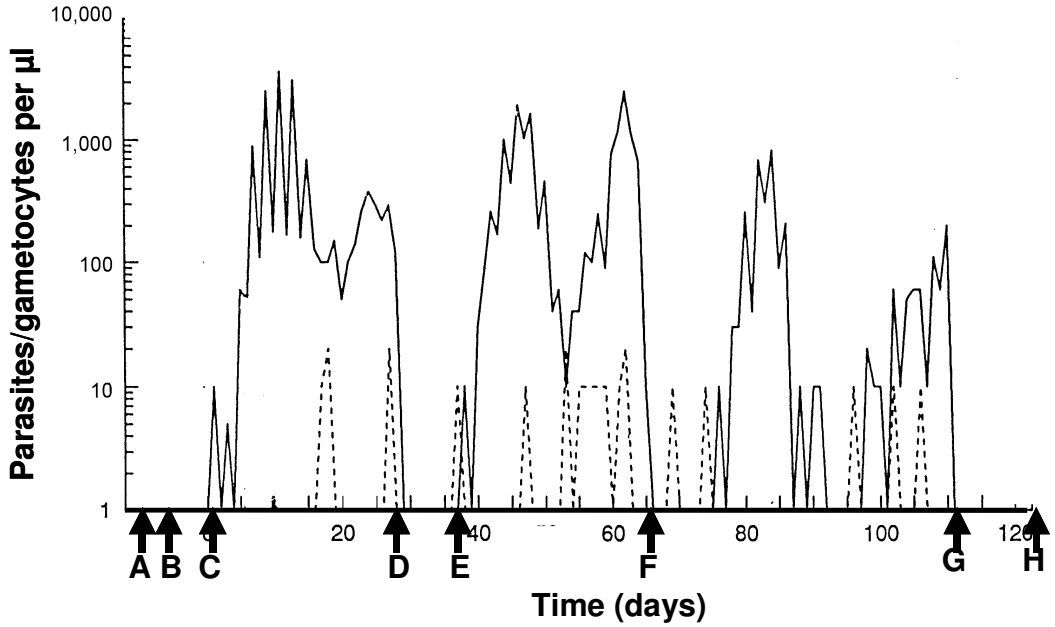


Figure 2.8 Changes in parasite density with time in a malaria-therapy patient: Full lines represent graph for parasite, broken lines represent graph for gametocytes. The letters **A-H** denote specific points in time.

In contrast, the unlinked analysis is not subject to this source of bias. If detectibility is age- and time- independent and the true process follows the dynamics described by equation (1), then the observed process will also follow these dynamics. This implies that the model for unlinked data remains the same whether or not detectibility is taken into account. It is therefore appropriate for estimating the total duration of infection (H-B in fig. 2.8).

Our study thus provides convincing evidence that malaria infections in endemic areas actually persist on average for much longer than Macdonald (1950b) claimed and that it is the cohort analysis that can be biased. It also suggests that duration is not very dependent on exposure in highly endemic communities. However, studies where transmission is interrupted cannot tell us

what is the infection duration when superinfection is occurring. Moreover, very few of the individuals were in the youngest age categories, so we cannot be confident that very young children and naïve individuals have similar duration to people with more exposure. A critical assumption of this paper was that residual transmission was negligible. A key question that could not be adequately addressed using these datasets is that by how much is the duration of infection overestimated due to the fact that re-infection is not accounted for. A graphical display of the raw data however provides convincing evidence of a general exponential decay pattern, therefore indicating that the results are not greatly altered, though it is possible that a better fit could be obtained with more complex decay models. In the Garki project for instance, it was shown (Molineaux and Gramiccia, 1980) that after the spraying program, re-infection did continue at a level of about one sixth its original value. Using the cross sectional analysis for the Garki data, we obtained estimated infection duration of 1329 days suggesting that even if this low level of re-infection was accounted for, the estimated duration of infection will still be much longer than have been thought of in the past.

The average durations that we estimated are also much longer than those seen in malariatherapy patients in whom the whole course of infection was observed (Molineaux *et al*., 2001).

Field-based parasite typing studies now make it possible to study duration of infection in the presence of superinfections. Thus it was found (Bruce *et al*., 2000b) that the mean duration of episodes of positivity (i.e. D-C or F-E in fig. 2.8) for the same (Bruce *et al*., 2000a) *P. falciparum* genotype to be approximately 60 days in Papua New Guinean children. Typing studies have found asymptomatic infections persisting for more than 12 months in Eastern Sudan (Babiker *et al*., 1998). It was also found (Franks *et al*., 2001) that a single parasite genotype of *P.*

falciparum asymptomatic infections could persist for as long as 40 weeks. However few typing studies have attempted to calculate population averages of overall duration.

Recently a model which allows for imperfect detectibility have been proposed (Smith and Vounatsou, 2003) in analysing infection dynamics in longitudinal data where parasites have been typed by PCR and in the presence of new inoculations. These analyses covered only a narrow age range of hosts. Further analyses of longitudinal parasite typing data are needed, covering the whole age distribution, and from areas of different endemicity.

2.6 Acknowledgements

Wilson Sama is in receipt of a stipend from the Stipendiumkommission of the Amt für Ausbildungsbeiträge of the Canton of Basel. The authors would like to thank Klaus Dietz and Louis Molineaux for their helpful comments on an earlier draft of the manuscript; Penelope Vounatsou for assistance with the statistical software. David Bradley, Louis Molineaux and Paulette Rosé helped in the recovery of data from old records.

CHAPTER 3

An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect

Wilson Sama[1], Seth Owusu-Agyei[2], Ingrid Felger[1], Penelope Vounatsou[1], & Tom Smith[1].

[1]Swiss Tropical Institute, Basel, Switzerland.

[2]Kintampo Health Research Centre, Kintampo, Ghana

This paper has been published in

Statistics in Medicine 2005; **24:**3269-3288

The method described in this Chapter is based on work submitted towards the Post-Diploma in Applied Statistics at the University of Neuchâtel, Switzerland. Hence it is included in this thesis just for reference purposes.

3.1 Abstract.

Immigration-death models are proposed to analyse the infection dynamics in longitudinal studies of panels of heavily parasitized human hosts where parasites have been typed at regular intervals by PCR. Immigration refers to the acquisition of a new parasitic genotype, occurring at rate λ, and death refers to the clearance of a parasitic genotype (with rate μ). The models assume that corresponding to each observed process which is the detection or failure to detect a parasitic genotype, is an underlying true process which is hidden as a result of imperfect detection. We consider: (i) a model in which no distinction is made between the different members of the human population, who collectively represent the habitat of the parasites, and (ii) a model that allows for the accrual of infections with age. The models are fitted to a panel data set of malaria genotype of parasites belonging to the *msp2* FC27 and 3D7 allelic families from a study of the dynamics of *Plasmodium falciparum* in Northern Ghana. Maximum likelihood estimates suggest that on average any individual residing in this holo-endemic area will acquire 16 new infections per year (95%CI, 15–18) (defined by their single locus genotypes) and that infection with any of these genotypes lasts on average 152 days (95%CI, 138–169). We estimate that an average of 47% (95%CI, 42–51) of the parasite types present in the host are detected in a finger-prick blood sample. This model provides a basis for analyses of how these quantities vary with the age, and hence the immune status of the host.

3.2 Introduction.

Plasmodium falciparum malaria is the most important eukaryotic parasite of humans, with hundreds of millions of people infected at any one time, causing at least 1 million deaths per annum mainly in sub-Saharan Africa. Humans develop a partially protective immunity against *Plasmodium falciparum* malaria after repeated infections. In high transmission areas, asymptomatic carriers are numerous. Infections can persist for long periods at very low densities, and are difficult to detect by optical microscopy techniques.

Two important factors in models of malaria transmission are the extent of superinfection (infection with a different parasite clone) and the length of time for which clones of malaria parasites persist in the partially immune host. These determine to a large extent the likely effects of vaccines, of impregnated bed nets, and of residual spraying with insecticides on malaria transmission. Mathematical models have played a role in understanding malaria epidemiology and targeting interventions since the days of Ross (1911) who assumed that an infected individual could not be infected again until after complete recovery from the initial infection. Therefore the whole population can be divided into two groups, infected and susceptible.

Ross' assumption implies that an initial infection confers absolute immunity against superinfections (concomitant immunity). It is also the appropriate formulation on the assumption that the superinfections are 'wasted', as they would be if all parasites were equivalent. Macdonald subsequently developed a model which made the converse assumption. He postulated that "The existence of infection is no barrier to superinfection, so that two or more broods of organisms may flourish side by side, the duration of infection due to one being unaltered by others" (Macdonald, 1950a). Unfortunately, the mathematical formulation of this problem proposed by Macdonald and his colleague Irwin (the Macdonald-Irwin model) was erroneous (see for example Fine (1975)) since they imply that individual

infections queue at a single server counter to be terminated. Subsequently a number of alternative formulations were proposed (Molineaux *et al.*, 1988; Dietz, 1988).

The introduction of the polymerase chain reaction (PCR) technique in 1985 (Saiki *et al.*, 1985) increased the sensitivity with which parasites can be detected. PCR-genotyping of a highly polymorphic locus like the merozoite surface protein 2 (*msp2*) locus with subsequent Restriction Fragment Length Polymorphism (RFLP) analysis allows discrimination between different alleles (Felger *et al.*, 1993) and makes it possible to distinguish parasites arising from different infection events and to enumerate concurrent infections. Such methods have demonstrated that multiple concurrent infections[1] are indeed found in many individuals in endemic areas (Beck *et al.*, 1997; Contamin *et al.*, 1996; Felger *et al.*, 1999a; Owusu-Agyei *et al.*, 2002). They allow comparison of consecutive samples and characterisation of newly appearing genotypes in longitudinal studies.

Macdonald's postulate provides a basis for studying the dynamics of infections where the parasites have been typed by PCR. His intended model (Fine, 1975) has been fitted to PCR typing data from the field in a study of Tanzanian infants (Smith *et al.*, 1999c). PCR typing provides direct evidence of the extent and persistence of superinfections (Owusu-Agyei *et al.*, 2002; Barker *et al.*, 1994; Brockman *et al.*, 1999; Snounou *et al.*, 1993). PCR detects more infections than optical microscopy and therefore provides a better representation of reality and more direct estimates of both recovery and inoculation rates. However *P. falciparum* parasites are absent from the peripheral blood for part of the erythrocytic cycle and so even PCR fails to detect some of the parasite genotypes in the host, because there is no template in the blood sample. Studies of the daily dynamics of *Plasmodium falciparum* genotypes have found that individual types continually disappear and then reappear in the peripheral blood (Farnert *et*

[1] By multiple infections we mean the number of distinct genotypes identified using any specific analytical scheme. This represents a minimal estimate of the actual number of clones circulating in an individual, which will generally be higher than the number identified [8]

al., 1997), so the failure to detect a particular type in a blood sample does not mean that it is absent from the host.

The need to allow for such false negatives in the analysis of the transition dynamics of infections has long been recognised (Nedelman, 1985). The models developed and fitted by Dietz *et al.* (1974) and Nedelman (1984) made some allowance for imperfect detectability, but in the absence of PCR data it was not possible to validate the assumptions about the proportion of infections which were sub-patent. Estimates of the duration of infection can be made using PCR data, with allowance for periods when the parasite density was too low for the infection to be detected (Nedelman, 1985; Smith *et al.*, 1999b; Bruce *et al.*, 2000b; Smith and Vounatsou, 2003). Such estimates have been based on either exploratory analyses of recurrence of the same genotype over periods up to two months (Bruce *et al.*, 2000b) or using a Markov model for the recovery process, ignoring re-infection with the same genotype (Smith *et al.*, 1999b).

In this study we propose the use of immigration-death models as postulated by Macdonald (1950a) and later revised by Fine (1975), to analyse the infection dynamics in longitudinal studies where parasites have been typed by PCR. Immigration refers to the acquisition of a new parasitic genotype, occurring at rate λ, and death refers to the clearance of a parasitic genotype (with rate μ). The objective of this analysis is to obtain population averages for the duration of the persistence of any given genotype, and to estimate the infection rates and the probability that the PCR detects each of the genotypes present in the host (the detectability, s). We assume that parasite clearance has a constant intensity μ, equivalent to assuming the duration of infection to follow an exponential distribution with mean $\frac{1}{\mu}$. We also assume that corresponding to each observation is an underlying true state which is hidden because not all the parasite clones present in the host are detected by the PCR. These enable us to compute the likelihood for each of the observations as a function of λ, μ, and s. We use both

maximum likelihood methods, and Markov chain Monte Carlo (MCMC) simulation in order to obtain point and interval estimates for $\frac{1}{\mu}$.

The description of the materials and methods is presented in section 2. The results are presented in section 3. Concluding remarks and possible extensions of the model are discussed in section 4.

3.3 Materials and Methods.

3.3.1 Study site: The data analysed in this work was generated from a malaria study carried out in the Kassena-Nankana District (KND) located in the Upper East Region of Ghana, bounded on the north by part of the border between Ghana and Burkina- Faso. This site is served by the Navrongo Health Research Centre (NHRC), which uses the Navrongo Demographic Surveillance System (NDSS) to monitor the population dynamics of the district. The district has a human population of about 141,000 living in roughly 14000 compounds, mostly dispersed in rural areas. About 20,000 people live in Navrongo town, the administrative capital (Nyarko *et al.*, 2002).

3.3.2 Study design: The study was nested within a panel survey of malaria in which a sample of approximately 300 people provided finger-prick blood samples at two monthly intervals for a period of 10 months (July 2000 – May 2001). The survey was carried out with a target sample size of 256 people, comprising a cluster sample of the population of KND. Sixteen "index" compounds were selected at random from the 14,000 compounds within the 4 different geographic zones within the KND, making use of the NDSS. For each index compound, 2 people in each of the following age categories were selected: <1; 1-2; 3-4; 5-9; 10-19; 20-39; 40-59; 60+. Volunteers were recruited sequentially into each age category until the required number was made up. Blood samples were collected from all the participants on DNA Isocode stix (Owusu-Agyei *et al.*, 2002).

3.3.3 Laboratory methods: An age-stratified sub-sample of 69 individuals with complete data was selected from the compounds that were first visited. Deoxyribonucleic acid (DNA) was extracted from all the samples selected on DNA-Isocode stix irrespective of microscopy results and the *msp2* gene of *P. falciparum* for these samples was genotyped by the Polymerase Chain Reaction-Restriction Fragment Length Polymorphism (PCR-RFLP) approach of Felger *et al.* (1993). Individual genotypes were identified by allele-specific patterns.

3.3.4 Data: The total number of worldwide existing *msp2* alleles (genotypes) has not been determined precisely because the number of polymorphic variants is unlimited. Due to intragenic repeats, new variants can continuously arise *de novo*, others may disappear. More than 200 entries of *msp2* alleles have been submitted to the Genbank Felger *et al.* (1997). With an increasing number of sequenced PCR product, the number of new alleles will also increase.

In the KND so far 70 alleles of the 3D7 and FC27 allelic families have been detected by the PCR-RFLP genotyping scheme. This number is expected to increase with a larger number of samples analyzed.

The data for this analysis consists of the genotype panel data of *P. falciparum* from the 69 individuals selected from all age groups as described above. Genotype data are available from all 69 subjects at two months intervals (6 samples per individual including baseline). A total of 70 different genotypes of the FC27 and 3D7 allelic family were identified (Table 3.1). In some individuals, only one specific genotype was observed while in others, as many as 26 were observed during the entire study. The median number of genotypes observed per individual during the entire survey was 11, with first and third quartile values of 7 and 18 respectively. At baseline a total of 183 infections were observed in the 69 individuals.

Let n_{ij} denote the total number of genotypes observed in individual *i* at survey *j*. Then the mean number of genotypes observed in the population over all possible surveys is given by $(\sum_i \sum_j n_{ij})/(69\times 6) = 3.16$. The mean number of genotypes observed within an individual, *i*, at any given survey (and in particular at baseline) is $\sum_j n_{ij}/6$. The average number of distinct genotypes observed in the population at survey *j* is given by $\sum_i n_{ij}/69$. The number of infections acquired in the population decreases as we move from the rainy season through the dry season indicating a strong seasonal variation (Figure 3.1).

Table 3.1. Number of individuals, number of samples, and proportion of samples with a given genotype.

Genotype*	*N*	*n*	*p*	Genotype*	*N*	*n*	*p*
3D7 260	1	1	0.0024	3D7 375	7	7	0.0169
3D7 305	1	1	0.0024	Nav 7	5	7	0.0169
Ifa 10	1	1	0.0024	3D7 385	7	8	0.0193
Ifa 45	1	1	0.0024	3D7 280	7	9	0.0217
Ifa 50	1	1	0.0024	3D7 300	9	9	0.0217
Nav 10	1	1	0.0024	3D7 480	7	9	0.0217
Nav 13	1	1	0.0024	3D7 490	8	9	0.0217
Nav 16	1	1	0.0024	3D7 365	10	10	0.0242
Nav 17	1	1	0.0024	3D7 345	10	11	0.0266
Nav 20	1	1	0.0024	K 1	8	11	0.0266
Nav 21	1	1	0.0024	3D7 335	10	12	0.0290
Nav 4	1	1	0.0024	3D7 470	11	12	0.0290
Ifa 1	2	2	0.0048	Nav 5	7	12	0.0290
Ifa 52	2	2	0.0048	3D7 325	11	13	0.0314
Nav 11	2	2	0.0048	Nav 1	9	13	0.0314
Nav 18	2	2	0.0048	3D7 420	14	15	0.0362
Nav 19	1	2	0.0048	3D7 450	15	16	0.0386
Nav 2	2	2	0.0048	3D7 440	17	22	0.0531
Nav 3	1	2	0.0048	3D7 430	22	24	0.0580
Nav 9	1	2	0.0048	3D7 400	23	29	0.0700
3D7 395	3	3	0.0072	3D7 310	24	30	0.0725
Nav 8	2	3	0.0072	3D7 350	27	32	0.0773
Wos 7	1	3	0.0072	3D7 390	28	35	0.0845
3D7 315	3	4	0.0097	3D7 410	27	37	0.0894
3D7 355	4	4	0.0097	D 10	25	39	0.0942
3D7 405	4	4	0.0097	Wos 10	20	39	0.0942
3D7 510	3	4	0.0097	3D7 360	29	40	0.0966
Ifa 31	3	4	0.0097	3D7 370	37	43	0.1039
Ifa 38	3	4	0.0097	3D7 380	39	48	0.1159
3D7 500	4	5	0.0121	3D7 320	32	50	0.1208
Ifa 13	4	5	0.0121	3D7 340	39	51	0.1232
Nav 6	3	5	0.0121	3D7 330	44	64	0.1546
3D7 290	5	6	0.0145	Wos12	51	110	0.2657
3D7 460	6	6	0.0145	Wos 6	54	159	0.3841
3D7 270	3	7	0.0169	Wos 3	58	198	0.4783

* Genotype names follow references (Felger *et al.*, 1999a; Felger and Beck, 2002)

n is the number of samples (out of 414 samples, that is 69 individuals times 6 samples) with a given genotype. *p* is the proportion of samples with a given genotype. *N* is the number of individuals during the entire 6 surveys with a given genotype.

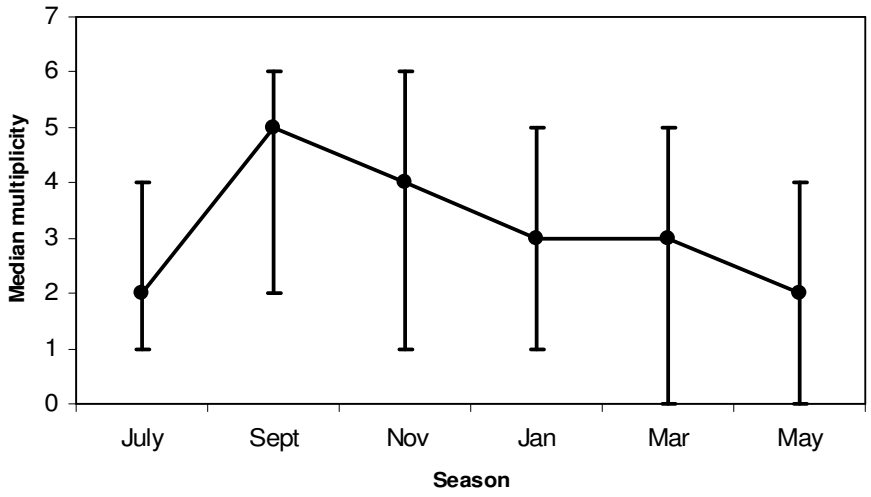


Figure 3.1. Seasonal variation of multiplicity of infections in the population. This represents the median number of genotypes observed in the population at each of the surveys. The period July marks the beginning of the rainy season. The error bars denote lower and upper quartiles.

We consider the parasitological status of these N individuals ($N = 69$) at 6 distinct time-points, 2 months apart. For each genetically distinct infection, this provided a series of presence and absence records in each of the 6 samples. The observed history of the infection was then summarised as a sequence of binary indicator variables using a 1 to denote presence, and 0 to denote absence of the genotype. Thus the sequence 010100 denoted observation of the genotype in the 2nd and 4th samples but not in any of the others. The data analysed consist of the frequencies of the 63 such sequences that are possible (the sequence consisting only of zeros (that is 000000) cannot be counted). We assign an index to each of these possible sequences by evaluating the 6 digit binary representation as the corresponding decimal number, which we refer to as a pattern. For instance the observed binary sequence 010100 corresponds to pattern 20 (twenty)[2].

[2] Converting the decimal number 20 into a binary number gives 10100 which is of length less than six. Whenever this occurs the desired length of six is simply obtained by adding zeros from the left. This is strictly just for notational convenience.

Table 3.2 shows the complete dataset for just two individuals. The first individual was infected with a total of 4 genotypes while the second individual was infected with 6 genotypes. The binary sequence for each of the corresponding genotype is outlined according to whether the individual was infected at the specific survey or not. The complete dataset for the population of 69 individuals consists of 827 such representations. We are interested in modelling the absolute frequencies of the observed patterns. For example in Table 3.2, the frequency of pattern 16 is five while that of pattern 26 is two.

Table 3.2. Example dataset for two individuals

Subject	Genotype	Survey1	Survey2	Survey3	Survey4	Survey5	Survey6	Pattern
1	Wos 3	0	0	1	0	1	1	11
1	Wos 6	0	1	0	0	0	0	16
1	3D7 340	0	1	1	0	1	0	26
1	3D7 370	0	1	1	0	1	0	26
2	Wos 10	0	1	0	0	0	0	16
2	Wos 3	1	1	0	0	0	0	48
2	Wos12	0	1	0	0	0	0	16
2	3D7 380	0	0	1	0	0	0	8
2	3D7 420	0	1	0	0	0	0	16
2	3D7 480	0	1	0	0	0	0	16

Four genotypes were observed for the first subject and six for the second during the entire six surveys. 0 denotes absence and 1 denotes presence of genotype at a given survey.

3.3.5 Model:

We model the dynamics of the parasite population as an immigration-death process. Immigration refers to the acquisition of a new parasitic genotype, occurring at rate λ, death refers to the clearance of a parasitic genotype (with rate μ).

We describe the process with the differential equation (1):

$$\frac{dn(t)}{dt} = \lambda - \mu n(t) \qquad (1)$$

where λ is the per capita recruitment rate, that is the expected number of newly acquired infections per person per year, μ is the rate (per year) at which infections are cleared and $n(t)$ is the expected total number of genotypes within an individual at time t. For the purpose of formulating a base model on which we hope to elaborate in future work, the rates are assumed to be time and population homogeneous. We also assume in this base model that the different genotypes are exchangeable and thus μ to be the same for each genotype, and to be unaffected by co-infections.

In the dataset to be analysed we consider the discrete time process with time interval of constant duration τ years (τ = 2 months = 1/6 year). We define $m(\tau)$ as the expected number (net) of new infections acquired per person in an interval. Solving equation 1 with initial value $m(0) = 0$ gives:

$$m(\tau) = \frac{\lambda}{\mu}(1 - e^{-\mu\tau}) . \qquad (2)$$

Since τ , λ and μ are constants, this quantity, $m_0 = m(\tau)$, is the same for each individual and $M_0 = m_0 N$ is the expected number (net) of infections acquired by a population of size N in a single interval.

Let r denote the probability that an infection is cleared in an interval of duration τ. It follows from the assumption of exponential durations that:

$$r = 1 - e^{-\mu\tau}. \qquad (3)$$

We consider two different models; the basic model and the age-dependent model:

(i) In the basic model we assume that all the *N* individuals in the population are exchangeable and define n_0 to be the expected number of infections present at baseline in each individual and $N_0 = n_0 N$ to be the expected number of infections present at baseline in the population. Thus the basic model assumes that the expected number of infections acquired at any time *t* is always constant.

(ii) The age-dependent model allows for the accumulation of infection with age (equation 1). Let $n(a_k)$ denote the expected number of infections within an individual, *k*, of age, a_k, at baseline. If we assume that the expected number of infections at birth is zero, then by substituting a_k for *t* we obtain a closed form solution of the equation 1:

$$n(a_k) = \frac{\lambda}{\mu}(1 - e^{-\mu a_k}) \qquad (4)$$

This distinguishes between the different members of the human population but it still assumes that the dynamic processes of acquisition and clearance of infections are homogeneous.

3.3.6 Likelihood Computations:

The observed sequences may differ from the true sequences because the observation process is imperfect, with a constant probability of detection, *s*, conditional on an infection being present in the host. We assume that re-infection with a genetically identical parasite clone is a rare event, and thus that observed sequences containing negative samples between two positive samples (for example as in the third sample of the sequence 010100 and as in the third and fourth samples of the sequence 110010) invariably result from failure to detect the infection. A certain degree of bias is certainly introduced in our estimates as a result of this assumption because the sampling interval is rather long. However Table 3.1 and its graphical representation (Figure 3.2) suggest that most genotypes occur only infrequently. Hence only a

subset of the 63 possible observed sequences therefore correspond to possible true sequences. This subset consists of 21 sequences; that is there are only 21 possible true sequences. We will assume that the frequency of each of these observed patterns is Poisson distributed about a certain expectation which would be derived in the section that follows.

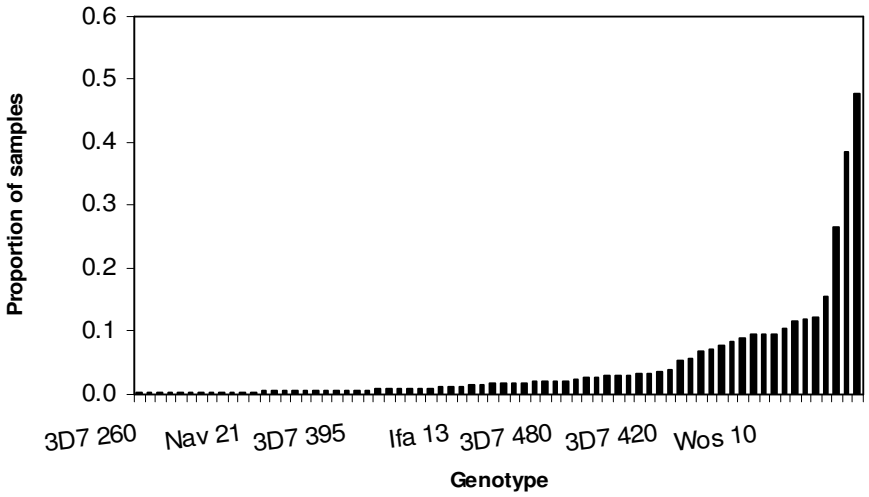


Figure 3.2. The proportion of samples with a given genotype. The data for this plot is given in Table 3.1.

In order to work out the expected frequency of the true patterns in the population we divide the set of patterns into two groups, those where the infection was present at baseline and those where it was not. We further sub-divide the patterns without infection at baseline into those where new infections were acquired during the second survey and those where new infections were acquired during the third survey and so on. The tree diagram (Figure 3.3) illustrates the derivation of the expected frequency, $f_T(j)$, for 11 of the 21 true patterns, j. The derivations of the rest are quite similar and are not presented. However the expected frequency of a true pattern in the population using the basic model can be generalised as follows:

$$\begin{aligned} f_T(j) &= N_0^{v_j} M_0^{b_j} r^{c_j} (1-r)^{d_j} \\ &= N_0^{v_j} M_0^{b_j} (1-e^{-\mu\tau})^{c_j} (e^{-\mu\tau})^{d_j} \end{aligned} \qquad (5)$$

where,

$$v_j = \begin{cases} 1 & \text{if the sequence correspoding to pattern } j \text{ starts with a "one"} \\ 0 & \text{if it starts with a "zero"} \end{cases}$$

$$b_j = \begin{cases} 1 & \text{if there is a transition from 0 to 1 in the sequence corresponding to pattern } j \\ 0 & \text{if there is none} \end{cases}$$

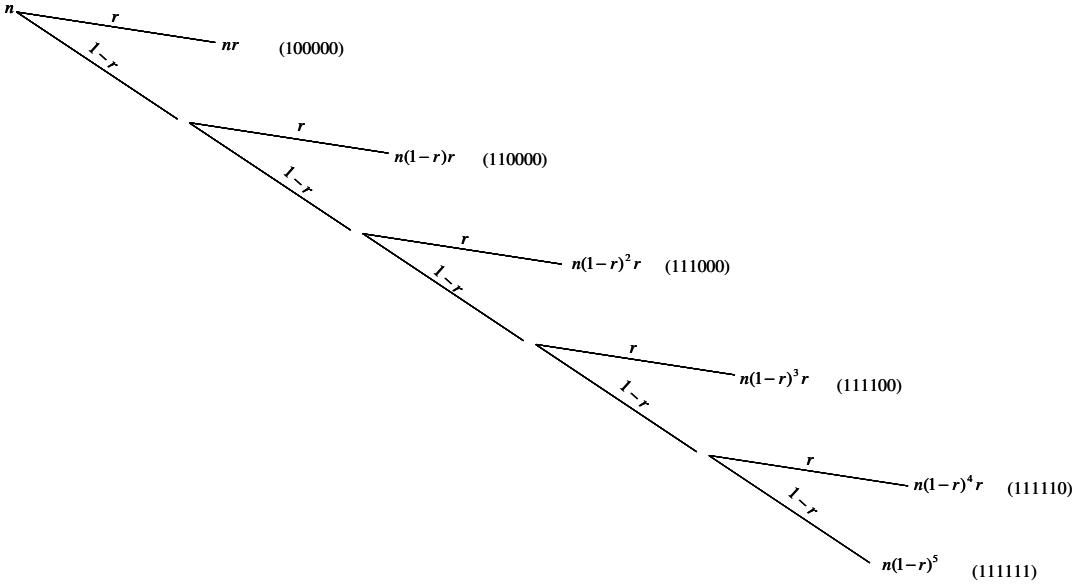
$$c_j = \begin{cases} 1 & \text{if there is a transition from 1 to 0 in the sequence corresponding to pattern } j \\ 0 & \text{if there is none} \end{cases}$$

d_j = the number of transitions from 1 to 1 in the sequence corresponding to pattern j

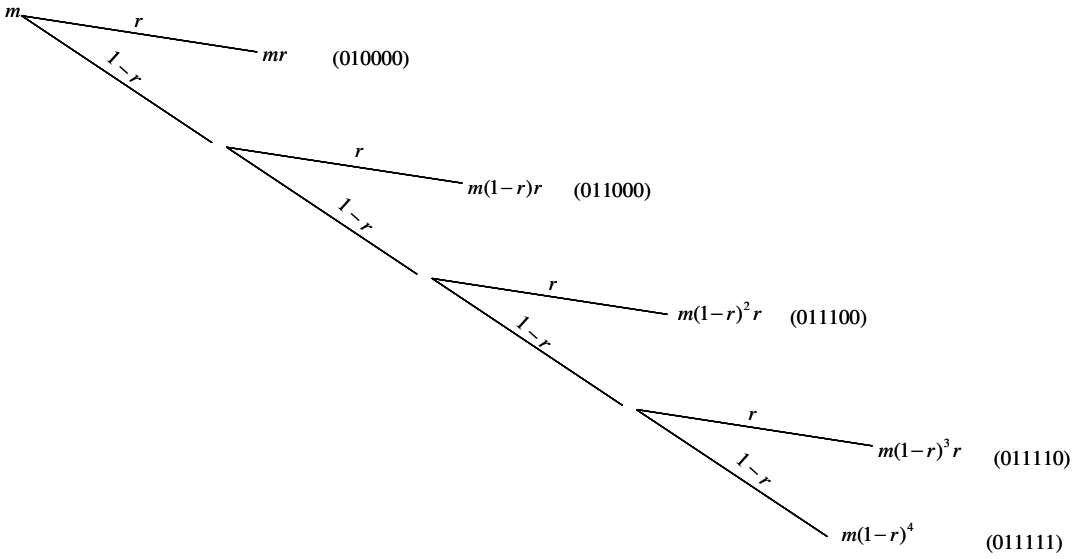
The parameter r of equation 5 has been substituted by its equivalent expression given in equation 3.

Figure 3.3. Tree diagrams demonstrating the derivation of the expected true frequencies for 11 out of the 21 possible true patterns. n and m respectively denotes the expected number of infections present at baseline and the expected (net) number of infections acquired per interval. Hence n will be replaced by N_0 and m by M_0 for the basic model; while $n = n(a_k)$ and $m = m_0$ for the age-dependent model.

(a) Expected frequency of true patterns where genotype was present at baseline.



(b) Expected frequency of true patterns where genotype was present at second survey.



To work out the expected frequency of an observed pattern, *i*, let X_O denote the observed process which is the detection or failure to detect a parasitic genotype in a sample at any time point, and let X_T denote the unobserved process, that is the underlying true presence of the genotype in the host. We define the detectability *s*, as $s = P(X_O = 1 \mid X_T = 1)$. This is different from the sensitivity of the PCR which is the conditional probability of detecting parasites present in the sample. The primers used for PCR have been shown to be specific for *P. falciparum* (Felger *et al.*, 1999a, 1994). We therefore assume that the test has perfect specificity, that is, $P(X_O = 1 \mid X_T = 0) = 0$.

Let Y_O denote the observed pattern and let Y_T denote the unobserved true pattern. Any observed pattern, *i,* can actually arise from a certain number of true patterns. Using the perfect specificity assumption, it is easy to list all the unobserved true patterns, *j*, that can give rise to an observed pattern, *i.* For example, due to imperfect detectability, the observed pattern 9 (001001) can arise either from the true patterns 15 (001111) or 31 (011111) or 63 (111111) so that the expected frequency of the observed pattern 9 in the population will depend on the

expected frequencies of the true patterns 15, 31, and 63, in the population. Let $S_{ij} = P(Y_O = i \mid Y_T = j)$ denote the probability of observing pattern *i* conditional on the true pattern being *j*. For instance, $S_{9,15} = P(Y_O = 9 \mid Y_T = 15) = s^2(1-s)^2$; similarly $S_{9,31} = s^2(1-s)^3$ and $S_{9,63} = s^2(1-s)^4$. The expected frequency of observed pattern 9, $f_O(9)$, in the population is thus given by

$$\begin{aligned} f_O(9) &= S_{9,15} f_T(15) + S_{9,31} f_T(31) + S_{9,63} f_T(63) \\ &= s^2(1-s)^2 M_0(1-r)^3 + s^2(1-s)^3 M_0(1-r)^4 + s^2(1-s)^4 N_0(1-r)^5 \end{aligned}$$

In general, for any given observed pattern, *i,* and a true pattern *j*,

$$S_{ij} = I_{ij} s^{z_i} (1-s)^{z_j - z_i} \qquad (6)$$

where,

$$I_{ij} = \begin{cases} 1 & \text{if the observed pattern } i \text{ can arise from the true pattern } j \\ 0 & \text{otherwise} \end{cases}$$

z_i = number of "ones" in the binary sequence corresponding to observed pattern *i*
z_j = number of "ones" in the binary sequence corresponding to true pattern *j*

The expected frequency of observed pattern *i* in the population is therefore given by

$$f_O(i) = \sum_j S_{ij} f_T(j) \qquad (7)$$

This enables us to compute the Poisson log-likelihood for each of the possible observed patterns as a function of M_0, N_0, *s*, and $\boldsymbol{\mu}$.

Let X_i, i = 1, 2, 3,....63 be independent random samples of the random variable X, where X denote the frequency of occurrence of a given pattern. Let the observations x_i be a realization of X_i, and assume that each X_i is Poisson distributed with mean $f_O(i)$; that is $X_i \sim Poisson(f_O(i))$. The contribution to the likelihood of observation x_i obtained from the whole population and the *i*th pattern is given by

$$L(M_0, N_0, s, \mu; x_i) = \frac{\exp(-\sum_j S_{ij} f_T(j))(\sum_j S_{ij} f_T(j))^{x_i}}{x_i!} \qquad (8)$$

Therefore the likelihood for the complete data is given by

$$L(M_0, N_0, s, \mu; x) = \prod_{i=1}^{63} \frac{\exp(-\sum_j S_{ij} f_T(j))(\sum_j S_{ij} f_T(j))^{x_i}}{x_i!} \qquad (9)$$

The generalisation of these equations (5, 7, 8, 9) to the age-dependent model is straight-forward. For instance, we now refer to the expected frequency of the true pattern, *j*, in an individual, *k*, of age, a_k, as.

$$\begin{aligned} f_{T_k}(j) &= [n(a_k)]^{v_j} m_0^{b_j} r^{c_j} (1-r)^{d_j} \\ &= (\frac{\lambda}{\mu})^{v_j+b_j} (1-e^{-\mu a_k})^{v_j} (1-e^{-\mu\tau})^{b_j+c_j} (e^{-\mu\tau})^{d_j} \end{aligned} \qquad (10)$$

and the expected frequency of the observed pattern, *i*, for an individual, *k*, of age, a_k, as

$$f_{O_k}(i) = \sum_j S_{ij} f_{T_k}(j) \qquad (11)$$

We have substituted m_0, r, and $n(a_k)$ of equation 10 by their equivalent expressions respectively given in equations 2, 3 and 4.

Similarly if x_{ki} is a realisation of the random variable X_{ki} where X_{ki} denote the frequency of occurrence of pattern, *i,* for individual *k* and we again assume a Poisson error function for the frequency of the observed patterns, that is $X_{ki} \sim Poisson(f_{O_k}(i))$, then the contribution to the likelihood of observation x_{ki} obtained from the *k*th individual and the *i*th pattern is given by

$$L(s, \lambda, \mu; x_{ki}) = \frac{\exp(-\sum_j S_{ij} f_{T_k}(j))(\sum_j S_{ij} f_{T_k}(j))^{x_{ki}}}{x_{ki}!} \qquad (12)$$

Therefore the likelihood for the complete data $x = (x_{ki})$ is given by

$$L(s, \lambda, \mu; x) = \prod_k \prod_i \frac{\exp(-\sum_j S_{ij} f_{T_k}(j))(\sum_j S_{ij} f_{T_k}(j))^{x_{ki}}}{x_{ki}!} \qquad (13)$$

We use both maximum likelihood methods and Bayesian inference (using MCMC simulation) to obtain point and interval estimates for *s*, λ, and μ, and compare the appropriateness of these two methods for application in future work extending these models.

The maximum likelihood estimation was implemented in FORTRAN 95 (*Compaq Visual Fortran v6.6. Compaq Computer Corporation, Houston 2001*) using the quasi-Newton algorithm (Gill and Murray, 1976) from the NAG FORTRAN library (*NAG Fortran Manual, Mark 19, NAG Ltd, 1999*). Confidence intervals were calculated by inverting the observed information matrix (see for instance Davison (2003)).

Within the Bayesian framework, we assume non-informative Log-normal priors for the parameters M_0, N_0, λ, μ, and a Uniform(0,1) prior for the parameter *s*. To analyse the sensitivity to the priors of M_0, N_0, λ and μ, we first changed the precision assigned to the Log-normal distributions and secondly evaluated the effects of using Gamma priors for these parameters. The models were fitted using WinBUGS version 1.4 (Spiegelhalter *et al.*, 2003) employing a Metropolis algorithm. The chains were run several times with different starting values, with a single chain monitored during each run and convergence assessed in BOA (Smith, 2003) using Heidelberger and Welch (1983) tests. The quoted results are based on samples of 5000 values from the posterior densities following a burn in of 10000 iterations.

3.4 Results.

The Cramer-Von-Mises statistics from the Heidelberger and Welch stationary test were respectively 0.05 ($P = 0.70$), 0.13 ($P = 0.11$), 0.13 ($P = 0.15$), 0.16 ($P = 0.07$) for M_0, μ, N_0, *s*, using the basic model and 0.16 ($P = 0.17$), 0.36 ($P = 0.14$), 0.31 ($P = 0.11$) for λ, μ, and *s*, using the age-dependent model, suggesting convergence of the chains.

The results obtained from the basic and the age-dependent model using both fitting procedures are shown on Table 3.3 (a and b respectively). The Bayesian estimates are

insensitive to the prior specifications and there is little difference in the results obtained from the basic and age-dependent model. The reason for this can easily be explained by looking at Figure 3.4. Equilibrium is attained between the ages of 4 and 5 years (Figure 3.4) and since only 14/69 of the sample (contributing 152/827 sequences) were aged less than 4 years, most of the information is contributed by individuals above this age. The value for $N_0 / 69$ (estimated as 7.40) for the basic model is similar to the equilibrium value $\frac{\lambda}{\mu}$ (estimated as 6.80) from the age-dependent model.

Table 3.3 a. Parameter estimates from the basic model.

Parameter	Method of estimation				
	MLE	Bayesian Inference			
		Log-normal(0,10^v) priors on M_0, N_0, μ, and a Uniform(0, 1) for s.			Gamma(0.001 ,0.001) priors on M_0, N_0, μ and a U(0, 1) for s
		$v = 6$	$v = 4$	$v = 2$	
M_0	146 (129-163)	146 (128-163)	146 (129-163)	146 (128-163)	146 (129-163)
N_0	511 (418-604)	510 (421-608)	515 (426-612)	514 (428-615)	511 (425-608)
s	0.45 (0.40-0.50)	0.46 (0.41-0.51)	0.45 (0.40-0.50)	0.45 (0.40-0.50)	0.46 (0.41-0.51)
λ (per year)	15.4 (13.3-17.7)	15.4 (13.2-17.7)	15.4 (13.3-17.7)	15.4 (13.2-7.7)	15.4 (13.3-17.7)
μ (per year)	2.40 (2.06-2.75)	2.41 (2.06-2.77)	2.38 (2.06-2.74)	2.39 (2.06-2.75)	2.41 (2.07-2.77)

The mean value of the parameter estimates with the 95% confidence intervals (and 95% credible intervals) in brackets. λ was estimated as $\lambda = \frac{M_0 \mu}{69(1 - e^{-\frac{\mu}{6}})}$ using equation 2, while the rest of the parameters were estimated directly from the likelihood.

Table 3.3 b. Parameter estimates using the age-dependent model.

Parameter	Method of estimation				
	MLE	Bayesian Inference			
		Log-normal(0,10^{v}) priors on λ, μ, and a Uniform(0, 1) for *s*.			Gamma(0.001, 0.001) priors on λ, μ, and a U(0, 1) for *s*
		$v = 6$	$v = 4$	$v = 2$	
M_0	155 (143-166)	155 (144-166)	154 (143-167)	154 (143-165)	155 (143-166)
N_0	453 (451-456)	453 (439-478)	452 (443-473)	455 (439-478)	453 (439-471)
s	0.47 (0.42-0.51)	0.47 (0.42-0.51)	0.47 (0.42-0.51)	0.46 (0.42-0.51)	0.47 (0.42-0.51)
λ (per year)	16.3 (14.8-17.8)	16.3 (14.8-18.0)	16.3 (14.7-18.0)	16.3 (14.7-17.8)	16.3 (14.8-18.0)
μ (per year)	2.40 (2.16-2.64)	2.40 (2.06-2.74)	2.40 (2.07-2.72)	2.38 (2.04-2.71)	2.40 (2.08-2.74)

M_0 and N_0 were estimated as $M_0 = \frac{69\lambda}{\mu}(1-e^{-\frac{\mu}{6}})$, $N_0 = \sum_k \frac{\lambda}{\mu}(1-e^{-\mu a_k})$, using equations 2 and 4 respectively, while the rest of the parameters were estimated directly from the likelihood.

Distinct plots of the age variation in the observed multiplicity of infections at each of the surveys were similar to the scatter plot of the observed mean multiplicity at any given survey (earlier denoted as $\sum_j n_{ij}/6$) in Figure 3.4 and hence are not presented. A similar feature was also observed for the plot of the age variation in the total number of distinct genotypes found within each individual during the entire study.

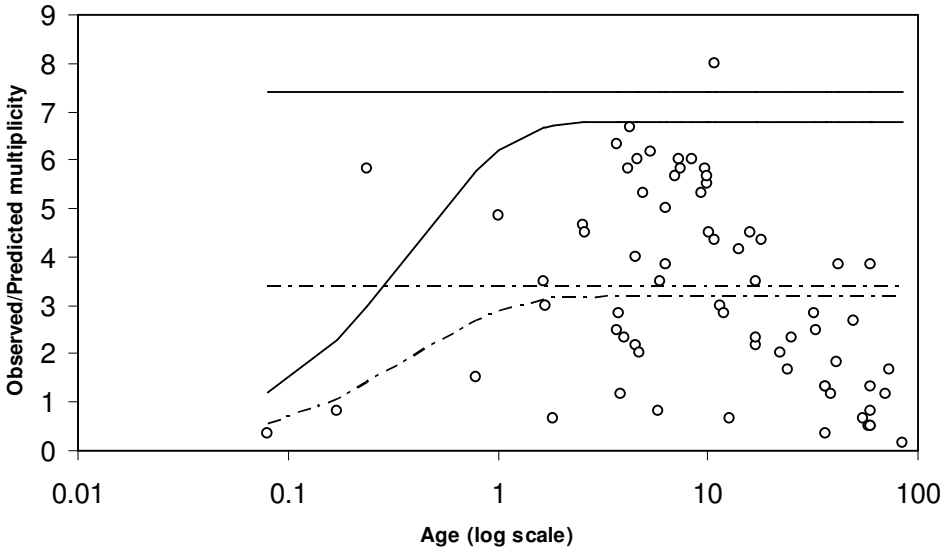


Figure 3.4. Mean multiplicity of infection at baseline* as a function of age. The circles represent the observed mean multiplicity (given by $\sum_j n_{ij}/6$). The horizontal broken line and the broken curve represent the predictions for the observed mean multiplicity using the basic (given by $sN_0/69$) and age-dependent (given by $s\frac{\lambda}{\mu}(1-e^{-\mu a_k})$) models respectively. The horizontal full line and the full curve represent the predictions for the true mean multiplicity using the basic (given by $N_0/69$) and age-dependent (given by $(\frac{\lambda}{\mu}(1-e^{-\mu a_k}))$ model respectively. *It could as well be "at any given survey" since the mean was calculated over all the surveys.

The frequencies of most of the 63 patterns were close to the predictions (Figure 3.5). The largest residual corresponded to pattern 16 (010000) which occurred 147 times (expected frequency 79.8). Four other patterns had residual frequencies >20 or <-20. All of these five patterns represented infections that were observed at only one survey, thus suggesting that the model fits least well for infections that persist for only a short time. When we assessed the ratio between the observed and expected frequencies, Pattern 63 (111111), which occurred 19

times, but with an expected frequency of only 0.6, gave a very poor fit. A possible explanation for these discrepancies is that the true distribution of the durations of infections has a greater variance than is assumed by the exponential model.

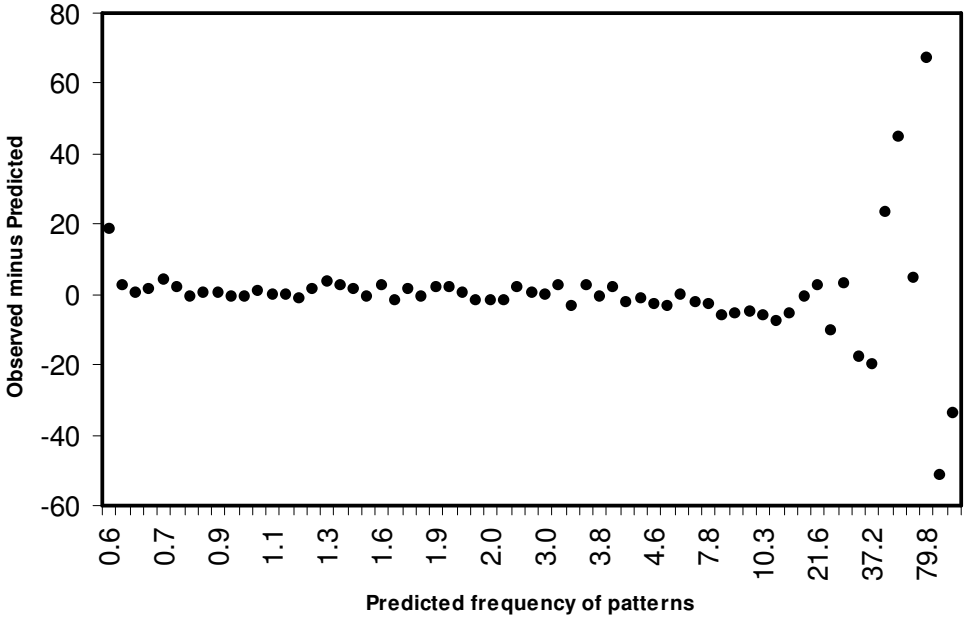


Figure 3.5. A plot of the difference between the observed and predicted frequency of the 63 patterns against the predicted frequency of the patterns using the basic model.

Our results suggests that on average any individual residing in a holo-endemic area will acquire about 16 new genotypes per year and that infection with any of these genotypes will last approximately 152 days (95%CI, 133 – 178) using the maximum likelihood method for the basic model. The same results using the maximum likelihood method for the age-dependent model is 152 days (95%CI, 138–169). It was estimated that the probability of clearing an infection with any given genotype within a two months period was 0.33 (95%CI, 0.29 – 0.37). The predicted value for the total observed multiplicity (that is 183) in the whole population at baseline was 511 (95%CI, 418 – 604) using the basic model.

* The estimate for the clearance rate μ (or the duration of infection, $\frac{1}{\mu}$) stated in the work above is for a single infection. We now illustrate how to obtain the expression for the total duration of infection for an individual with multiple infections. Since the clearance rate for a single infection is μ, it follows that the rate of converting from two infections to one infection is 2μ, and in general from *n* infections to (*n* – 1) infections is $n\mu$. It thus follows that the total duration of *n* infections is $\sum_{K=1}^{n} \frac{1}{k\mu}$. In our results above, we estimated an average duration of 152 days for a single infection. The estimated expected multiplicity of infections in the study population of 69 individuals is 7. Applying the formula above, this gives an estimate of 394 days for the expected total duration of infections.

3.5 Discussion.

Estimates of the duration of malaria infection are of particular practical importance in the evaluation of effects on malaria control of interventions, such as mosquito nets, residual insecticide spraying and vaccines. However few studies have attempted to estimate this quantity from field data.

The major challenges arise because newly acquired infections cannot be distinguished from pre-existing ones by optical microscopy. Consequently estimates from microscopy represent average durations for a combination of new and pre-existing infections (see for example Sama *et al.* (2004)). The bias is increased by treating temporarily sub-patent infections as having been cleared, leading to estimates of the duration of parasitaemic episodes rather than of the total duration of infection. Microscopy readings at very frequent intervals indicated that asymptomatic episodes of patent *P. falciparum* parasitaemia in children in a Papua New Guinean (PNG) village lasted only 3-27 days (Bruce *et al.*, 2000a).

* This paragraph is not included in the published work.

Analyses of longitudinal typing data have considerable potential for elucidating the dynamics of parasitic diseases. They can provide realistic estimates of infection and recovery rates even in endemic areas where most individuals are already infected, because they make it possible to distinguish new infections from old ones. Thus in the PNG study (Bruce *et al.*, 2000a) it was found that the mean duration of episodes of positivity for individual *P. falciparum* genotypes was approximately 60 days (Bruce *et al.*, 2000b). Duration of parasitaemic episodes does not reflect the total duration of infections, since a single infection embraces sub-patent intervals interspersed between parasitaemic episodes (Sama *et al.*, 2004). Typing studies have found parasite genotypes persisting asymptomatically for at least 40 weeks in Ghana (Franks *et al.*, 2001) and for more than 12 months in Eastern Sudan (Babiker *et al.*, 1998).

Our approach gives an estimate for the population average of total duration of infection with any specific genotype. Although malaria modelling has long made use of immigration-death models, (see for instance Macdonald (1950a), for a seminal idea in this field) other models have not estimated the detectability (see for example, (Bekessy *et al.*, 1976; Richard *et al.*, 1993)).

Our analysis ignores the possibility of re-infection with the same genotype. When we disaggregated the parasite population into genotypes, most of them were infrequent (Table 3.1, Figures 3.2) so super-infection with the same genotype is relatively rare. However, since the estimate for the detectability of the PCR is quite low (47%), the genotypes may be more frequent than Table 3.1 suggests, though other studies have also estimated low detectability of 51% (Smith *et al.*, 1999b) and have shown that it could be as low as 20% for older individuals (Smith and Vounatsou, 2003). Misclassification of genotypes could lead to low estimates of detectability, but DNA sequencing of PCR products indicates that misclassification of the

different patterns is infrequent, although sometimes the whole pattern might be assigned to the wrong bin.

The resolution of the PCR-RFLP genotyping is limited, as with all genotyping techniques. The method of choice would be direct sequencing of each PCR product but the high frequency of mixed genotype infections precludes a sequencing approach. In order to optimize our laboratory technique for the task of longitudinal tracking of individual genotypes as required in the present study, we are currently developing a more accurate sizing technique which will eventually be tested on the same set of samples to assess the degree of misclassifications and its impact on detectability of the PCR.

Since the sampling interval was long (2 months), we were concerned that re-infections with the same genotypes may have biased the results. Wos3, Wos6 and Wos12 genotypes were the most frequent (Table 3.1 and Figure 3.2). We repeated the analysis for the basic model with a reduced datasets obtained by successively removing the most frequent genotype (Wos3), followed by the two most frequent genotypes (Wos3 and Wos6), up to the eight most frequent genotypes. Apart from effects on the detectability parameter, s, the estimates are approximately of the same order with a slight increase in the width of the confidence limits (Table 3.4). Our estimates of the duration of infection were thus little affected by this assumption. The drop in the detectability parameter was expected because by consistently removing the most frequent genotypes, the number of positive samples is reduced.

Table 3.4 Sensitivity analysis of parameter estimates to the assumption that re-infection with a specific genotype is a rare event. This was obtained from basic model using Bayesian inference.

Parameter	Number of most frequent* genotypes removed from the analysis								
	None+	1	2	3	4	5	6	7	8
M_0	145.9 (128.2-162.6)	158.5 (139.5-180.0)	172.9 (145.4-201.5)	183.0 (149.4-213.9)	176.9 146.1-208.5	165.6 (130.2-195.8)	161.8 (127.6-193.6)	152.1 (116.8-183.7)	141.7 (115.1-169.2)
N_0	509.7 (421.3-608.1)	530.3 (419.7-655.3)	607.4 (462.2-778.7)	632.8 (470.7-835.0)	592.9 (422.8-826.2)	570.7 (392.7-801.4)	572.8 (388.7-822.5)	526.0 (346.7-781.8)	437.2 (288.9-645.9)
S	0.46 (0.41-0.51)	0.38 (0.32-0.44)	0.29 (0.24-0.35)	0.25 (0.20-0.31)	0.25 (0.19-0.31)	0.24 (0.18-0.31)	0.23 (0.17-0.29)	0.23 (0.16-0.30)	0.24 (0.18-0.32)
μ (per year)	2.41 (2.06-2.77)	2.57 (2.13-3.07)	2.56 (1.99-3.13)	2.59 (1.89-3.24)	2.68 (1.82-3.446)	2.57 (1.73-3.48)	2.58 (1.73-3.52)	2.55 (1.67-3.43)	2.67 (1.86-3.59)

The mean value of the parameter estimates with the 95% credible intervals in brackets. * See Table 3.1 for the most frequent genotypes. + None indicates that the analysis was done with the complete dataset.

It is possible to fit models (Hidden Markov Models (MacDonald and Zucchini, 1997)) which simultaneously estimate both the incidence of new infections and persistence, allowing for imperfect detectability and for re-infection with the same genotype (Smith and Vounatsou, 2003) but we do not adopt this approach because the models are highly parameterized due to the large number of genotypes identified so that the results obtained from such an approach may be highly unstable due to the sparsity of the data.

It is important to understand how the dynamics of malaria infections vary with age and as immunity is acquired. The present analysis does not take all these factors into consideration. We have assumed a homogeneous human and parasite population as a result of the

assumption of constant parameters. A number of studies in endemic areas have reported that infections are of relatively short duration in very young children. For example Walton (1947) found that the average duration of infection in infants in Freetown, Sierra Leone, was little over 3 months, at a time when there was relatively little transmission there. More recently it was estimated that the duration of infections with parasites belonging to the msp2 FC27 allelic family increased with age using data collected from Tanzanian children aged 6-30 months (Smith *et al.*, 1999c; Smith and Vounatsou, 2003). An increase in duration with age during the first two years of life has also been reported in a study with Ghanaian children (Franks *et al.*, 2001).

In older individuals there was a poor fit of our model (Figure 3.4), indicating that the infection rate, detectability or clearance rate is age-dependent. Since acquired immunity reduces parasite density it is evident that the detectability decreases with age, while the infection rate clearly varies with season. Macdonald's original model (Macdonald, 1950a) assumed no age-dependence in infection or clearance rates but it is known that as children grow older they are more frequently bitten by mosquitoes and hence presumably acquire infections at a faster rate (Port *et al.*, 1980). Very likely infection rates decrease after a certain age as a result of acquired immunity.

It was assumed in the age-dependent model that the mean number of infections acquired per individual is time homogeneous. Figure 3.6 shows the time series representation of this event for 16 randomly chosen individuals, together with the prediction for the observed value. The message here is that there can be a lot of variation in this number between different individuals and between different time points (perhaps due to seasonal variation in infection rates). A complete representation for the 69 individuals will surely show a very complex structure from which possible realistic trends need to be carefully examined. Possible

extensions of the model to include seasonal variation in the infection rate are aspects under consideration.

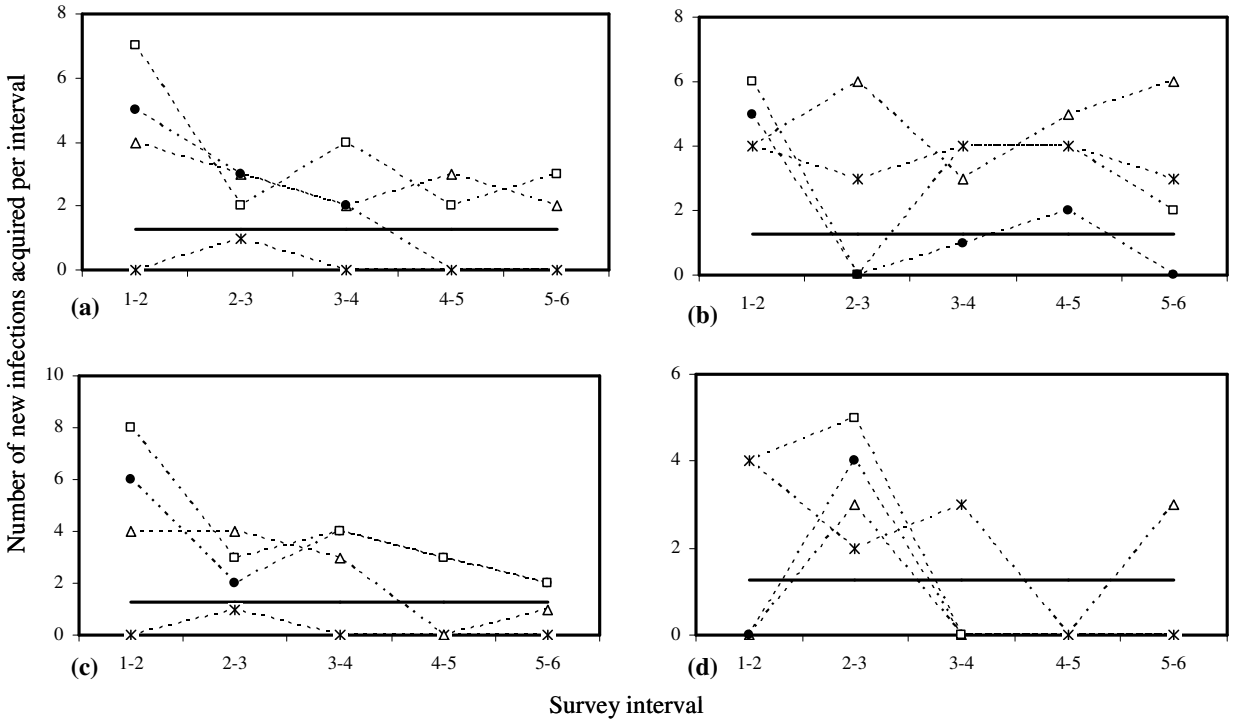


Figure 3.6. (a)-(d). The number of observed infections acquired per interval for 16 (four individuals per graph) randomly selected individuals of different ages from the population of 69 individuals. Each of the four broken lines represents an individual. The full line is the prediction for the observed value.

It would also be desirable to extend our model to incorporate effects of other explanatory variables such as the use of insecticide treated nets which have been shown to reduce infection rates [23], to investigate equivalence of genotypes and interactions between them, and to include alternatives to the exponential distribution for the survival times of the infections. Our model however provides a basis for such extensions.

3.6 Acknowledgements.

The authors appreciated comments from Louis Molineaux, Neal Alexander and Anthony Davison, and are particularly grateful to Klaus Dietz for his suggestions which helped to considerably improve the paper. We also thank the staff of the Navrongo Research Health Centre involved in the field work and the villagers for their co-operation. Wilson Sama is in receipt of a stipend from the Stipendiumkommission of the Amt für Ausbildungsbeiträge of the Canton of Basel, Switzerland.

CHAPTER 4

Age and seasonal variation in the transition rates and detectability of *Plasmodium falciparum* malaria

Wilson Sama[1], Seth Owusu-Agyei[2], Ingrid Felger[1], Klaus Dietz[3], & Tom Smith[1].

[1]Swiss Tropical Institute, Basel, Switzerland.

[2]Kintampo Health Research Centre, Kintampo, Ghana

[3]Institut für Medizinische Biometrie, Tübingen, Germany.

This paper has been published in

Parasitology 2006; **132:**13-21

4.1 Abstract

The effect of acquired immunity on the duration of *Plasmodium falciparum* infections is unclear, although this is an important term in models of malaria transmission. It is problematical to determine the duration of infections because of the difficulty of distinguishing persisting infections from new ones, and because parasite densities are often transiently below the limit of detection. We recently developed a dynamic model for infection incidence, clearance, and detection of multiple genotype *Plasmodium falciparum* infections and fitted it to a panel dataset from a longitudinal study in Northern Ghana. We now extend this model to allow for seasonal and age variation in infection rates and also age dependence in clearance and in detectability of infections. These models indicate that there is seasonal variation in the infection rate, and age dependence in detectability. The best fitting models had no age dependence in infection or clearance rates, suggesting that acquired immunity mainly affects detectability.

4.2 Introduction

The dynamics of *Plasmodium falciparum* malaria infections in endemic areas are crucial determinants of the effects of preventative interventions but are difficult to study because of the challenge of distinguishing persisting infections from new ones, and because parasite densities are often transiently below the limit of detection.

Where infectious agents are endemic, levels of immunity generally increase with age, and it is therefore to be expected that the duration of infections will be lower in older hosts than in younger ones. While this pattern has been assumed in many models of malaria transmission (Aron, 1988; Dietz *et al.*, 1974), the empirical evidence for it is weak (Kitua *et al*. 1996; Sama *et al.*, 2004; Smith *et al*. 1999b; 1999c; Smith and Vounatsou, 2003; Walton, 1947).

We recently reported a model (Sama *et al*., 2005) to estimate infection and recovery rates (and hence the duration of infections) from repeated observations of the presence or absence of *Plasmodium falciparum* genotypes in the same group of individuals in the Kassena- Nankana district (KND) in Northern Ghana. In developing this approach we treated the infection and recovery rates as constant. It was assumed that the laboratory test used to detect the infections agent has imperfect detectability, but the detectability was assumed to be constant across the whole population. We also assumed that these parameters are the same for all genotypes.

We now extend our previous model (Sama *et al*., 2005) by explicitly parameterising the detectability as well as the infection and recovery processes as functions of age. We also extend our model to allow for seasonal variation in infection rates.

4.3 Materials and Methods

4.3.1 Field surveys

The KND, is a highly endemic malarious area showing a peak in transmission during the short wet season between May to September and corresponding seasonality in prevalence and clinical incidence (Baird *et al*., 2002; Binka *et al*., 1994; Koram *et al*., 2003). We analysed *P. falciparum* genotype data of 69 individuals (Sama *et al*., 2005), ranging in age from 1 month to 84 years. Each individual contributed six blood samples, with the first survey carried out in July 2000 and the remaining surveys at regular intervals of 2 months subsequently, each survey lasting for 8 or 9 days. Samples were analysed for presence or absence of *P. falciparum* merozoite surface protein 2 genotypes using a PCR-RFLP (Polymerase Chain Reaction-Restriction Fragment Length Polymorphism) technique (Felger *et al*., 1999a). There were a total of 70 *msp2* genotypes observed, 33 belonging to the FC27 and 37 to the 3D7 allelic family.

As an example of the structure of the dataset analysed, consider an individual observed with 4 genotypes during the entire six surveys, whose data was represented by the following four sequences: (001001), (010000), (001110), (111100), corresponding to the four genotypes observed and indicating that the first genotype was observed during the third and sixth survey, while the second genotype was observed only during the second survey, and so on. The complete dataset consisted of 827 observed sequences of this nature.

4.3.2 Model of parasite dynamics

We previously used the immigration-death model corresponding to that originally proposed by Macdonald (1950a) to describe $n(a)$, the expected number of distinct

genotypes (or the expected true multiplicity of infections) within an individual of age *a* by the following equation.

$$\frac{dn(a)}{da} = \lambda - \mu n(a) \qquad (1)$$

where λ is the infection rate, that is the rate at which new infections are acquired, and μ is the clearance rate, that is the rate at which infections are cleared, both assumed to be homogeneous across the population. We now allow these parameters λ, μ, to vary by age and season, *t*, so that equation 1 takes the form.

$$\frac{dn(a,t)}{da} = \lambda(a,t) - \mu(a)n(a,t) \qquad (2)$$

The general solution of this equation is given by:

$$n(a,t) = e^{-\int \mu(a)da}\left[K + \int \lambda(a,t)e^{\int \mu(a)da}da\right] \qquad (3)$$

where K is the constant of integration obtained by substituting the initial condition $n(0,t) = 0$.

(a) ***Infection process***

We evaluate three alternative forms for $\lambda(a,t)$

(i) $\lambda(a,t) = \lambda$, i.e. we treat the infection rate as constant.

(ii) $\lambda(a,t) = \beta_0 e^{\beta_1 a}$, i.e. we assume that the infection rate is a monotonic function of age, where β_0 is the infection rate at birth while β_1 is the change in infection rate (on the logarithmic scale) for a unit increase in age.

(iii) $\lambda(a,t) = \lambda_i$, i.e. seasonal variation in infection rates. i = 1, 2, …, 6 indexes the two month period of the year at time *t*, corresponding to the inter-survey interval and the

parameters $\lambda_1, \lambda_2, ..., \lambda_6$ account for seasonal variation in the infection rate, with the rates treated as constant within each inter-survey period. In this model $\lambda(a,t)$ is assumed to follow a recurring annual cycle and the same vector of parameters, λ_i, is applied for each year of life of the individual preceding the observation period.

(b) ***Clearance process***

We evaluate two alternative forms for $\mu(a)$

(i) $\mu(a) = \mu$

(ii) $\mu(a) = \mu_0 e^{\mu_1 a}$

μ_0 is the clearance rate at birth while μ_1 is the change in clearance rate (on the logarithmic scale) for a unit increase in age.

(c) ***Observation process***

The observed sequences may differ from true sequences because the observation process is imperfect. We assume one of the following forms for the detectability, *s*, the probability of detecting an infection in a blood sample conditional on it being present in the host.

(i) Constant detectability, $s(a) = s$;

(ii) Age-dependent detectability, $\text{logit}\left[s(a)\right] = s_0 + s_1(a - \bar{a})$, where $\bar{a}$ is the mean age (20.12) and s_0 and s_1 are parameters to be estimated.

The prediction for the observed mean multiplicity (or simply the expected multiplicity) is obtained by multiplying equation 3 by *s*(*a*), that is:

$$\tilde{n}(a,t) = s(a)e^{-\int \mu(a)da}\left[K + \int \lambda(a,t)e^{\int \mu(a)da}da\right] \qquad (4)$$

The expected (net) true number of infections acquired during the interval t to $t+\tau$ in an individual of initial age a is then $m(a+\tau,t+\tau)$ where:

$$\frac{dm(a,t)}{da} = \lambda(a,t) - \mu(a)m(a,t) \qquad (5)$$

and the initial condition is $m(a,t) = 0$ (Sama *et al*., 2005). In this paper, all units of measurement for time and age are years and for rates are year^{-1}.

We assumed that re-infection with a genetically identical parasite clone is a rare event, and thus that observed sequences containing negative samples between two positive samples (for example as in the third sample of the sequence 010100 and as in the third and fourth samples of the sequence 110010) invariably result from failure to detect the infection. We also assumed that our test had perfect specificity. By maintaining these two assumptions, the derivation of the Poisson-likelihood for the frequency of each observed sequence is the same as previously described in (Sama *et al*., 2005). The solutions of equations 2 and 5 are important components in this Poisson likelihood function. The models considered here are thus fitted to the whole genotype histories through time of the individual patients as in Sama *et al.* (2005).

The models with the specifications a(i), b(i), and c(i) above have been discussed in (Sama *et al*., 2005). We now consider all the remaining possible combinations of (a), (b) and (c) (see Table 4.1). The models are fitted using maximum likelihood, employing a fixed order Runge Kutta method (Shampine, 1994) for the numerical integration of equation 3 and the quasi-Newton algorithm (Gill and Murray, 1976) for the maximization

process. Confidence intervals were obtained by inverting the observed information matrix (Davison, 2003). The programming was implemented in Fortran 95 (*Compaq Visual Fortran Version 6.6. Compaq Computer Corporation. Houston, Texas,* 2001)

The improvement of the fit gained by considering the likelihood of a fuller model L_2 containing $p+q$ parameters with respect to the likelihood of a reduced nested model L_1 containing only p parameters was compared using the likelihood ratio test. Comparison of non-nested models was done using the Akaike Information Criterion (AIC).

Table 4.1. Different models evaluated

Model	Clearance rate, $\mu(a)$	Infection rate, $\lambda(a,t)$	Detectability, $s(a)$	AIC*	p†
M1	μ	λ	$s(a)=s$	2709.6	3
M2	μ	λ	$\text{logit}\left[s(a)\right]=$ $s_0+s_1(a-\bar{a})$	2599.3	4
M3	μ	$\lambda_i,\ i=1,2,...,6$	$s(a)=s$	2644.3	8
M4	$\mu_0 e^{\mu_1 a}$	λ	$s(a)=s$	2620.5	4
M5	μ	$\beta_0 e^{\beta_1 a}$	$s(a)=s$	2643.9	4
M6	μ	$\lambda_i,\ i=1,2,...,6$	$\text{logit}\left[s(a)\right]=$ $s_0+s_1(a-\bar{a})$	2528.3	9
M7	$\mu_0 e^{\mu_1 a}$	λ	$\text{logit}\left[s(a)\right]=$ $s_0+s_1(a-\bar{a})$	2601.0	5
M8	μ	$\beta_0 e^{\beta_1 a}$	$\text{logit}\left[s(a)\right]=$ $s_0+s_1(a-\bar{a})$	2601.3	5
M9	$\mu_0 e^{\mu_1 a}$	$\lambda_i,\ i=1,2,...,6$	$s(a)=s$	2548.3	9
M10	$\mu_0 e^{\mu_1 a}$	$\beta_0 e^{\beta_1 a}$	$s(a)=s$	2621.3	5
M11	$\mu_0 e^{\mu_1 a}$	$\lambda_i,\ i=1,2,...,6$	$\text{logit}\left[s(a)\right]=$ $s_0+s_1(a-\bar{a})$	2529.5	10
M12	$\mu_0 e^{\mu_1 a}$	$\beta_0 e^{\beta_1 a}$	$\text{logit}\left[s(a)\right]=$ $s_0+s_1(a-\bar{a})$	2602.7	6

*Comparison of the fit of models with different parameter specifications using the Akaike Information Criterion (AIC). A lower value of AIC indicates a better fit.

† p is the number of parameters estimated.

4.4 Results

An exploratory analysis was done to assess the number of new infections gained and the proportion of existing infections that were lost. An infection present in survey X but absent in the consecutive survey X+1 was considered as "Loss" (+, -), while "Gain" (- +) was noted when an infection was present in survey X but absent in the previous survey X-1. The total number of infections gained was calculated over all possible consecutive surveys. The total number of infections lost at survey X+1 among those initially infected at survey X was also calculated over all possible consecutive surveys. The results were summarized by age group and by survey intervals. The number of new infections acquired increased to the age of 5-9 years and then dropped for older age groups. An increase in age of the proportion of infections lost was also observed (Fig. 4.1a). However when these quantities were assessed by season, the proportion of infections lost remained fairly constant throughout the year, while there was a seasonal effect on the number of infections gained with the maximum values observed during the first two surveys, reflecting the high transmission rate during the wet season (Fig. 4.1b).

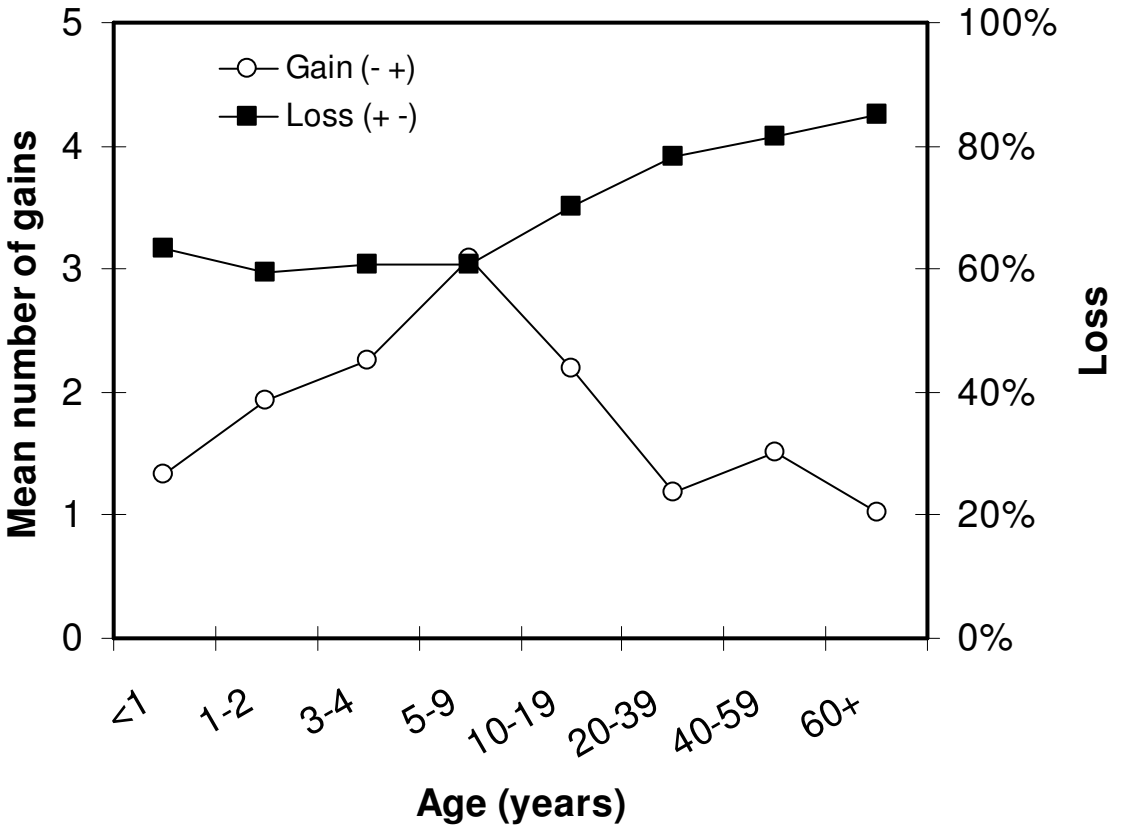


FIGURE 4.1 (a): Mean number of newly acquired infections (Gains), and proportion of infections lost (Loss) among those initially infected, per person-interval by age group.

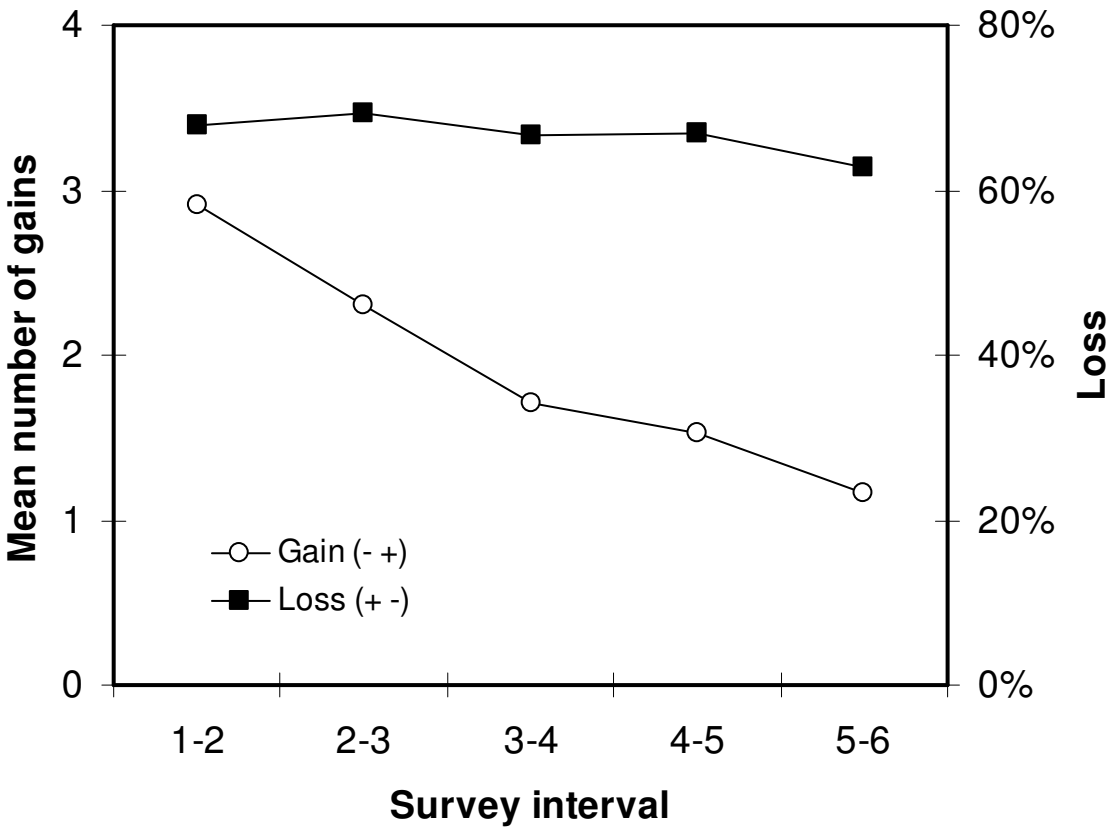


FIGURE 4.1 (b): Mean number of newly acquired infections (Gains), and proportion of infections lost (Loss) among those initially infected, per person by survey interval.

We considered a total of 12 models (Table 4.1). M1 is the model we reported previously (Sama *et al*., 2005) in which it was assumed that there was no age or seasonal variation in the parameters. Likelihood ratio tests indicated that all four of extended models M2, M3, M4 and M5 in which age or seasonal effects were added, demonstrated better fits than M1 indicating that both age and seasonal variation are important (Fig. 4.2).

All models where detectability was expressed as a function of age (M2, M6, M7, M8, M11, M12), estimated a significant age effect, and all the estimates of s_1 from these models are similar, indicating decreases in the detectability with age (Table 4.2). The likelihood ratio tests comparing each of these models with the corresponding reduced models with age- independent detectability all showed statistically significant differences (Fig. 4.2).

When the infection rate, λ, was assumed constant, its value (of about 17 infections gained per annum) was rather insensitive to the other parameters in the model and somewhat higher than the rate of acquisition of apparent new infections in the raw data (Fig. 4.1). Although model M5 suggested that there was a strong decrease of infection rate with age, adjustment for age dependence in detectability gave an improvement in fit, and the best model (by Akaike's criterion) among those where the infection rate was allowed to vary with age was M8 (Table 4.1), in which there is a slight tendency for the infection rate to decrease with age (Table 4.2). This reflects a rather strong correlation of -0.78 between the estimates of the parameters λ_1 and s_1 in model M8, indicative of moderate collinearity.

Among models where the clearance rate varies by age, the best fit was M11 (Table 4.1), in which there is a slight increase in clearance rate with age (Table 4.2). This increase was much less than that in model M4 in which both infection rate and detectability were assumed constant. The lower estimate of μ_1 in M11 reflects a rather strong positive correlation of 0.87 between the estimates of μ_1 and s_1 in this model.

Neither the age trend in infection nor clearance rates was statistically significant (as measured by likelihood ratio tests comparing models M8 and M11 with their reduced models, Fig. 4.2). This is because, when age-dependence in the detectability was allowed for, no significant improvement in fit could be achieved by including age effects in either λ or μ (compare M2 with M7 and M8 in Fig. 4.2; or M6 with M11) or in both of these parameters (compare M12 with M2, M7 or M8). Similarly, including age dependence in both λ and μ (M10) did not improve the fit significantly compared with a model with age dependence in only μ (M4).

The overall best fitting model by Akaike's criterion is M6, which included seasonal variation in λ as well as age dependence in *s*. Indeed, all models with seasonal patterns in the infection rate fitted better than reduced models with constant λ. The highest two-month specific infection rate estimated from model M6 was 31 new infections per year, during the period June to August. The rates gradually reduce as we move towards the dry season (the low transmission period) and peaks again as we enter the wet season. This pattern is the same for all the models (M3, M6, M9, and M11) where the infection rates were allowed to vary seasonally with similar values estimated for each of the six different infection rates λ_i in the different models (Table 4.2).

The estimates of the parameters describing the detection process, and those measuring seasonality in infection rates were insensitive to the other variables in the model (Table 4.2).

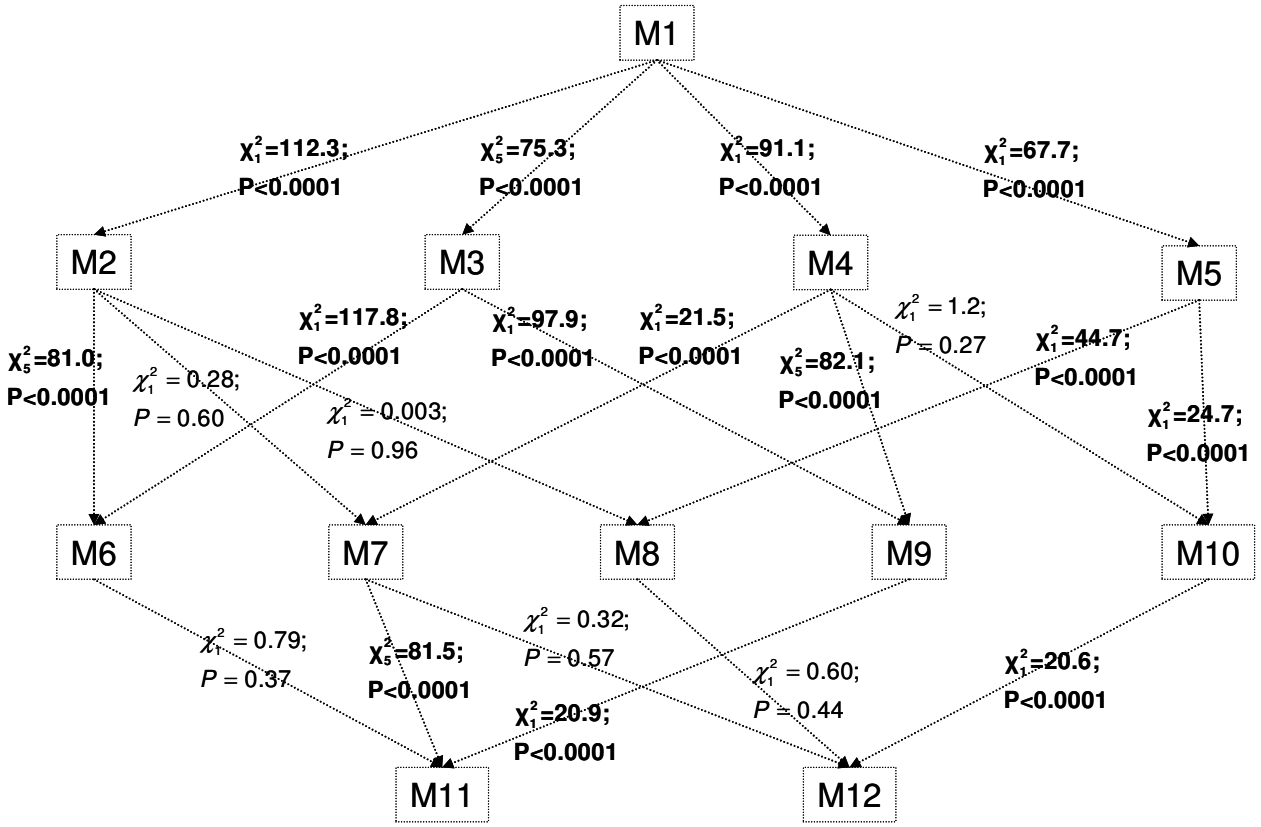


FIGURE 4.2. Flow-chart of nested models with likelihood ratio statistics (and P-values) comparing the models. M1⋯▸M2: M2 is nested within M1

TABLE 4.2. Parameter estimates from the different models in Table 4.1.

Para-meters	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
μ	2.4 (2.2,2.6)	2.3 (1.9,2.6)	2.6 (2.3,3.0)		2.4 (2.0,2.7)	2.6 (2.2,2.9)		2.3 (2.1,2.5)				
μ_0				1.8 (1.5,2.1)			2.3 (1.9,2.8)		2.0 (1.7,2.5)	1.9 (1.5,2.2)	2.4 (2.0,2.8)	2.4 (2.0,2.9)
μ_1				0.020 (0.016,0.024)			-0.003 (-0.014,0.008)		0.020 (0.016,0.024)	0.017 (0.011,0.023)	0.004 (-0.005,0.013)	-0.007 (-0.024,0.011)
λ	16.3 (14.8,17.8)	17.5 (15.7,19.3)		17.1 (15.4,18.8)			17.5 (15.7,19.2)					
β_0					21.1 (18.6,23.5)			17.5 (16.2,18.8)		18.0 (15.4,20.5)		18.0 (15.4,20.5)
β_1					-0.015 (-0.019,-0.011)			-0.0002 (-0.007,0.006)		-0.003 (-0.010,0.0028)		-0.003 (-0.014,0.008)
s	0.47 (0.42,0.51)		0.49 (0.44,0.54)	0.46 (0.42,0.51)	0.46 (0.41,0.51)				0.48 (0.43,0.53)	0.47 (0.42,0.52)		
s_0		-0.40 (-0.61,-0.20)				-0.33 (-0.53,-0.14	-0.44 (-0.68,-0.20)	-0.40 (-0.58,-0.23)			-0.29 (-0.50,-0.08)	-0.45 (-0.69,-0.21)
s_1		-0.032 (-0.038,-0.026)				-0.035 (-0.040,-0.029)	-0.036 (-0.049,-0.022)	-0.032 (-0.041,-0.023)			-0.030 (-0.042,-0.018)	-0.036 (-0.050,-0.022)
λ_1			25.6 (20.7,30.4)			31.1 (25.0,37.2)			28.6 (23.2,34.1)		31.2 (25.1,37.2)	
λ_2			24.6 (18.8,30.5)			22.9 (16.7,29.2)			23.5 (17.6,29.4)		23.0 (16.7,29.2)	
λ_3			13.0 (7.9,18.0)			13.8 (8.5,19.1)			13.4 (8.3,18.5)		13.8 (8.4,19.1)	
λ_4			13.0 (8.5,17.5)			13.1 (8.4,17.8)			13.4 (8.7,18.0)		13.2 (8.4,17.9)	
λ_5			4.6 (1.2,8.0)			5.4 (1.8,9.0)			4.5 (0.9,8.0)		5.3 (1.6,8.9)	
λ_6			13.2 (2.5,24.0)			21.6 (7.8,35.5)			19.2 (5.6,33.0)		21.0 (7.2,34.9)	

The graphical fits are consistent with the results of formal evaluation of the 12 models by AIC. The frequency of the 63 observed sequences are closer to predictions when the overall best model, M6, is used than when M1 is used (Fig. 4.3). This is closely followed by the fits from model M11 (which is not significantly different from M6), and the fits from M3 and M9. The largest residual using M6 corresponded to sequence (010000) which occurred 147 times (expected frequency 107.8). Three other sequences had residual frequencies >20 or <-20. All of these four sequences represented infections that were observed at only one survey, thus suggesting that the model fits least well for infections that persist for only a short time. When we assessed the ratio between the observed and expected frequencies, sequence (111111) which occurred 19 times, but with an expected frequency of only 0.9, gave a very poor fit. A possible explanation for these discrepancies is that the true distribution of the durations of infections has a greater variance than is assumed by the exponential model.

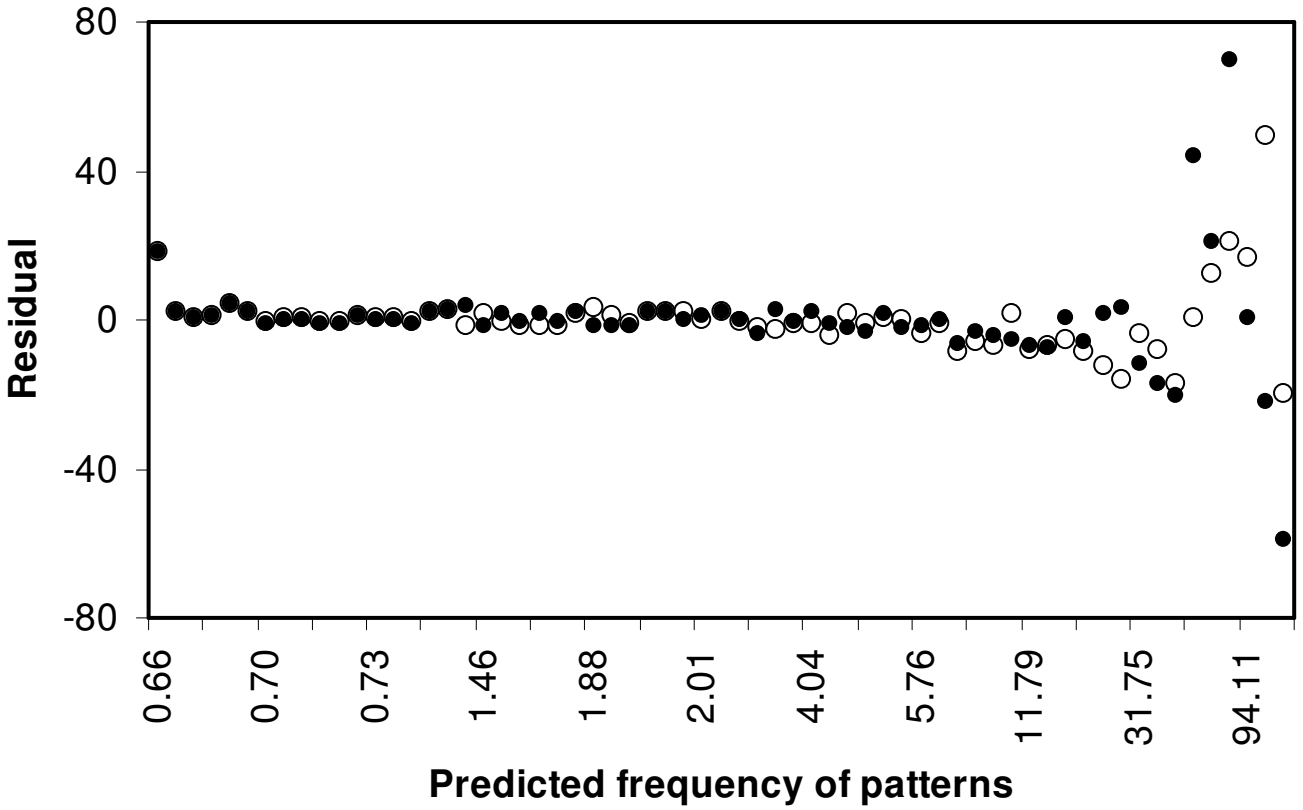


FIGURE 4.3. Plot of the difference between the observed and predicted frequency of the 63 patterns against the predicted frequency.

●: model M1, (identical to the model of Sama et al, (2005));

○: model M6, (best fit).

Though the models are fitted to the frequency of the observed sequences and not to the observed mean multiplicity, equations 3 and 4 nevertheless predict the true numbers of distinct genotypes, as well as the numbers of distinct genotypes observed in each sample (observed mean multiplicity). There is a tendency for the observed mean multiplicity of infection to rise and then fall again as age increases. The plots of the expected multiplicity (equation 4) against age are indistinguishable for models M4, M5, M8, and M10, while that of M6 is also similar to M9, and these plots tend to follow the pattern in the observed data. The plots for the expected true multiplicity (equation 3) for models M4, M5, and M10 are indistinguishable, but differ from that of M8, while that of M6 differs from that of M9. There is a greater tendency to follow the pattern in the observed data with the plots of the expected true multiplicity using models M4, M5, M9 and M10, than for M8 and M6 that were reported as the best models above (see Fig. 4.4). Because of this discrepancy, we were concerned that the conclusion obtained for the clearance rate may be an artifact of the type of function considered for the clearance rate. We attempted a more flexible function by assuming that

$$\mu(a) = \mu_0 e^{\mu_1 a + \mu_2 a^2}, \text{ where } \mu_0 > 0, \ -\infty < \mu_1, \mu_2 < \infty$$

This however gave no significant improvement of fit compared with any of models M4, M7, M10, M9, M11, and M12.

We also attempted the following two logistic functions for the clearance rate.

$$\mu(a) = \mu_0 + (\mu_1 - \mu_0)\frac{1}{1 + (\frac{a^*}{a})^k}, \quad \text{where} \quad \mu_0, \mu_1, a^* > 0, \ k < 0.$$

$$\mu(a) = \frac{\mu_0 \mu_1}{\mu_0 + (\mu_1 - \mu_0)e^{-\delta a}}, \quad \text{where} \quad \mu_0, \mu_1, \delta > 0.$$

However the parameters in all the models fitted by substituting the clearance rate in models (M4, M7, M9, M10, M11, and M12) with the above two functional forms were not identifiable.

Similarly we attempted the following flexible functional form for the infection rate in models M5, M8, M10, and M12;

$\lambda(a) = \beta_0 e^{\beta_1 a + \beta_2 a^2}$, where $\beta_0 > 0,\ -\infty < \beta_1, \beta_2 < \infty$

The results were also not significantly different from the initial models.

We also attempted the following form for the infection rate where it is allowed to vary both by age and season,

$\lambda(a) = \lambda_i e^{\lambda a}$, where $-\infty < \lambda < \infty,\ \lambda_i > 0,\ i = 1, 2, ..., 6$

and evaluated models M1, M2, M4, M7 with this expression substituted for the infection rate. The model obtained from M1 by replacing the infection rate with the above expression gave better results (AIC = 2572.0) than M3 and M5 where the infection rates were allowed to vary only by age or by season. Similarly the model obtained from M4 (AIC = 2547.4) also gave a better fit than M10 where the infection were allowed to vary only by age; and gave a similar fit to M9 where the infection rate was allowed to vary only by season. However the parameters in the models obtained from M2 and M7 (which represents the most interesting case since the detectability is allowed to vary by age) were not identifiable. The models obtained from M1 and M4 indicated a seasonal variation in infection rates with a slow decrease with age and that obtained from M4 also indicated a significant increase in clearance rate with age.

The best model by the AIC criterion remained M6.

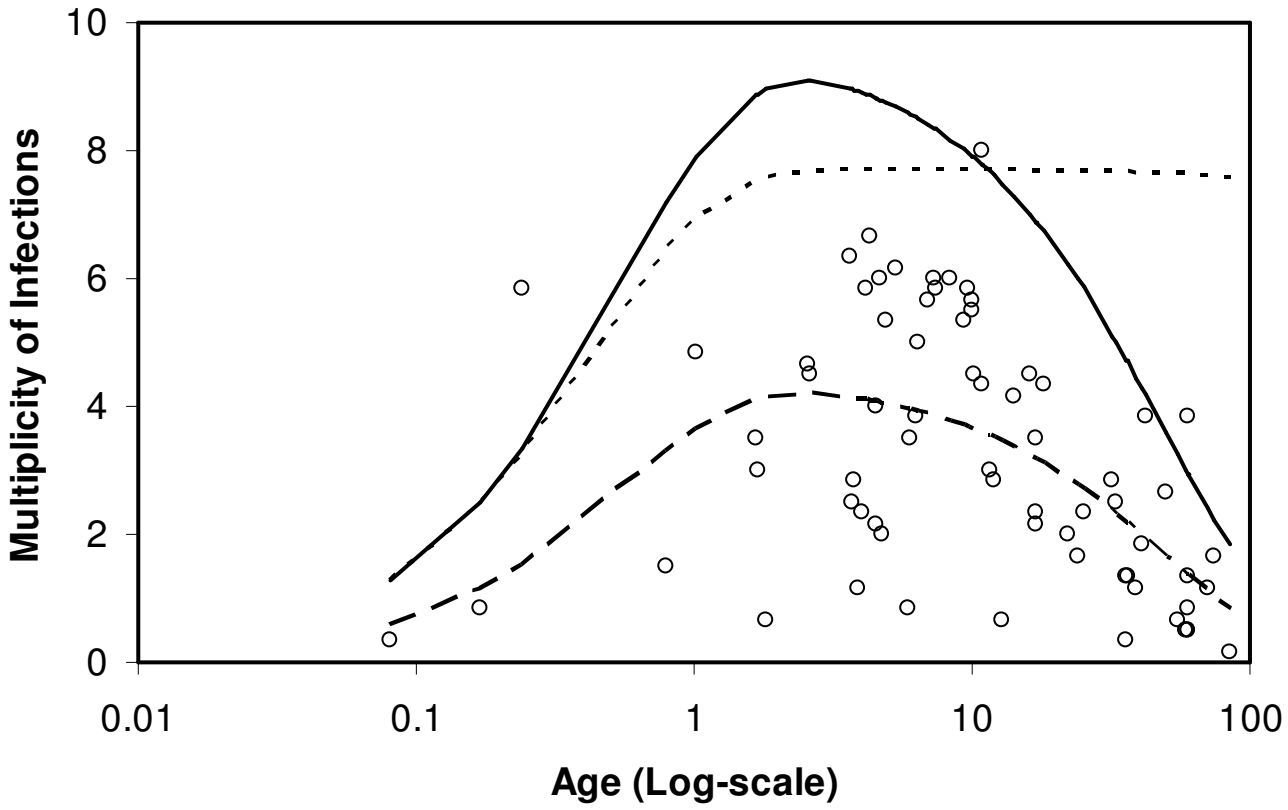


FIGURE 4.4(a): Mean Multiplicity of infections at baseline.

○: observed mean multiplicity;

— —: fit to the observed mean multiplicity ($\tilde{n}(a,0)$) using model M5;

——: predicted true mean multiplicity ($n(a,0)$) using model M5;

- - -: predicted true mean multiplicity ($n(a,0)$) using model M8;

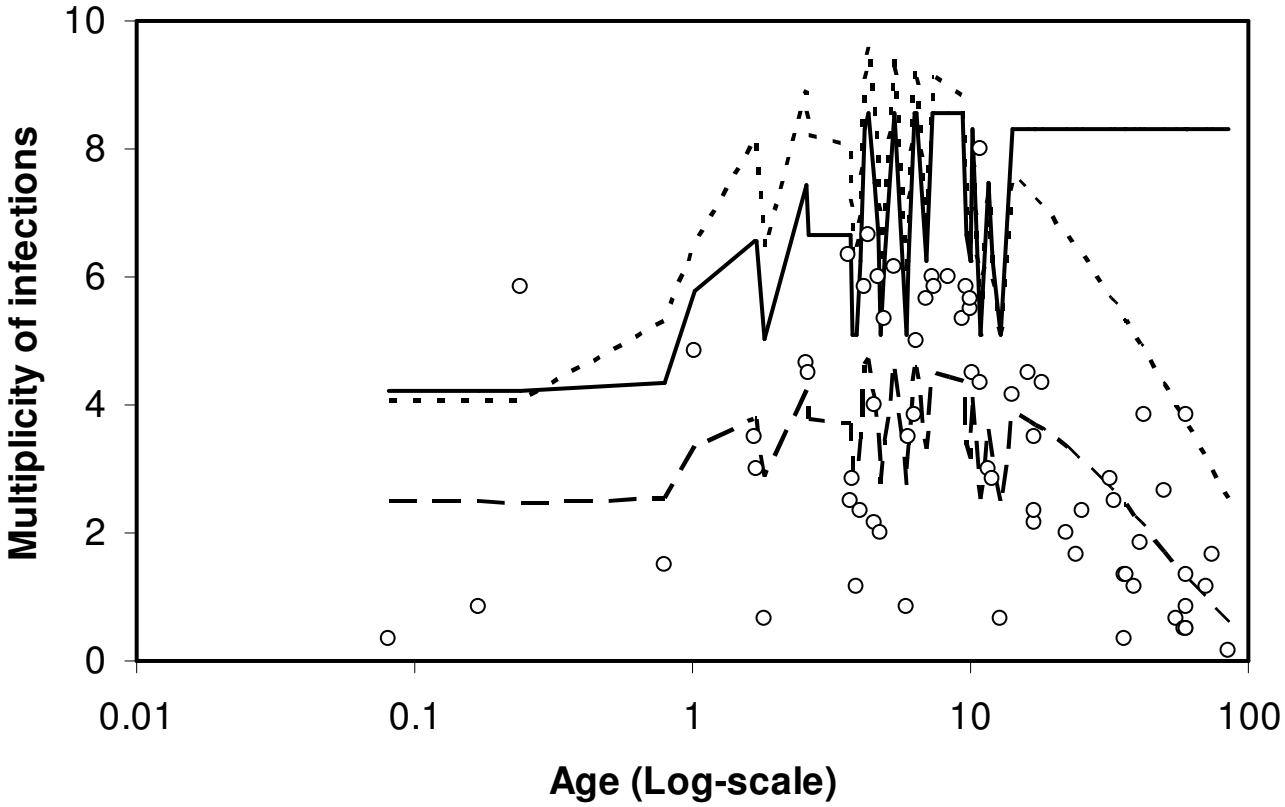


FIGURE 4.4(b): Mean Multiplicity of infections at baseline.

○: observed mean multiplicity;

– – : fit to the observed mean multiplicity ($\tilde{n}(a,0)$) using model M6;

——: predicted true mean multiplicity ($n(a,0)$) using model M6;

- - -: predicted true mean multiplicity ($n(a,0)$) using model M9;

As in our previous work (Sama *et al*., 2005) our analysis assumed that re-infection with a given genotype is a rare event. In other to assess the impact of this assumption, we excluded the most frequent genotype from the dataset, and repeated the analysis of models M1-M12. The infection and clearance rates obtained from this modification were very similar to those from the complete dataset, while detectability estimates were rather lower, (e.g. reduced to $s = 0.38$ in model M1). The AIC results were similar to those in the full dataset.

4.5 Discussion

An important gap in knowledge of the epidemiology of malaria is in the effect of acquired immunity on the duration of infection. The rate at which *Plasmodium falciparum* infections are cleared is often thought to be highly affected by acquired immunity but the empirical evidence for this is weak. Most studies on this are based on data obtained from light microscopy, a technique which cannot distinguish between persisting and new infections. The failure to detect low density infections and the inability to distinguish the concurrent infections within a host at a given time also potentially lead to bias in estimates of infection and recovery rates from light microscopy data (for instance as in the studies of (Bekessy *et al.*, 1976; Kitua *et al*., 1996; Sama *et al*., 2004)). A single continuously infected individual, appears by microscopy neither to contribute to the pool of new infections nor to that of infections that have been cleared. Durations estimated from microscopy data will also highly depend on whether the data were derived from repeated cross-sectional studies or from longitudinal studies (Sama *et al*., 2004).

With the availability of PCR genotype specific transition data, it is now apparent that continuously infected individuals are experiencing repeated super-infection and clearance of specific genotypes, and PCR analyses can therefore give more reliable estimates of infection and recovery rates. The estimate of the overall infection rate in our study, as might have been expected, is much higher than in studies using microscopy data. Estimates of duration of infection from PCR, which are for a single monoclonal infections (Sama *et al*., 2005; Smith *et al*. 1999c; Smith and Vounatsou, 2003), tend to be shorter than those from microscopy data (Bekessy *et al*., 1976; Kitua *et al*., 1996; Macdonald and Göckel, 1964; Sama *et al*., 2004) which are for polyclonal combinations of old and new infections.

Our previous model for *P. falciparum* genotype data from a study in Northern Ghana (Sama *et al*., 2005) assumed that the transition rates and the detection process were homogeneous. The present analysis extends this to allow different age dependences in the clearance rate, infection rate, and detectability and thus for age-dependence in multiplicity of infection. Some of the models also allow for seasonal variation in the infection rates. Age-dependence in multiplicity is generally found in highly endemic areas (Ntoumi *et al*., 1995; Smith *et al*., 1999a) but its relationship to acquired immunity has been unclear. No other PCR study has analyzed transition rates with a complete age representation of the population.

We find that there is a decrease in detectability with increasing age. Since the parasite densities decrease with age and it is expected that a genotype is less likely to be detected if the density is low, the decrease with age in detectability was anticipated. Detectability varied from about 60 percent in younger individuals to 10 percent in adults.

Immunity is known to increase with age and it has often been assumed that duration of infection must decrease as immunity is acquired. Some studies have found that infections are of very short duration in infants (Walton, 1947; Kitua *et al*., 1996; Smith *et al*., 1999c) and one study in young children estimated that duration increases with age (Smith and Vounatsou, 2003). In our analysis, the models that did not include age dependence in detectability estimated significant increases in the clearance rate with age. However our best fitting models allow for age dependence in detectability and found no residual effect of age on clearance rates, but we have to treat this conclusion with caution because the parameter estimates for the age effect on clearance are rather strongly correlated with those for age dependence in detectability.

The estimated duration of infection was not greatly modified by allowing for seasonal variation in the infection process, although infections were found to be highly seasonal. The use of six different parameters for the infection rates allowed us to obtain season-specific estimates for the infection rates with the highest infection rate estimated from our best model, M6, is 31 new infections per year, during the period June to August which is the peak transmission season in this area (Baird *et al.,* 2002; Binka *et al*., 1994). This implies that individuals may acquire as many as 5 new infections during this period.

In separate models (M5, M8, M10, M12), we allowed the infection rate to vary only by age, the best model among these was M8 which indicated a decrease in infection rate with age, but this was not statistically significant (see Fig. 4.2 for comparison of M8 and M2). Some studies have suggested that infection rates in children are higher than in older people (Rogier and Trape, 1995), as would be expected if immunity to infection is acquired. However the attractiveness of the human host to mosquitoes in general

increases with host size (Port *et al.*, 1980), which should lead to an increase in infection rates with age (Smith *et al*., 2004). The overall best model among the 12 models we considered was M6 which included season- but not age-dependence in infection rates.

Our results should be interpreted cautiously because none of our models allows simultaneously for seasonal variation in infection rates or detectability, variation of infection rates with age, or variation of clearance rate with age. Further work needs to be done to capture all these variations and this will need a substantial amount of information, covering a wide range of age groups, from different endemic, geographic, and cultural settings. The present models also ignore heterogeneity between genotypes, which may also be important.

We assumed age to be the only factor leading to heterogeneity in infection rates across the population. However, on 26% of visits respondents indicated that they slept under bednets. It is likely that these nets moderated the infection rate but had little effect on the clearance rate. Patterns of treatment for febrile illness in the KND are complex (Owusu-Agyei *et al*., submitted) often involving ineffective medications or inappropriate dosing. We did not record treatments used by the participants during the study or attempt to model effects of treatment but it is clear that the use of anti-malarial treatments will increase the clearance rate and so may have biased downwards our estimates of duration.

A further improvement would be to directly incorporate information about parasite densities. The clearest effect of acquired immunity is in reducing the mean parasite density, and this must account for a large element of the decrease in detectability with age. The question of how to incorporate the extra information contributed by the parasite density while avoiding identifiability problems deserves further attention. There is likely

to be a change in both detectability and clearance rate if this information is well accounted for.

4.6 Acknowledgements

We thank the staff of the Navrongo Research Health Centre involved in the field work and the villagers for their co-operation. We also thank Beatrice Glinz for her contribution to the genotyping. 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.

CHAPTER 5

Comparison of PCR-RFLP and GeneScan–based genotyping for analyzing infection dynamics of *Plasmodium falciparum*

Nicole Falk[1], Nicolas Maire[1], Wilson Sama[1], Seth Owusu-Agyei[2], Tom Smith[1], Hans-Peter Beck[1], & Ingrid Felger[1].

[1]Swiss Tropical Institute, Basel, Switzerland.

[2]Kintampo Health Research Centre, Kintampo, Ghana

This paper has been accepted for publication in the
American Journal of Tropical Medicine and Hygiene

5.1 Abstract

Parameters describing the infection dynamics of *Plasmodium falciparum* are important determinants of the potential impact of interventions and are potential outcome measurements for malaria intervention trials. Low parasite densities, periodic sequestration of parasites, and the presence of multiple concurrent infections make it essential to use molecular techniques to estimate the force of infection and duration of infections in endemic areas. We now compare two approaches for tracking individual genotypes of the highly polymorphic merozoite surface protein 2 (i) fluorescence-labeled PCR and GeneScan-sizing, and (ii) restriction-fragment-length-polymorphism (RFLP). We analyze samples from a longitudinal field study in Ghana and use statistical approaches that allow for imperfect detectability. The two methods gave broadly similar estimates of parasite dynamics, but GeneScan is more precise and can achieve a higher throughput. The analysis of parasite dynamics indicated an average duration of infection of 210 days by GeneScan versus 152 days by PCR-RFLP in the study population in Kassena-Nankana, Northern Ghana. This reflects the good performance of GeneScan-based genotyping for studies of parasite infection dynamics.

5.2 Introduction

Plasmodium falciparum parasite densities fluctuate over time in the course of an infection and the late stages of the parasite cytoadhere to endothelial receptors and sequester in deep tissues, leading to a 48 hour cycle in the appearance of parasites in the peripheral circulation. These features of malaria parasite biology complicate the analysis of infection dynamics. This background of persisting but transiently detectible parasitaemia makes it inappropriate to use standard microscopy techniques to determine the rates of infection, or the clearance rates of infections in the absence of interventions. These are essential parameters for understanding the transmission dynamics of the parasite (Smith and Vounatsou, 2003).

Allelic discrimination is possible by molecular means and multiple concurrent infections are usually found in blood samples from areas endemic for malaria (Ntoumi *et al.*, 1995 ; Felger *et al.*, 1999a; Smith *et al.*, 1999d). Highly polymorphic marker genes amplified by PCR can be used to track individual clones in longitudinal sample sets. Despite the lower detection limit of PCR relative to microscopy, some clones still seem to disappear but are in fact persisting, and genotyping often detects clones recurring periodically at intervals of 48 hours (Farnert *et al.*, 1997; Bruce *et al.*, 2000a). We have previously analyzed infection dynamics of individual clones as determined by PCR-RFLP using transition models based on the relative frequencies of different patterns of infection defined by presence/absence at successive surveys (Smith *et al.*, 1999b; Sama *et al.*, 2005). These models take into account the imperfect detectability of clones sequestered at the time of sampling.

We now compare two molecular techniques for identifying individual clones. The first is restriction fragment length polymorphism of PCR fragments (PCR-RFLP) has been used in a number of studies to track parasite clones longitudinally. The marker gene merozoite surface protein 2 (*msp2*) is highly polymorphic due to intragenic repeats, with more than 50 genotypes identified by RFLP in each of its two allelic families 3D7 and Fc27 (Felger *et al.*,

1994; Felger *et al.*, 1999a). The pattern of restriction fragments clearly identifies individual clones, and particular patterns are recognized also within mixed infections. However, in samples with more than 5 concurrent genotypes, the superimposed patterns are increasingly more difficult to analyze. The second technique is based on sizing PCR fragments by an automated sequencer using the GeneScan™ program. This technique uses fluorescent labeled PCR primers specific for the *msp2* allelic families. FC27- and 3D7-type *msp2* PCR fragments are identified by the two fluorescent markers 6-Fam and VIC, respectively. The use of a different dye for each allelic family increases resolution of fragment sizing. GeneScan analysis software uses a size standard added to each sample after PCR to create an internal calibration curve to determine the size of each PCR fragment. In comparison with PCR-RFLP, this technique increases throughput and avoids subjectivity in analyzing the readout.

We report analyses of *P. falciparum* infections in 100 individuals of all ages from the Kassena-Nankana District of Ghana sampled at 2-monthly intervals over one year. We compare the performance of genotyping by PCR-RFLP versus GeneScan for the analysis of the force of infection (clonal acquisition rate) and duration of individual infections based on the patterns of persistence of individual genotypes.

5.3 Materials and Methods

5.3.1 Study site and population.

The population of Kassena-Nankana District (KND) in northern Ghana is plagued by high infant and maternal mortality rates (Binka *et al.*, 1994). In KND malaria transmission is seasonal, but even in the absence of any rain, widespread transmission of malaria continues. For a molecular epidemiological survey of *P. falciparum* multiplicity of infection and infections dynamics among asymptomatic inhabitants of a holoendemic malarious area, a cluster sample of the KND population was drawn by selecting 16 index compounds at random

from the 14 000 within the district. From each index compound, two people in each of the following age categories were selected: <1, 1-2, 3-4, 5-9, 10-19, 20-39, 40-59, 60+. Blood samples were collected on DNA ISOCode^{TM} Stix (Schleicher & Schuell) in intervals of two months resulting in a total of 6 samples per participant (R1-R2-R3-R4-R5-R6). Blood samples of the first time point (R1) were collected in June/July 2000 and genotyping results of R1 were presented previously (Owusu-Agyei *et al.*, 2002). Informed consent was obtained from participants by signature or thumbprint in the presence of a witness. Ethical clearance for this study was obtained from the Ghana Health Service Ethics Committee. For the present study on comparison of two genotyping methods, a subset of the 1848 samples collected were used, amounting to 600 samples deriving from 100 individuals.

5.3.2 DNA isolation and genotyping.

10 µl whole blood were dotted onto ISOCode^{TM} Stix. Irrespective of microscopy results, 6 samples from each of 100 individuals were screened for presence of *P.falciparum* by PCR. Processing of Stix, PCR conditions and RFLP procedures have been described in detail (Felger *et al.*, 1999a; Felger and Beck, 2002). In brief, primary and nested PCR were performed for RFLP analysis using primer pair S2/3 (5'-GAAGGTAATTAAAACATTGTC-3'/ 5'-GAGGGATGTTGCTGCTCCACAG-3') and S1/4 (5'-GAGTATAAGGAGAAGTATG-3'/5'-CTAGAACCATGCATATGTCC-3'), respectively, followed by restriction digest with *HinfI* and *DdeI* as described elsewhere (Felger *et al.*, 1999a; Felger and Beck, 2002).

For GeneScan analysis, 2 µl of primary PCR product were amplified in nested PCR with the fluorescent-labeled family-specific primer M5 (FC27-specific: 5'-6-FAM-GCATTGCCAGAACTTGAA-3') or N5 (3D7-specific: 5'-VIC-CTGAAGAGGTACTGGTAGA-3') at a concentration of 100 nM and 200 nM, respectively. The non-fluorescent-labeled forward primer S_{Tail} (5'-TTATAATATGAGTATAAGGAGAA-

3') was modified at the 5' end by adding a 7 bp tail in order to avoid non-template-directed addition of a single nucleotide to the 3' end of a blunt-end double-stranded DNA ("plus-A-artefact") (Brownstein *et al.*, 1996). The cycle conditions were 5 min at 94°C followed by 30 cycles of 30 sec at 94°C, 1 min at 50°C and 1 min at 70°C and a final elongation for 7 min at 70°C.

0.5 µl nPCR product were combined with 10 µl ROX-labelled size standard (diluted 1:10 with H_2O to minimize pipetting errors). Samples were dried and sent to the Genomics Core Laboratory of MRC Clinical Science Centre in London. Highly deionized formamide was added and after denaturation samples were analyzed on an ABI PRISM 3700 genetic analyzer.

5.3.3 Determination of detection limits.

Genomic DNA was isolated from *P. falciparum* in vitro cultures (FC27 and 3D7 strains) as described previously (Beck, 2002). DNA concentration was determined photometrically. Tenfold dilutions of genomic DNA were amplified in PCR reactions containing 1000, 100, 50, 10, 5, 1, and 0.1 genomes of either 3D7 or FC27 DNA, or a mixture of both.

5.3.4 Data analysis.

An in-house generated computer program was used to process the output of the GeneScan analyzer. The main tasks of this program were:

(1) Cut-off determination. A sample-specific cut-off was used to separate real signals from noise and to allow for variability between GeneScan runs. The cut-off was established by GeneScan analysis of sequenced *msp2* reference alleles (cloned *msp2* fragments or single clone infections). The presence of a single template per reaction made it possible to determine background fluorescence levels to 300 arbitrary fluorescence units. A sample-specific cut-off was determined by multiplying the arithmetic mean of peak heights of the size standard

signals per sample by a constant (the empirically chosen cut-off of 300 units devided by the mean peak hight of size standard peaks of all samples). Since with our primers none of the *msp2* sequences available at Genbank would give rise to a PCR fragment of less than 216 bp, peaks with a measured size of less than 200 bp were not considered.

(2) Elimination of bleeding and "plus-A-artifacts". Spectral overlap of the fluorescent dyes labelling the family-specific primers caused bleeding in case of strong signals caused by PCR fragments present at very high concentrations (>5000 fluorescent units). Although the dyes emit light at different wavelengths, some overlap exists despite using a GeneScan software matrix file to remove spectral overlaps.

Taq polymerase has terminal deoxynucleotidyl transferase activity to add an extra nucleotide, usually adenine ("plus-A"), at the 3' end of PCR products. This results in two populations of amplified products with a size difference of one nucleotide. Despite the use of a tailed primer that supports "plus-A" addition, we observed in case of high intensity peaks also a small peak about 1 nucleotide prior to the actual peak. Our software eliminated peaks due to bleeding and "plus-A-artifact".

(3) Elimination of PCR artifacts. Allele-specific PCR artifacts were detectable when particular alleles were present at very high concentration. Such artifacts are likely due to the intragenic repeats of *msp2* alleles, which can facilitate abberant annealing of an incompletely synthesized strand to the repeat region. We determined allele-specific PCR artifacts by analyzing cloned *msp2* alleles at high DNA concentrations. Artifacts were omitted from subsequent analyses.

(4) Genotype calling. For analysis of longitudinal sets of samples, a persisting genotype must be accurately identified in sequential blood samples. There were slight variations among repeated size-determinations of identical fragments depending on the concentration of the amplified fragment. To allow for inaccuracies in size determination, peaks were assigned to

size bins with a width of 2.4 bp. Since a coding region is genotyped, fragment sizes must differ by multiples of 3 bp.

5.3.5 Statistical analysis.

For comparing the two genotyping methods, SAS statistical software Release 8.2 (SAS Institute Inc., Cary, NC, USA) was used. Infection dynamics were analyzed by calculating the frequency of gains, losses and persistence of infecting clones. An infection present in survey at time t, but not seen in the subsequent survey t+1 was considered as "loss" (+ −), whereas "gain" (− +) was noted when an infection was observed in round t but not in the previous round t−1. Where infections were observed in consecutive surveys this was recorded as "persistence".

We analyzed the infection dynamics using methods that allow for imperfect detection using the method of Smith *et al.* (1999b). In further analysis, we fitted an immigration-death model to the full sequences of six observations (Sama *et al.*, 2005). This provided estimates of the rate of new infection, λ, clearance rate, μ, and detectability.

5.4 Results

5.4.1 Limit of Detection and Evaluation of GeneScan.

A serial dilutions of genomic DNA from parasite culture strains FC27 and 3D7 were performed to determine the detection limit of GeneScan-based *msp2* genotyping. After primary PCR, nested PCR was performed either independently for each primer set (3D7 simplex PCR and FC27 simplex PCR) or with a combination of FC27- and 3D7-specific primers (duplex PCR). While both the FC27-and 3D7-specific simplex PCR detected one parasite genome per reaction, performing a duplex PCR with equal amounts of both templates reduced the detection limit to 5 genomes per reaction **(Table 5)**. This sensitivity was considered sufficient, and considerably reduced costs justified duplex PCR for field samples.

The effect of excess DNA of the alternative allelic family was assessed for duplex PCR. 3D7-specific amplification remained unaffected by the presence of 1000 FC27 genomes. But FC27-specific amplification detected only 10 genomes in the presence of 1000 3D7 genomes. This difference is probably due to the smaller amplicon size of 3D7 (267 bp) versus FC27 (358 bp) (accession numbers M28891, J03828).

Prior to application to field samples, the technique was validated on a panel of sequenced reference alleles (data not shown). The standard deviation (SD) of repeated sizing of the same fragment was only about 0.06 and 0.17 nucleotides for FC27 and 3D7 DNA, respectively. For the most frequent allele in the Ghanaian field samples a standard deviation of SD= 0.14 was calculated.

When comparing all *msp2* sequences submitted to Genbank until 2004, we detected two genotype pairs of the same allelic family sharing the same amplicon size (accession numbers U07001/U16842, AY534506/AF010461). Thus GeneScan cannot discriminate these alleles. This represents the limitation of our method.

Table 5: Sensitivity of detection of *msp2* PCR fragments by agarose gel electrophoresis compared to GeneScan (GS).

Number of templates/reaction *)	1000		100		50		10		5		1		0.1	
Detection technique	Gel	GS	Gel	GS	Gel	GS	Gel	GS	Gel	GS	Gel	GS	Gel	GS
Fc27 amplified in simplex PCR	+	+	+	+	nd	nd	+	+	+	+	+	+	-	-
3D7 amplified in simplex PCR	+	+	+	+	nd	nd	+	+	+	+	-	+	-	-
Fc27 amplified in duplex PCR	+	+	+	+	nd	nd	+	+	+	+	-	-	-	-
3D7 amplified in duplex PCR	+	+	+	+	nd	nd	+	+	+	+	-	-	-	-
Fc27 in presence of 1000 3D7 copies	+	+	+	+	+	+	-	+	-	-	-	-	-	-
3D7 in presence of 1000 Fc27 copies	+	+	+	+	+	+	+	+	+	+	-	-	-	-

*) 1 template corresponds to 1 genome of a FC27 or 3D7 parasite from an *in vitro* culture.

5.4.2 Longitudinal genotyping in field samples from Ghana.

From 100 individuals enrolled at baseline, 550 blood samples were collected during the one year follow-up in two-monthly intervals. 78.2% were found to be positive for *P. falciparum* by PCR. For 99 individuals GeneScan analysis was successfully performed and 96 individuals were analyzed by PCR-RFLP, accounting for 1405 and 1084 observed clonal infections, respectively. The discrepancy is due to differences in the resolution of these methods. Problems occurred in particular with the PCR-RFLP method when interpreting superimposed RFLP patterns.

A total of 164 different *msp2* alleles were distinguished by GeneScan analysis, 116 and 48 belonged to the 3D7 and FC27 allelic family, respectively. Frequencies of 3D7-type alleles were all <3% (**Figure 5.1a**). Some FC27-alleles occurred at very high allelic frequencies, the most frequent one reaching 14% (**Figure 5.1b**).

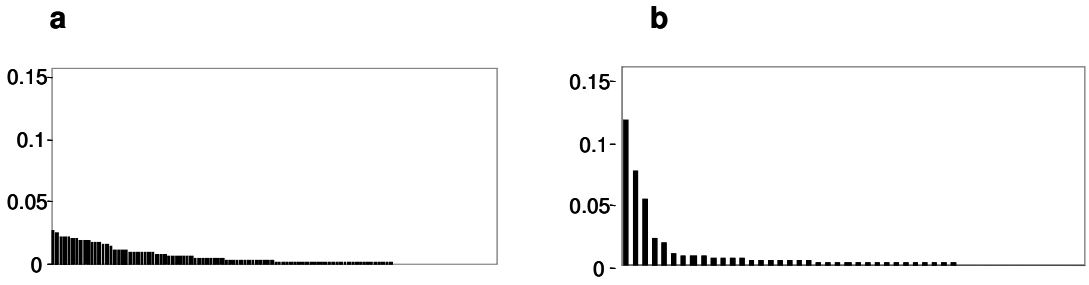


Figure 5.1: Frequencies of 164 different *msp2* genotypes detected by GeneScan in 6 two-monthly surveys of 99 individuals from Ghana (total number of parasite clones detected n = 1405). a. Genotypes belonging to the 3D7 allelic family (n = 116). b. FC27-type genotypes (n = 48).

5.4.3 Multiplicity of infection.

Overall MOI assessed by Genescan was higher than by PCR-RFLP with a mean multiplicity of 6.6 (95% confidence interval: 5.6 - 7.6) in age group 5-9 years, while RFLP analysis only detected a mean of 5.0 (95% CI: 4.5 - 5.5) infections. The age trends were similar with the two methods, with average MOI increasing until the age of 5-9 and decreasing during adolescence and adulthood (**Figure 5.2** a). Mean multiplicity in each age group was analyzed separately for 3D7- and FC27-type alleles (**Figure 5.2 b. and c**). 3D7-type infections were more frequent in all age groups than FC27-type infections.

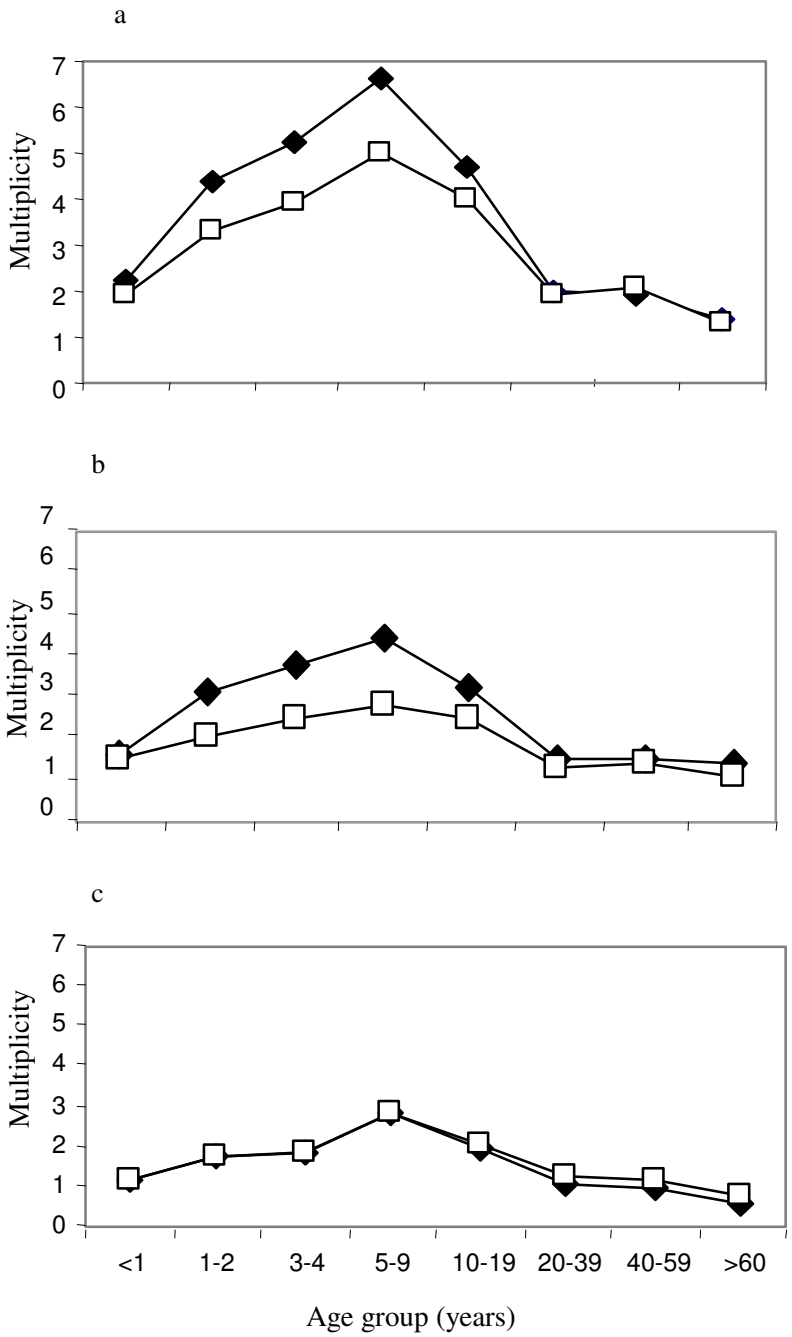


Figure 5.2: Age dependency of mean multiplicity. a. Overall multiplicity assessed by GeneScan and PCR-RFLP. b. 3D7-type *msp2* alleles. c. FC27-type *msp2* alleles.

5.4.4 Infection dynamics.

To describe the longitudinal genotyping data, we determined the numbers of transitions between genotypes of consecutive samples. A transition is either a loss of a genotype or a gain (depicted as patterns "+ -"and "- +", respectively). When comparing both genotyping methods a similar age distribution of transitions (**Figure 5.3 a and b**) was found. For both methods the gain of new infections was highest in children aged 5-9. The loss of infection was highest in people older than 60 years of age and the persistence of infecting clones was less frequent in older age groups.

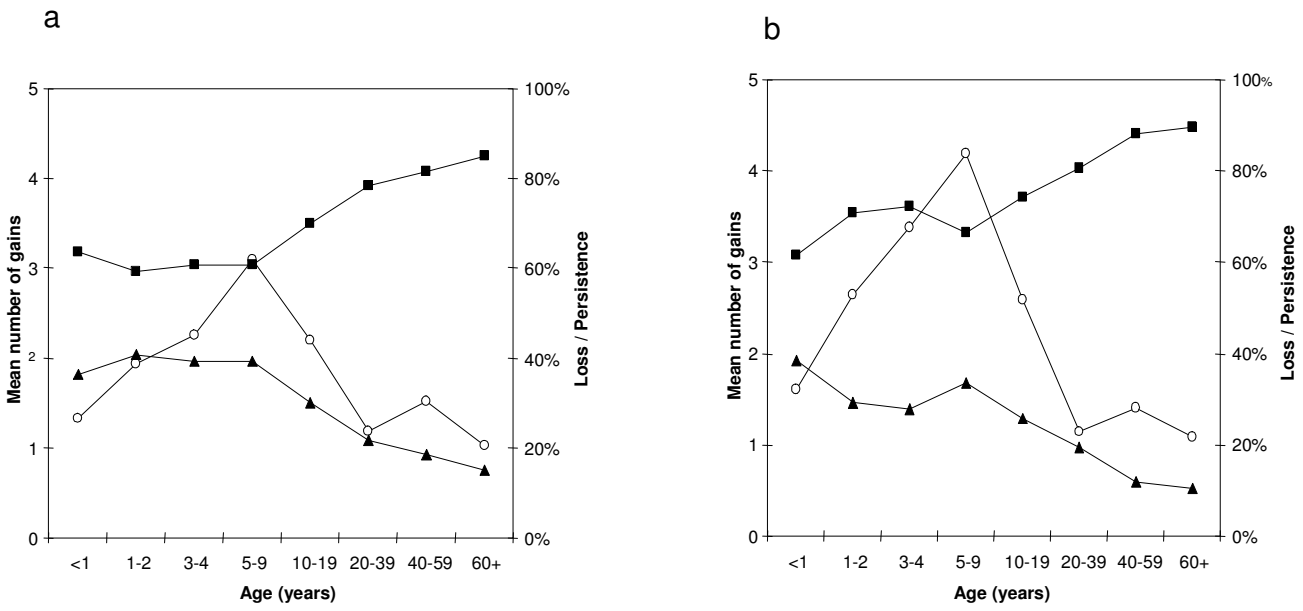


Figure 5.3: Number of newly acquired, lost or persisting infections per person-interval by age group, determined with RFLP(a) and GeneScan (b).

Gain (- +), Loss (+ -), Persistence (+ +)

The pattern of gains and losses of genotypes over the course of the 12 months of the study exhibited seasonal variation (**Figure 5.4**). The number of gains was highest in first to second survey reflecting the higher transmission rate during the wet season. Gains became gradually less frequent in the following surveys. Loss and persistence rates remained unchanged throughout the year. Comparison of both genotyping methods showed that new alleles

(gain – +) were more frequently detected by GeneScan than by RLFP. The number of losses was slightly higher in samples analyzed with RFLP.

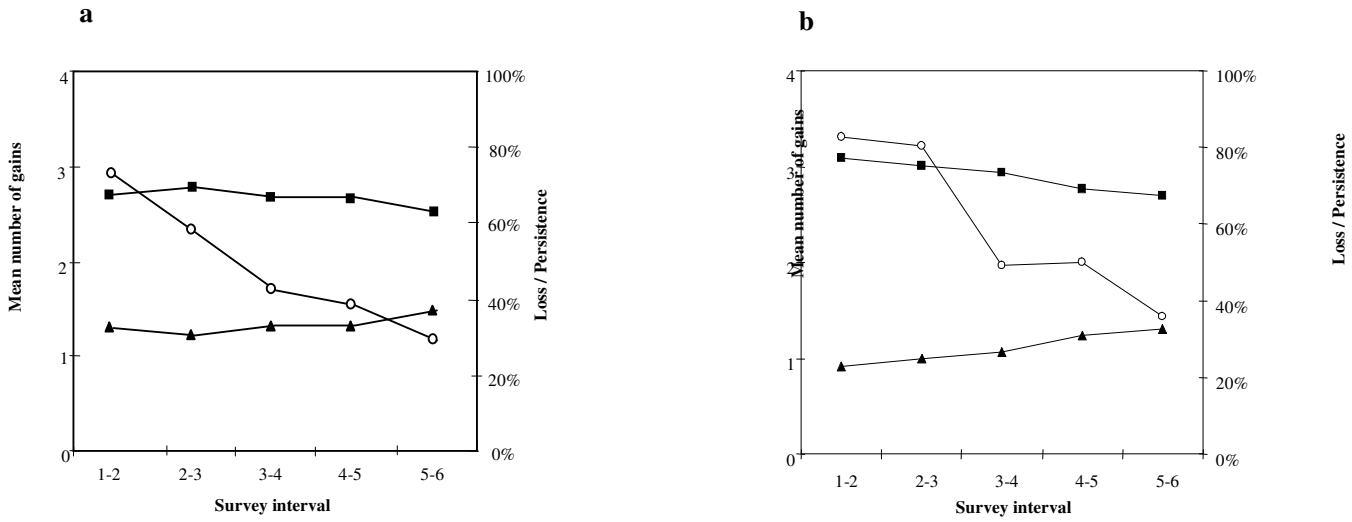


Figure 5.4: Frequencies of transition types by survey interval determined by RFLP (a) or GeneScan (b).

Gain (- +), Loss (+ -), Persistence (+ +)

5.4.5 Statistical analysis and modeling.

The above description of simple transitions does not take into account the imperfect detection of sequestered parasite clones. Statistical models can be applied to estimate the true presence and the actual persistence of genotypes (Smith and Vounatsou, 2003; Sama *et al.*, 2005). While paired samples do not allow to identify the detectability (the detected proportion of all genotypes actually present), this can be estimated from sequences of three or more consecutive samples from the same individual. We analyzed our data according to an approach proposed by Sama *et al.* (2005) making use of an immigration death model. Immigration refers to the acquisition of a new parasitic genotype with rate λ, and death refers to the clearance of a parasitic genotype with rate μ. The model assumes that corresponding to each observed process which is the detection or failure to detect a parasitic genotype is an underlying true process which is hidden as a result of imperfect detection by the parasitological diagnostic tool (Genescan or RFLP). For GeneScan, this approach estimates a detectability of 0.35 (95% CI: 0.31 - 0.39) that is an average of 35% of the parasite present in the host are detected in a finger-prick blood sample. For RFLP the detectability was 0.47 (0.42 - 0.51). This estimation of the detectability was also done based on a simpler approach (Smith *et al.*, 1999b) which led to comparable results (not shown).

Based on the genotype acquisition and clearance rates λ and μ, respectively, we determined the average duration of an infection. We estimated a value of 19.6 (17.3 - 22.0) per year for λ and 1.7 (1.4 - 2.1) per year for μ, which corresponds to an average duration of infection of 210 (176 - 256) days when samples were analyzed by GeneScan. For RFLP analysis, λ and μ were found to be 16.3 (14.8 - 18.0) and 2.4 (2.1 - 2.7), respectively, which corresponds to an average duration of infection of 152 (133 - 177) days. Our comparison of both genotyping techniques and both statistical approaches showed that the turnover rate for 3D7- type alleles

was almost twice as high as for the FC27-type alleles and, that infections with FC27-type alleles persisted longer.

5.5 Discussion

The *P. falciparum msp2* gene is highly polymorphic and therefore qualifies as ideal marker gene for analysis of parasite dynamics. The PCR-RFLP genotyping technique that has been applied in many descriptive and analytical studies (Genton *et al.*, 2002; Felger *et al.*, 1999b; Fraser-Hurt *et al.*, 1999; Irion *et al.*, 1998). GeneScan-based *msp2* genotyping overcomes some limitations of PCR-RFLP. The major advantage lies in the accuracy of identifying particular genotypes in complex mixtures and in different samples. Misclassification of the allelic family is prevented by using fluorescently-labeled primers, allele frequencies can be determined precisely and throughput is increased considerably. Alternative genotyping approaches were also based on discrimination of amplicon size but genotypes represented discrete bins, spanning 20 nucleotides and comprising several alleles of similar size (Snounou *et al.*, 1999; Ranford-Cartwright *et al.*, 1993). For some loci this identifies only few alleles making it necessary to use statistical models to estimate allele frequencies (Hill and Babiker, 1995). Our new methodology produces precise genotype frequencies which are essential to estimate the rate of reinfection with the same genotype.

In our study we have investigated several parameters, MOI by age, different transition states by allelic family, and seasonal effects on these transitions. In both data sets, mean MOI was low among infants, increased steadily until 5-9 years and started to decrease again during adulthood. This finding is consistent with the previously published cross-sectional results from the first survey (Owusu-Agyei *et al.*, 2002). The distribution of the three transition types (gain, loss and persistence of clones) was similar in by both methods, with acquisition of infections being most frequent within the first and second transition interval, reflecting the increase in transmission rate during the second half of the wet season. The number of

transitions, however, was different between the two laboratory methods. By GeneScan more genotypes were gained whereas by PCR-RFLP the clearance rate was higher. Both these findings might be due to higher sensitivity of GeneScan. In particular, 3D7-type alleles are difficult to determine in polyacrylamide gels used for PCR-RFLP, but easy to identify by using family-specific fluorescent primers in GeneScan analysis, resulting in a higher number of acquisitions. The higher percentage of losses by PCR-RFLP can also be explained by higher sensitivity of GeneScan since the transition pattern "+ - - " by RFLP analysis was found to be "+ - + " by GeneScan.

Another parameter calculated from transition rates was the duration of infections. The turnover rate of 3D7 infections was higher in both data sets. This suggests that FC27 alleles are more stable over time and more resistant to elimination under selective pressure by the host's immune system than 3D7 alleles.

5.5.1 Sensitivity and detectability.

The sensitivity of PCR-RFLP versus GeneScan technique and their performance in a molecular epidemiological field study can be compared by the total number of parasite clones detected by each technique. Compared to PCR-RFLP we found an increase by 22.8% for GeneScan-based genotyping. This is also reflected by the parameter "mean MOI" which by GeneScan showed an increase of 1.6 infections at the peak of the age distribution. We found that the increased sensitivity in field samples was due to higher precision in discriminating 3D7-type genotypes. This can be explained by the scrambled repeat structure of 3D7-type alleles giving rise to only small differences in fragment sizes which can hardly be discriminated by PCR-RFLP in gel electrophoresis. We conclude that the increase in detected clones by GeneScan-typing was accounted for by problems of PCR-RFLP in resolving the complex 3D7-type patterns in multiple infections.

Even optimal sensitivity of *P. falciparum* detection does not reflect the whole parasite population in an infected individual. Sequestration of late stage parasites causes a 48 hours periodicity in detectability. Such fluctuations were observed when monitoring the daily dynamics of *P. falciparum* clones (Farnert *et al.*, 1997). Thus, imperfect detection of some of the *P. falciparum* clones concurrently present in a host is a consequence of the parasite's life cycle, and the estimation of persistence is complicated even using molecular methods. Therefore mathematical models need to be applied to estimate the molecular parameters based on such imperfect data.

Both the approaches we used (Smith *et al.*, 1999b; Sama *et al.*, 2005) take into account the frequencies of transitions. A substantial frequency of the pattern {+-+} indicates imperfect detection. The detectability is our estimate of that proportion of the duration of an infection during which the densities are high enough to be detected, averaged over all the infections known to be present. Paradoxically, we found a lower detectability by GeneScan (35%) than by PCR-RFLP (47%), despite the fact that GeneScan had detected more genotypes. We arrived at the same conclusion when the method of Smith *et al.* (1999b) was used. This can be explained by different detection limits. **Figure 5.5** shows the density profile of three hypothetical infections. Curve A represents an infection that persists at detectable density and is detected in almost the same proportion by both techniques. Curve B represents an infection that occasionally reaches detectable density, but persists sub-patently for a long time. A proportion of such infections is detected by GeneScan, but not by RFLP. Curve C represents an infection that persists for a short time and is also more likely to be detected by GeneScan. It is likely that patterns B and C contribute to a reduced overall estimate of GeneScan detectability compared to RFLP which never sees infections such as B or C.

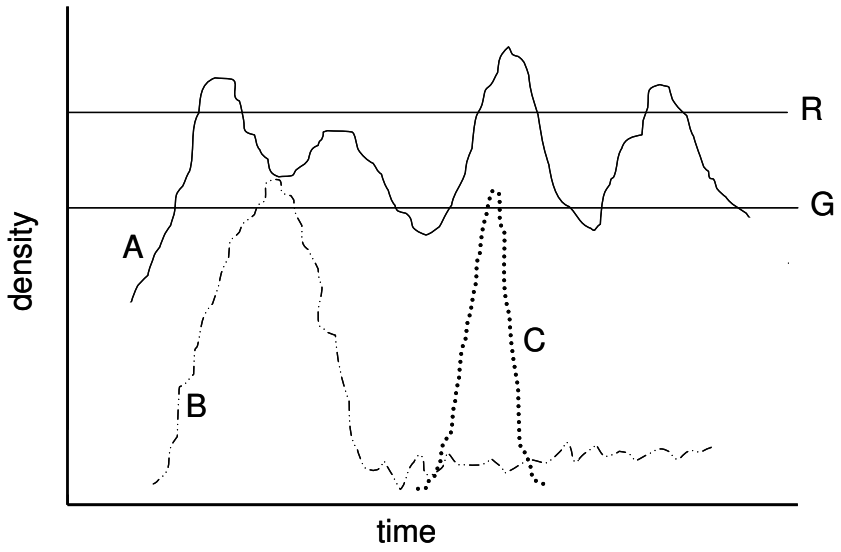


Figure 5.5: Illustration of detection limits of two methods, PCR-RFLP (R) and GeneScan (G); and the implication for detection of infections with three different density profiles, A, B, C.

5.5.2 Relevance of molecular parameters.

The molecular parameters we have used to describe multiple *P. falciparum* infections longitudinally can give new insights in malaria epidemiology. The force of infection can be studied with respect to age, seasonal variation or as an effect of an intervention. Also, the duration of infections in endemic areas can only be estimated by molecular means because of frequent superinfection. Effects of interventions on duration of infections can provide a valuable outcome measurement.

The potential of GeneScan-based *msp2* genotyping has recently been shown for the discrimination of recrudescent from new infections in drug efficacy studies (Jafari *et al.*, 2004). We anticipate further applications of molecular parameters in a range of intervention studies against malaria (Felger *et al.*, 2003). In malaria vaccine trials, molecular monitoring has been applied to detect selective effects if the vaccine was polymorphic (Bojang *et al.*, 2001; Fraser-Hurt *et al.*, 1999). Molecular infection dynamics parameters could further

contribute to describe the parasitological outcomes of vaccine trials and perhaps elucidate biological effects of candidate vaccines.

5.6 Acknowledgements.

Sincere thanks to the community members of the KND, especially the participants and the parents of the children consenting on their behalf. We thank the staff of the Navrongo Health Research Center for their field assistance, especially Lucas Amenga-Etego and Victor Asoala, and Béatrice Glinz-Szára and André Tiaden for genotyping.

Financial support. Swiss National Science Foundation (grant 3300C0-105994).

CHAPTER 6

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

Wilson Sama[1], Klaus Dietz[2], & Tom Smith[1].

[1]Swiss Tropical Institute, Basel, Switzerland.

[2]Institut für Medizinische Biometrie, Tübingen, Germany.

This paper has been accepted for publication in

Transactions of the Royal Society of Tropical Medicine and Hygiene

6.1 Abstract.

Survival time data of *P. falciparum* infections from deliberate infection of human subjects with *P. falciparum* between 1940 and 1963 as a treatment for neurosyphilis in the United States (Georgia) has 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 distribution. 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 program for eliminating or reducing malaria.

6.2 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 interval of times between successive arrivals (and also between arrival and departure) follows 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 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 an extensive literature on mathematical models of malaria with some fundamental departures from the view of malaria epidemiology as conceived by Ross (see for instance the 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 re-infection and the common diagnostic tool (light microscopy) in the field are not able to differentiate between infections derived at different points in time. Polymerase Chain Reaction (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, 1999a). However given that even the PCR is not 100% sensitive, the continual appearance and disappearance of specific types in consecutive blood samples in longitudinal studies makes it difficult to characterize such types as new infections 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. Due 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 *P. falciparum* infections deliberately induced as a treatment for neurosyphilis in the mid-20th century in the USA, at a time where there were no antibiotics for treatment of neurosyphillis (Collins and Jeffery, 1999; Jeffery and Eyles, 1955). Though this data was collected from

non-immune individuals in a non-endemic area, it represents an excellent source of data for testing the assumption of an exponential distribution for the survival time.

We study the fits of four 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 all what follows we would use the word hazard to mean the clearance rate (or recovery rate) of infections.

6.3 Methods

6.3.1 Data

Malariatherapy data were collected in the United States (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 *Plasmodium falciparum* were inoculated either with sporozoites (generally through mosquito bite) or with infected blood. Microscopic blood examinations were performed almost daily.

Out of a total of 334 patients in our database, 157 were from Georgia hospital and 177 from S. Carolina. The average duration of infections in Georgia patients was 135.2 days (standard error of 8.8), and 75.4 (s.e. 4.2) in S. Carolina patients, irrespective of, whether they received treatment or not, and the time at which treatment was given. 99/157 Georgia patients received treatment, with a total of 540 days when treatment was given while 116/177 S. Carolina 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 S. 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. We did not observe any substantial difference in the mean (arithmetic) duration of infections, by varying the qualifying period. Among the 54 patients, 23 of them received some subcurative treatment before their last positive slide.

6.3.2 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 basic properties of the above distributions, for instance hazard rates, quantiles, see (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, this 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, the random fluctuations in any particular dataset will cause the point 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).

In order to obtain the theoretical quantiles, the parameters of the distributions considered have to be estimated. We estimate these parameters by maximum likelihood methods implemented in Fortran 95 (*Compaq Visual Fortran Version 6.6. Compaq Computer Corporation. Houston, Texas 2001*) using the quasi-Newton algorithm (Gill and Murray, 1976) from the NAG FORTRAN library (*NAG Fortran Manual, Mark 19, NAG Ltd., 1999*). 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.

6.4 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, while 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 the right (Figure 6.1). The Wilcoxon rank-sum test indicates that the two samples (those who received the SanteeCooper strain and those who received the ElLimon strain) comes from the same distribution (z = -0.193, P = 0.85).

The five parametric distributions considered are shown in Table 6 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.

The quantile-quantile plots suggest that the Gompertz distribution provides the best fit to the data. This is closely followed by the Weibull and the Gamma distributions. The worst fits are obtained from the exponential and log-normal distributions (Figure 6.1, Figure 6.2). This ranking is also confirmed by the AIC values; while the K-S test suggest that the Weibull fit is slightly better than the Gompertz (Table 6). The fits of the distributions depicted on Figure 6.1 seems to favour more the results from the K-S test. The tails of the probability density functions of the Gompertz and Weibull are shorter than the remaining three distributions considered (Figure 6.1). Plots of the hazard rates indicate that for the Gompertz, Weibull and gamma distributions, the hazard increases with the age of the

infection and the rate of increase is highest for the Gompertz, and lowest for the gamma. The hazard rates for the log-normal peaks and decreases again with increase age of infection (Figure 6.3).

Table 6. Estimates of parameters and expected lifetimes (with 95% confidence intervals in brackets), the Akaike information criterion (AIC), and the Kolmogorov-Smirnov statistics (with p-values in brackets), for some common parametric distributions for survival times, using malariatherapy data. $F(x)$ is the cumulative distribution function.

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

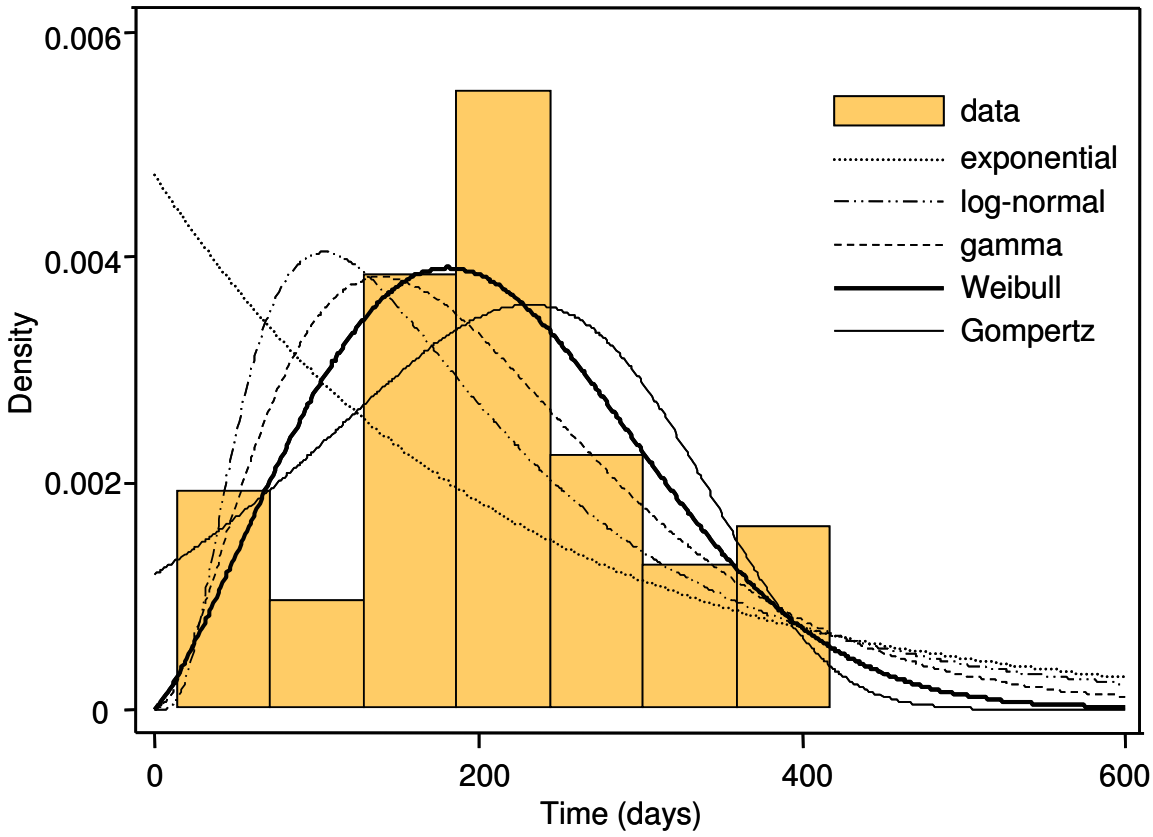


Figure 6.1. A histogram of the observed data and the probability density functions of exponential, log-normal, gamma, Weibull and Gompertz distributions. The parameters used for the distributions are maximum likelihood estimates from the fits of these distributions to the malariatherapy data.

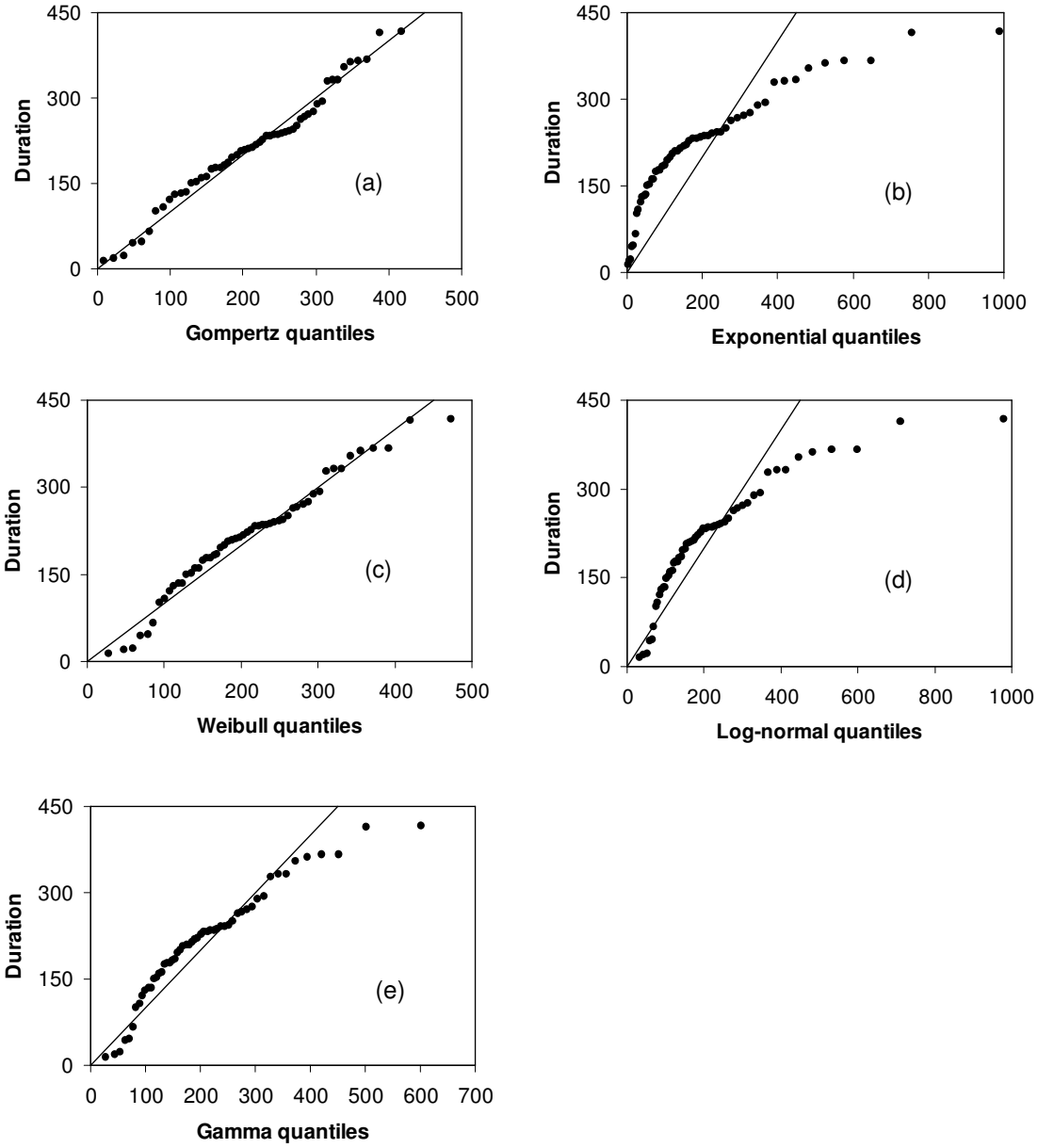


Figure 6.2. (a) - (e): Quantile-Quantile plots of the malariatherapy data. The straight line represents the reference line.

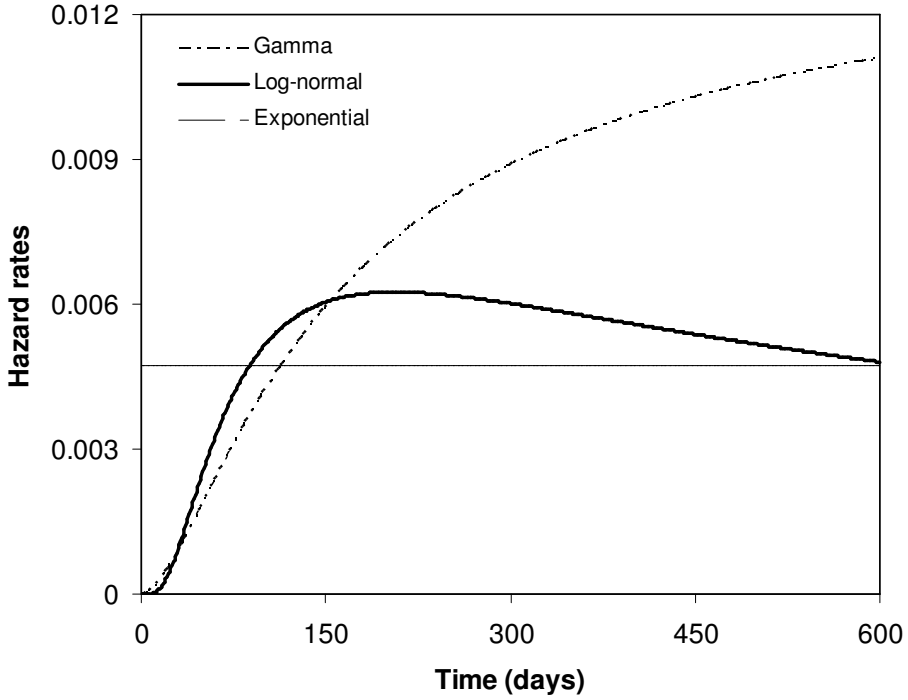


Figure 6.3(a) Plots of the hazard rates for the gamma, log-normal, and exponential distributions using the malariatherapy data.

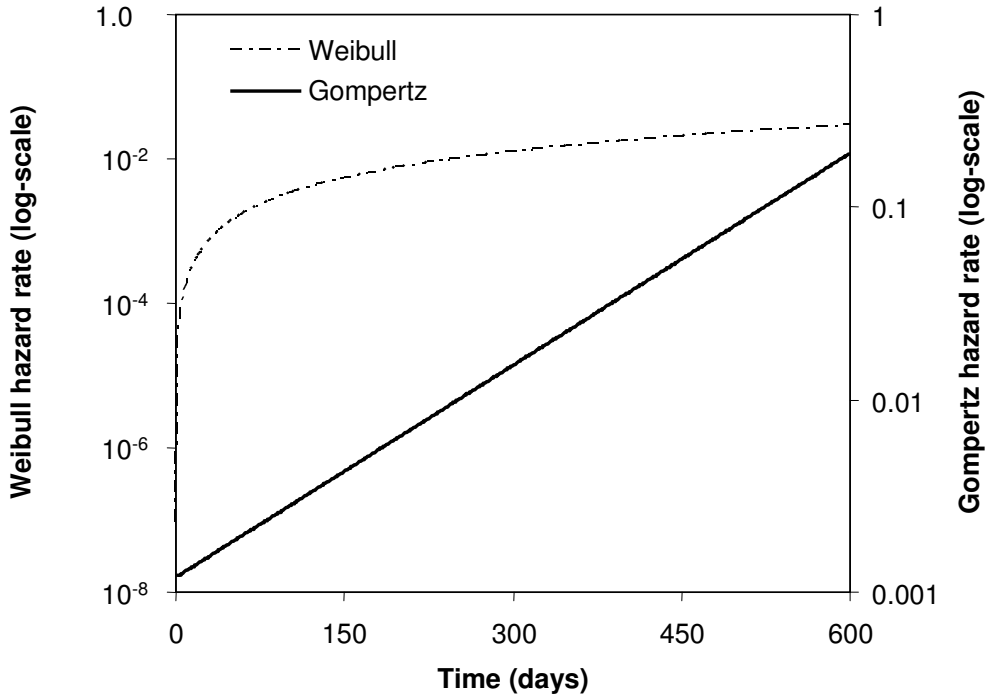


Figure 6.3(b) Plot of the hazard rate for Weibull and Gompertz distributions using the malariatherapy data.

6.5 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 United States 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 infections are not available. This however raises the question of the applicability of the results to endemic areas. It has been reported that infections in naïve host mimic the levels of parasitaemia reached by *P. falciparum* infections induced to treat neurosyphilis in the United States (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 make the scale parameter a suitable function of age.

Our results suggest that the best fit is obtained from the Gompertz or Weibull distribution. A moderate fit is obtained using the gamma distribution. We do not obtain good fits using the log-normal and exponential 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 important to know the correct distribution because its shape profoundly influences the length of monitoring needed in an intervention program for eliminating or reducing malaria. The density functions of the Gompertz and Weibull (Figure 6.1) 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 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 6.3). For instance if transmission is interrupted at a time of the year where people harbour more new infections (for instance the rainy season) than old ones, then the Gompertz and Weibull suggest that a longer period of monitoring is required after the intervention, while for the exponential, the hazard does not depend on the age of the infection. This 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 that we have 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 malariatherapy patients.

The theory of malaria eradication proposed in the 1950's (Macdonald, 1957) is partly based on length of times for which eradication campaigns should be carried out to ensure that there was no residual transmission. Most of this theory has assumed that the length of time for which duration of untreated infections will die away naturally has an exponential distribution. This theory has to 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 (Macdonald, 1950b) estimate of 200 days (based on the exponential distribution) which was used for the planning of 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 *Plasmodium falciparum* infections.

6.6 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.

CHAPTER 7

Discussion

As a result of acquired immunity, most inhabitants in malaria endemic areas harbour parasites within them without showing any signs or clinical symptoms of the disease and hence they are seldom treated unless when these individuals show clinical symptoms. Such individuals are often asymptomatic and they often serve as a reservoir of infection for the subsequent transmission season. Such asymptomatic individuals often complicate disease control strategies as it is often not known how long such infections persist within them.

The two processes of acquisition and clearance of malaria infections determine the prevalence of infection of the human population. The impact of a preventative intervention is largely a function of its effect on the force of infection and on the recovery rate from infection or equivalently, the average duration of infection. Hence in evaluating the effectiveness of certain intervention programs, it is important to know a priori the force of infection and the recovery rate. However until recently only few studies have attempted to estimate these quantities from field data. The main objective of this thesis was to find out how long such untreated malaria infections persist within the host in endemic areas and how often new infections arise. Since *Plasmodium falciparum* is the most prevalent and virulent form of the disease in endemic areas, our analysis of the

within host dynamics of malaria is in this work is solely based on datasets on *Plasmodium falciparum* malaria.

The datasets used in our analysis comes from distinct projects. The motivations for using these datasets have been previously explained in the various chapters and will be briefly mentioned here. We have used three historical malaria prevalence datasets from Garki, a district in Northern Nigeria; Pare-Taveta, a contiguous area bordering Tanzania and Kenya.; and West Papua, and analysed using light microscopy. Another source of data used in this work comes from a more recent study in Navrongo, a district in Northern Ghana where blood samples were analyzed using the PCR-RFLP technique (Felger *et al.*, 2002). The last source of data (commonly referred to as malariatherapy data) analysed in this work was collected from neurosyphilis patients. This is data on the duration of artificial infections with malaria in neurosyphilis patients.

A detailed discussion of the techniques used in analyzing the data and the findings in this work has been given separately in the corresponding Chapters. Here we provide a summary of the main contributions and an outlook and recommendations for future research.

The methods developed in this work has permitted us to (i) estimate the clearance rate of *Plasmodium falciparum* infections (and hence the average duration of infections, (ii) estimate the infection rate, that is how often new infections are acquired, (iii) assess the dependence of past exposure (or age or acquired immunity) on the clearance and infection rates, and (iv) to assess the importance of allowing for both seasonality in infection rate and imperfect detection of the PCR in estimating the duration of infection

in endemic settings, (v) study the suitable statistical distribution for the survival time of *Plasmodium falciparum* infection within the human host.

In Chapter 2, we have reviewed the use of exponential decay models in estimating the duration of infections and extended the method by allowing for random variations in the observed data. The original idea was that of Macdonald (1950b) who analysed *Plasmodium falciparum* data from a study in Puerto-Rico and came to the conclusion that infections with *Plasmodium falciparum* last about 200 days. We however obtained an average duration of above 600 days using the data from Pare-Taveta, West Papua and Garki. It should be noted that the surveys in Pare-Taveta and West Papua were cross-sectional surveys while that of Garki was a longitudinal survey. We carried out two separate analyses of the Garki Study, one in which we treated the data as coming from cross-sectional surveys and the other from a longitudinal survey. The results from the analysis as a longitudinal survey gave an average duration of 186 days, very close to what Macdonald (1950b) had earlier obtained. In these analyses there was a slight tendency for the infection duration to reduce with age in all the three areas. Our main conclusion from this Chapter was that the average duration of infection was much longer than was thought in the past.

The conclusions derived from Chapter 2 was based on the assumption that there was no re-infection going on. This was backed by the fact that there were intensive vector control measures prior to the surveys. However in the study analysed in Chapter 3, this assumption no longer holds because there was no vector control. This therefore gave us the possibility to account for new infections and hence get an estimate for the infection rate. In Chapter 3, we applied the model originally proposed by Macdonald (1950a) and

later corrected by Fine (1975) for the dynamics of superinfection. Macdonald developed this model at a time where there was no suitable data on it could be tested. With the advent of the PCR, it was now possible to fit such models to data derived from PCR studies. Using PCR genotyping that from the study in Navrongo (Owusu-Agyei *et al.*, 2002), we estimated an average duration of infection of 152 days for any single *Plasmodium falciparum* genotype for any individual living in this holo-endemic site. The estimated expected multiplicity of infections in the study population was 7, and an estimate for the expected total duration of 394 days. This estimate is less than the estimates of 600 days obtained in Chapter 2, but twice that of 200 days obtained by Macdonald (1950b).

It was also estimated that each individual will acquire about 16 new infections per year, and we estimated a detectability of 47%, that is about 47% of the parasite types present in the host is found in each blood sample.

We did not at this stage allow for any age-dependence in the parameters, hence all the estimates obtained were population based averages. We did not also allow for differences in infection rate or duration for the different genotypes, thereby assuming that they are exchangeable.

In Chapter 4, we extended the methods of Chapter 3 by allowing for age dependence in infection rate, clearance rate and detectability of infection. This was done by considering a number of suitable mathematical functions to describe the dependencies. We also accounted for seasonal variation in infection rates by assigning a different parameter for the infection rate at each survey-interval. This led to the evaluation of several different

models from which a model selection criteria was used to select the best model. The best fitting model indicated that there is seasonal variation in infection rate with high infection rate during the wet season and low infection rate during the dry seasons as anticipated. This model also indicated a decrease in detectability with age. The best model did not indicate any age dependence in clearance rate or infection rate. However this conclusion should be treated with caution because we found that there was collinearity between the parameters for the clearance rate and those for the detectability. For instance all models where the detectability was assumed to be constant estimated a significant increase in the clearance rate with age, meanwhile when the detectability was no longer assumed to be constant, this increase was no longer statistically significant. This leads us to speculate that the tendency is for the clearance rate to increase with age. However we must state that our method could not resolve this correlation problem between the detectability and the clearance rate, hence we cannot draw any firm conclusions regarding the dependence of past exposure or age on the clearance rate.

The PCR-RFLP technique has been shown to be more sensitive than light microscopy and hence provides a more realistic means of analysing infection dynamics. In Chapter 5, we compare a more recent method, the GeneScan, to PCR-RFLP and it is found that the GeneScan has a better performance compared to the PCR. For the same 69 individuals considered in Chapter 3 for the Navrongo dataset, a total of 70 distinct alleles were detected using the PCR, while a total of 119 genotypes were detected using the GeneScan technique. Applying the method developed in Chapter 3, an estimate of 210 days was obtained for the duration of infection of a single genotype, and 20 per year for the infection rate, using data from the GeneScan analysis. Considering that the expected

multiplicity of infection using the GeneScan analysis was 11, the total expected duration of infection is 634 days, very close to the results obtained from the analysis of the cross-sectional surveys in Chapter 2.

One critical assumption that we have used so far in the previous Chapters is that the duration of infection follows an exponential distribution. In Chapter 6, we test this assumption by fitting five commonly used parametric statistical distributions for survival time data to data (malariatherapy data) collected from neurosyphilis patients in Georgia. The best fit is obtained from the Weibull and Gompertz distribution suggesting that the exponential model is not realistic to approximate the survival time of *Plasmodium falciparum* infections. However the malariatherapy data was collected from naïve individuals and probably does not reflect the reality in endemic areas. Nonetheless it provides a useful first approximation. An alternative strategy would be to test a number of different distributions on data from endemic areas and then compare the fits of the models obtained from the different distributions.

In our report here we do not provide results using the conclusions obtained from Chapter 6, that is re-analysing our data with the Weibull and Gompertz as alternative distributions. However we present in the Appendix (Appendix A), a method we have developed that attempts to incorporate these assumptions.

APPENDIX A

Accounting for age of infection in estimating the clearance rate of *Plasmodium falciparum* infections

Wilson Sama[1], Seth Owusu-Agyei[2], Ingrid Felger[1], Klaus Dietz[3], & Tom Smith[1].

[1]Swiss Tropical Institute, Basel, Switzerland.

[2]kintampo Health Research Centre, Kintampo, Ghana

[3]Institut für Medizinische Biometrie, Tübingen, Germany.

The method described in this appendix has not been tested on data, hence it is therefore a subject for further research.

Introduction.

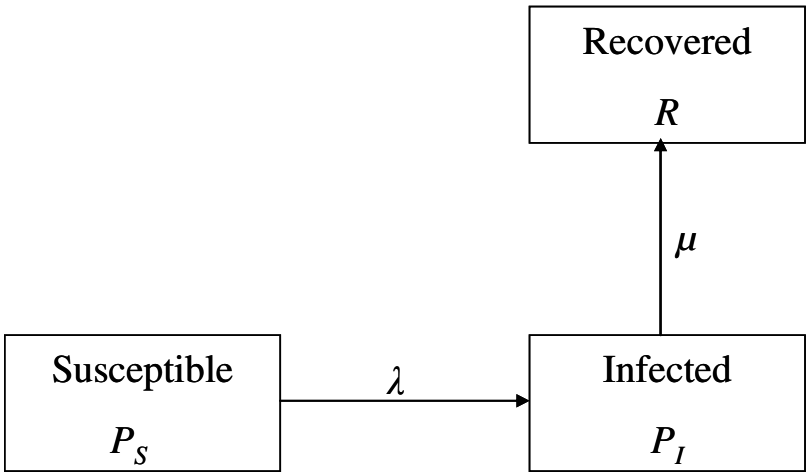
Consider a human population subjected to a wide diversity of infections with the possibility of being infected with more than one of such infections at any given time. An example of an infection with such diversity is *Plasmodium falciparum*, which we have recently studied. In our previous study we formulated in Chapter 3, a model for the infection dynamics by assuming a Poisson likelihood for the frequency of occurrence of sequences of observations and assuming an exponential distribution for the survival time of infections. We now consider an alternative formulation of the model allowing for the possibility that the survival time of each infection may depend on the age of the infection and assuming a multinomial likelihood for the observed process. We assume as in Chapter 3 that there are 63 possible observed patterns and 21 true patterns. For a start however, we will consider the pattern 000000, so that the 64^{th} observed pattern will refer to this pattern and the 22^{nd} true pattern will refer to the same pattern. This pattern is however not observed so we would discard it when deriving the probabilities for the 63 observed sequences. Our terminology to observed and true sequences (or patterns) is the same as in our previous work (see Chapter 3).

Method

Let $(n_{k,1}, n_{k,2}, ..., n_{k,63})$ denote the realization of a multinomial random variable with parameters $N_k, p_{k,1}, p_{k,2}, p_{k,3}, ..., p_{k,63}$, where $N_k = \sum_{i=1}^{63} n_{k,i}$ and $p_{k,i}$ is the probability that individual *k* is observed with pattern *i*. We demonstrate below how to derive the $p_{k,i}$. Firstly we derive the probabilities for the 22 true sequences, $q_{k,i}$ (we are now taking into

account the sequence 000000; this will be discarded later, so that we are left with only 21 true sequences as in Chapter 3).

Because of the wide diversity of this infection, one of our basic assumptions was that, an individual once infected cannot be re-infected with the same strain. This leads to the SIR (susceptible-infected-recovered) model depicted in the figure below for a single infection (strain or genotype), where clearance is via a Poisson process.



Let $P_S(t)$ and $P_I(t)$ be respectively the proportion susceptible and infected at time *t* with a single infection and assume a constant force of infection λ and a constant clearance rate μ. The assumption of a constant μ implies that the survival time of the infection is independent of its age. We described the changes over *t* of $P_S(t)$ and $P_I(t)$ by the differential equations:

$$\frac{dP_S(t)}{dt} = -\lambda P_S(t) \qquad (1)$$

$$\frac{dP_I(t)}{dt} = \lambda P_S(t) - \mu P_I(t) \qquad (2)$$

If the whole population is susceptible to a given infection at time $t = 0$, then

solutions of equations 1 and 2 are given by

$$P_S(t) = e^{-\lambda t} \qquad (3)$$

$$P_I(t) = \frac{\lambda}{\lambda - \mu}(e^{-\mu t} - e^{-\lambda t}) \qquad (4)$$

Let us now consider a more general situation where the clearance rate is not constant. Let *S* be a generic survival time random variable with probability distribution function *f*(*s*) and hazard function *h*(*s*), where *s* is the age of the infection. Then we simply have to replace equation 2 above by a partial differential equation. Before writing down the equation, first consider the following two sequences of true status:

0 0 0 1 1 0 and 1 1 1 1 0 0.

For the first sequence, we know that infection occurred at some time during the third interval. The idea here is to derive an expression for the probability of observing such a sequence and average the expression over all possible infection times. For the second sequence, all we know is that the individual was infected at baseline but we do not know the age of the infection at baseline. It can however not be greater than the age of the individual at baseline. Let a_k be the age of a subject *k* at baseline. If such a subject has an infection of age *s*, then the subject was infected at time $t = a_k - s$.

Let $T_{\max}$ be the maximum time of past exposure (this is to avoid the possibility of a subject of age, say 70 years at baseline having an infection which could be possibly 70 years old). Let t_k be the relevant time of past exposure for subject *k*, that is $t_k = \min(a_k, T_{\max})$ and let λ^- be the constant force of infection in the past, and let $\lambda(t)$ be the force of infection from baseline onwards.

Let $P(t,s)$ be the probability of a positive state at time *t* with an infection of age *s*. The partial differential equation governing $P(t,s)$ is given by:

$$\frac{\partial P}{\partial t}+\frac{\partial P}{\partial s}=-h(s)P(t,s) \qquad (5)$$

with initial and boundary conditions respectively given by

$$P(0,s)=g(s)=\lambda^{-}e^{-\lambda^{-}(t_k-s)-\int_0^s h(u)du} \qquad (6)$$

$$P(t,0)=v(t)=\lambda(t)\exp(-\int_0^t \lambda(u)du) \qquad (7)$$

where $h(s)$ is the intensity of the clearance process (hazard). The general solution for $P(t,s)$ is given by

$$P(t,s)=\begin{cases} g(s-t)\exp(\int_{s-t}^{s} h(u)du) & \text{if } t\le s \\ v(t-s)\exp(\int_0^s h(u)du) & \text{if } t\ge s \end{cases} \qquad (8)$$

We consider 3 different survival time distributions here, namely the Weibull, exponential and the Gompertz.

The pdf, $f(x)$, hazard, $h(x)$, and survival functions $S(x)$ of the Weibull are given by $f(x)=\alpha\gamma x^{\alpha-1}\exp(-\gamma x^{\alpha})$, $h(x)=\alpha\gamma x^{\alpha-1}$, $S(x)=\exp(-\gamma x^{\alpha})$. The mean survival time is given by: $E(X)=\dfrac{\Gamma(1+1/\alpha)}{\gamma^{1/\alpha}}$

The special case where $\alpha=1$ corresponds to the exponential distribution. Similarly for the Gompertz distribution:

$f(x)=\theta e^{\alpha x}\exp[\dfrac{\theta}{\alpha}(1-e^{\alpha x})]$, $h(x)=\theta e^{\alpha x}$, $S(x)=\exp[\dfrac{\theta}{\alpha}(1-e^{\alpha x})]$, $E(X)=\int_0^{\infty} S(x)dx$.

As an illustration of how to derive the probabilities, consider the sequence 000110. We recall our assumption (in Chapter 2) that for a true sequence, reinfection with a specific genotype is a rare event during the course of field work. Let τ denote the length of an inter-survey interval. If we account for seasonal fluctuation in the infection rate by letting $\lambda(t) = \lambda_j,\ j = 1,2,...,5$, where each λ_j is constant (e.g. λ_1 is the infection rate between the first and the second survey, then we use the solution of equations 3 and 8 to derive the probability of the sequence 000110 as:

$$e^{-\lambda^- t_k} e^{-(\lambda_1+\lambda_2)\tau} \lambda_3 e^{-\lambda_3(\tau-u)} [S(\tau+u) - S(2\tau+u)], \text{ where } 0 \le u \le \tau \qquad (9)$$

If we assume that *u* has a uniform distribution over [0, τ], then we can average the above expression over all possible infection times, and get the required solution as:

$$e^{-\lambda^- t_k} e^{-(\lambda_1+\lambda_2)\tau} \frac{1}{\tau} \int_0^{\tau} \lambda_3 e^{-\lambda_3(\tau-u)} [S(\tau+u) - S(2\tau+u)] du \qquad (10)$$

Equation (10) assumes that the individual was never infected before baseline (the first survey). If we relax this assumption, then the probability that the individual was infected and the infection is already terminated before the first survey is $\frac{1}{t_k} \int_0^{t_k} \lambda^- e^{-\lambda^-(t_k - x)} (1 - S(x)) dx$, so that a more realistic solution to the probability of the true sequence 000110 is

$$(e^{-\lambda^- t_k} + \frac{1}{t_k} \int_0^{t_k} \lambda^- e^{-\lambda^-(t_k - x)} (1 - S(x)) dx) e^{-(\lambda_1+\lambda_2)\tau} \frac{1}{\tau} \int_0^{\tau} \lambda_3 e^{-\lambda_3(\tau-u)} [S(\tau+u) - S(2\tau+u)] du \qquad (11)$$

Similarly the probability of observing the sequence 111100 is given by

$$\frac{1}{t_k} \int_0^{t_k} \lambda^- e^{-\lambda^-(t_k - u)} [S(3\tau+u) - S(4\tau+u)] du \qquad (12)$$

A third example we consider is the sequence 0 0 0 0 0 0, where no infection was observed. Such a sequence can arise either because the subject was never infected, or the infection is already terminated just before the first survey, or that the subject was never infected until the first survey but a transient infection occurred in the first survey-interval or that the subject was never infected until the second survey but a transient infection occurred in the second survey-interval, and so on. The probability of observing such a sequence is therefore given by:

$$q_{k,22} = \left(e^{-\lambda^- t_k} + \frac{1}{t_k}\int_0^{t_k} \lambda^- e^{-\lambda^-(t_k-u)}[1-S(u)]du \right)\left(e^{-(\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5)\tau} + \frac{1}{\tau}\int_0^{\tau} \lambda_1 e^{-\lambda_1(\tau-u)}[1-S(u)]du \right.$$

$$+e^{-\lambda^- t_k}e^{-\lambda_1\tau}\frac{1}{\tau}\int_0^{\tau} \lambda_2 e^{-\lambda_2(\tau-u)}[1-S(u)]du + e^{-\lambda^- t_k}e^{-(\lambda_1+\lambda_2)\tau}\frac{1}{\tau}\int_0^{\tau} \lambda_3 e^{-\lambda_3(\tau-u)}[1-S(u)]du +$$

$$\left. e^{-\lambda^- t_k}e^{-(\lambda_1+\lambda_2+\lambda_3)\tau}\frac{1}{\tau}\int_0^{\tau} \lambda_4 e^{-\lambda_4(\tau-u)}[1-S(u)]du + e^{-\lambda^- t_k}e^{-(\lambda_1+\lambda_2+\lambda_3+\lambda_4)\tau}\frac{1}{\tau}\int_0^{\tau} \lambda_5 e^{-\lambda_5(\tau-u)}[1-S(u)]du \right)$$

The expressions for all the sequences, $q_{k,j}$, derived in analogous manner as that of equation (11) are shown on the table below. Because of the long nature for the expression of the sequence 0 0 0 0 0 0, it is simply denoted as $q_{k,22}$ on the table.

To assure that the probabilities of the true sequences sums to unity, we normalize by defining for $k = 1,2,...,69,\ \ j, j' = 1,2,...,22$,

$$\tilde{q}_{kj} = \frac{q_{kj}}{\sum_{j'} q_{kj'}}$$

The reasoning used for deriving the probabilities of each of the observed patterns is analogous to that in Chapter 2. For instance if we assume that the detectability, π, is age dependent and consider a logistic function for this dependence, that is,

$\text{logit}(\pi_k) = \pi_0 + \pi_1(a_k - \bar{a})$, then the probability, S_{kij}, of observing pattern *i* in individual *k* given that the true pattern is *j* is given

$S_{kij} = I_{ij}\pi_k^{z_i}(1-\pi_k)^{z_j - z_i}$, where,

$$I_{ij} = \begin{cases} 1 & \text{if the observed pattern } i \text{ can arise from the true pattern } j \\ 0 & \text{otherwise} \end{cases}$$

z_i = the number of "ones" in the binary sequence corresponding to observed pattern *i*
z_j = the number of "ones" in the binary sequence corresponding to true pattern *j*

$k = 1,2,...,69,\ \ i = 1,2,...,64,\ \ j = 1,2,...,22$

To ensure the multinomial probabilities sum to 1, we define them by normalizing. So that, for $k = 1,2,...,69,\ \ i = 1,2,...,63,\ \ j = 1,2,...,21,\ j' = 1,2,...,22$

$$p_{ki} = \frac{\sum_j S_{kij}\tilde{q}_{kj}}{\left(\sum_i \sum_j S_{kij}\tilde{q}_{kj}\right) - \sum_{j'} S_{ki'j'}\tilde{q}_{kj'}}, \text{ where } i' = 64 \qquad (13)$$

Notice in the above equation that the expression for the 64th pattern (000000) is discarded as discussed earlier.

The multinomial likelihood for the complete data is:

$$\prod_{k=1}^{69} \binom{N_k}{n_{k,1}\ n_{k,2}\ \ldots\ n_{k,63}} p_{k,1}^{n_{k,1}} p_{k,2}^{n_{k,2}} \cdots p_{k,63}^{n_{k,63}}$$

All rate parameters are expressed as per year. Apart from π_0 and π_1 which take values on the real line, the remaining parameters are positive.

We can modify the expressions above by making the parameters of the Weibull distribution (and Gompertz distribution) depend on age of the individual. A reasonable choice for the Weibull would be to keep the shape parameter, α, fixed and allow γ to be a

function of age, that is $\gamma = \gamma(a_k)$ for individual k. So that $h(x) = \alpha\gamma(a_k)x^{\alpha-1}$ for the Weibull. Similarly, $h(x) = \theta(a_k)e^{\alpha x}$ for the Gompertz.

We illustrate the expressions for the probabilities of two sequences with the Weibull. We assume that γ depends on the age of the individual at the onset of the infection. The probability of observing the sequence 0 0 0 0 1 0 is given by

$$[e^{-\lambda^- t_k} + \frac{1}{t_k}\int_0^{t_k} \lambda^- e^{-\lambda^-(t_k-u)}(1-e^{-\gamma(a_k-u)u^\alpha})du]e^{-(\lambda_1+\lambda_2+\lambda_3)\tau}\frac{1}{\tau}\int_0^{\tau}\lambda_4 e^{-\lambda_4(\tau-s)}[e^{-\gamma(a_k+4\tau-s)s^\alpha} - e^{-\gamma(a_k+4\tau-s)(\tau+s)^\alpha}]ds$$

Similarly the probability of observing the sequence 1 1 1 1 1 0 is given by:

$$\frac{1}{t_k}\int_0^{t_k}\lambda^- e^{-\lambda^-(t_k-s)}[e^{-\gamma(a_k-s)(4\tau+s)^\alpha} - e^{-\gamma(a_k-s)(5\tau+s)^\alpha}]ds\,.$$

The expressions for the remaining true patterns can be derived in an analogous manner, and equation 13 applied again to derive those for the 63 observed patterns.

Expressions for the probabilities of the 22 true sequences.

Seq	probability	Seq	probability
000001	$q_{k,1} = F(\lambda_4)\int_0^{\tau} H(\lambda_5)S(u)du$	011000	$q_{k,12} = F(0)\int_0^{\tau} H(\lambda_1)[S(\tau+u) - S(2\tau+u)]du$
000010	$q_{k,2} = F(\lambda_3)\int_0^{\tau} H(\lambda_4)[S(u) - S(\tau+u)]du$	011100	$q_{k,13} = F(0)\int_0^{\tau} H(\lambda_1)[S(2\tau+u) - S(3\tau+u)]du$
000011	$q_{k,3} = F(\lambda_3)\int_0^{\tau} H(\lambda_4)S(\tau+u)du$	011110	$q_{k,14} = F(0)\int_0^{\tau} H(\lambda_1)[S(3\tau+u) - S(4\tau+u)]du$
000100	$q_{k,4} = F(\lambda_2)\int_0^{\tau} H(\lambda_3)[S(u) - S(\tau+u)]du$	011111	$q_{k,15} = F(0)\int_0^{\tau} H(\lambda_1)S(4\tau+u)du$
000110	$q_{k,5} = F(\lambda_2)\int_0^{\tau} H(\lambda_3)[S(\tau+u) - S(2\tau+u)]du$	100000	$q_{k,16} = \int_0^{t_k} G(\lambda^-)[S(u) - S(\tau+u)]du$
000111	$q_{k,6} = F(\lambda_2)\int_0^{\tau} H(\lambda_3)S(2\tau+u)du$	110000	$q_{k,17} = \int_0^{t_k} G(\lambda^-)[S(\tau+u) - S(2\tau+u)]du$
001000	$q_{k,7} = F(\lambda_1)\int_0^{\tau} H(\lambda_2)[S(u) - S(\tau+u)]du$	111000	$q_{k,18} = \int_0^{t_k} G(\lambda^-)[S(2\tau+u) - S(3\tau+u)]du$
001100	$q_{k,8} = F(\lambda_1)\int_0^{\tau} H(\lambda_2)[S(\tau+u) - S(2\tau+u)]du$	111100	$q_{k,19} = \int_0^{t_k} G(\lambda^-)[S(3\tau+u) - S(4\tau+u)]du$
001110	$q_{k,9} = F(\lambda_1)\int_0^{\tau} H(\lambda_2)[S(2\tau+u) - S(3\tau+u)]du$	111110	$q_{k,20} = \int_0^{t_k} G(\lambda^-)[S(4\tau+u) - S(5\tau+u)]du$
001111	$q_{k,10} = F(\lambda_1)\int_0^{\tau} H(\lambda_2)S(3\tau+u)du$	111111	$q_{k,21} = \int_0^{t_k} G(\lambda^-)S(5\tau+u)du$
010000	$q_{k,11} = F(0)\int_0^{\tau} H(\lambda_1)[S(u) - S(\tau+u)]du$	000000	$q_{k,22}$

$$F(\lambda_j) = [e^{-\lambda^- t_k} + \frac{1}{t_k}\int_0^{t_k} \lambda^- e^{-\lambda^-(t_k-u)}(1 - S(u)du]e^{-(\lambda_1+\lambda_2+\ldots\lambda_j)\tau}, \quad F(0) = e^{-\lambda^- t_k} + \frac{1}{t_k}\int_0^{t_k} \lambda^- e^{-\lambda^-(t_k-u)}(1 - S(u)du,$$

$$G(\lambda^-) = \lambda^- e^{-\lambda^-(t_k-u)},\ H(\lambda_j) = \lambda_j e^{-\lambda_j(\tau-u)}$$

Bibliography

Alonso P.L., Molyneux M.E., & Smith T. (1995) Design and methodology of field-based intervention trials of malaria vaccines. *Parasitology Today* **11,** 197-200.

Alonso P.L., Smith T., Schellenberg J.R., Masanja H., Mwankusye S., Urassa H., Bastos D.A., I, Chongela J., Kobero S., & Menendez C. (1994) Randomised trial of efficacy of SPf66 vaccine against *Plasmodium falciparum* malaria in children in southern Tanzania. *Lancet* **344,** 1175-1181.

Anderson R.M. (1982) *The Population Dynamics of Infectious Diseases. Theory and Applications.* Chapman & Hall, New York.

Anderson R.M. & May R.M. (1991) *Infectious Diseases of Humans: Dynamics and Control.* Oxford University Press, Oxford.

Aron JL (1988) Mathematical-modeling of immunity to malaria. *Mathematical Biosciences* **90,** 385-396.

Aron J.L. & May R.M. (1982) The population dynamics of malaria. In *Population Dynamics of Infectious Diseases.* (Anderson R.M., ed) Chapman and Hall, London, pp 139-179.

Babiker H.A., Abdel-Muhsin A.M., Ranford-Cartwright L.C., Satti G., & Walliker D. (1998) Characteristics of *Plasmodium falciparum* parasites that survive the lengthy dry season in eastern Sudan where malaria transmission is markedly seasonal. *Am. J. Trop. Med. Hyg.* **59,** 582-590.

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*, 2nd edition. Charles Griffin and Co. Ltd, London.

Bailey N.T.J. (1982) *The Biomathematics of Malaria*. Charles Griffin, London.

Baird J.K., Owusu-Agyei S., Utz G., Koram K., Barcus M.J., Binka F.N., Hoffman S.L., & Nkrumah F. (2002) Seasonal malaria attack rates in infants and young children in Northern Ghana. *Am. J. Trop. Med. Hyg* **66,** 280-286.

Barker R.H., Banchongaksorn T., Courval J.M., Suwonkerd W., Rimwungtragoon K., & Wirth D.F. (1994) *Plasmodium falciparum* and *P. vivax*: factors affecting sensitivity and specificity of PCR-based diagnosis of malaria. *Exp. Parasitol.* **79,** 41-49.

Beck H.P. (2002) Extraction and purification of *Plasmodium* parasite DNA. *Methods Mol. Med.* **72,** 159-163.

Beck H.P., Felger I., Huber W., Steiger S., Smith T., Weiss N., Alonso P., & Tanner M. (1997) Analysis of multiple *Plasmodium falciparum* infections in Tanzanian children during the phase III trial of the malaria vaccine SPf66. *J. Infect. Dis.* **175,** 921-926.

Beier J.C., Oster C.N., Onyango F.K., Bales J.D., Sherwood J.A., Perkins P.V., Chumo D.K., Koech D.V., Whitmire R.E., & Roberts C.R. (1994) *Plasmodium falciparum* incidence relative to entomologic inoculation rates at a site proposed for testing malaria vaccines in western Kenya. *Am. J. Trop. Med. Hyg.* **50,** 529-536.

Bekessy A., Molineaux L., & Storey J. (1976) Estimation of incidence and recovery rates of *Plasmodium falciparum* parasitaemia from longitudinal data. *Bulletin WHO* **54,** 685-691.

Binka F.N., Kubaje A., Adjuik M., Williams L.A., Lengeler C., Maude G.H., Armah G.E., Kajihara B., Adiamah J.H., & Smith P.G. (1996) Impact of permethrin impregnated bednets on child mortality in Kassena-Nankana district, Ghana: a randomized controlled trial. *Trop. Med. Int. Health* **1,** 147-154.

Binka F.N., Morris S.S., Ross D.A., Arthur P., & Aryeetey M.E. (1994) Patterns of malaria morbidity and mortality in children in northern Ghana. *Trans. R. Soc. Trop. Med. Hyg.* **88,** 381-385.

Bloland P., Slutsker L., Steketee R.W., Wirima J.J., Heymann D.L., & Breman J.G. (1996) Rates and risk factors for mortality during the first two years of life in rural Malawi. *Am. J. Trop. Med. Hyg.* **55,** 82-86.

Bojang K.A., Milligan P.J.M., Pinder M., Vigneron L., Alloueche A., Kester K.E., Ballou W.R., Conway D.J., Reece W.H.H., Gothard P., Yamuah L., Delchambre M., Voss G., Greenwood B.M., Hill A., McAdam K.P., Tornieporth N., Cohen J.D., & Doherty T. (2001) Efficacy of RTS,S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet* **358,** 1927-1934.

Bouvier P., Breslow N., Doumbo O., Robert C.F., Picquet M., Mauris A., Dolo A., Dembele H.K., Delley V., & Rougemont A. (1997) Seasonality, malaria, and impact of prophylaxis in a West African village .2. Effect on birthweight. *Am. J. Trop. Med. Hyg.* **56,** 384-389.

Breman J.G. (2001) The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.* **64,** 1-11.

Brockman A., Paul R.E., Anderson T.J., Hackford I., Phaiphun L., Looareesuwan S., Nosten F., & Day K.P. (1999) Application of genetic markers to the identification of recrudescent *Plasmodium falciparum* infections on the northwestern border of Thailand. *Am. J. Trop. Med. Hyg.* **60,** 14-21.

Brownstein M.J., Carpten J.D., & Smith J.R. (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques* **20,** 1004-1010.

Bruce M.C., Donnelly C., Narara A., Lagog M., Gibson N., Narara A., Walliker D., Alpers M.P., & Day K.P. (2000a) Age- and species-specific duration of infection in asymptomatic malaria infections in Papua New Guinea. *Parasitology* **121,** 247-256.

Bruce M.C., Galinski M.R., Barnwell J.W., Donnelly C.A., Walmsley M., Alpers M.P., Walliker D., & Day K.P. (2000b) Genetic diversity and dynamics of *Plasmodium falciparum* and *P. vivax* populations in multiply infected children with asymptomatic malaria infections in Papua New Guinea. *Parasitology* **121 (Pt 3),** 257-272.

Burattini M.N., Massad E., & Coutinho F.A. (1993) Malaria transmission rates estimated from serological data. *Epidemiol. Infect.* **111,** 503-523.

Carter R. & McGregor I.A. (1973) Enzyme variation in *Plasmodium falciparum* in The Gambia. *Trans. R. Soc. Trop. Med. Hyg.* **67,** 830-837.

Chambers J.M., Cleveland W.S., Kleiner B., & Tukey P.A. (1983) *Graphical Methods for Data Analysis*. New York.

Ciuca M., Chelaresu M., Sofletea A., Constantinescu P., Teriteanu E., Cortez P., Balanovschi G., & Ilies M. (1955) Contribution expérimentale a l'étude de l'immunité dans le paludisme. *Bucharest, Roumanie: Editions de l'Académie de la Republique Populaire Roumanie*.

Clements A.N. & Paterson G.D. (1981) The analysis of mortality and survival rates in wild populations of mosquitoes. *Journal Of Applied Ecology* **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,** 36-43.

Contamin H., Fandeur T., Bonnefoy S., Skouri F., Ntoumi F., & Mercereau-Puijalon O. (1995) PCR typing of field isolates of *Plasmodium falciparum*. *J. Clin. Microbiol.* **33,** 944-951.

Contamin H., Fandeur T., Rogier C., Bonnefoy S., Konate L., Trape J.F., & Mercereau-Puijalon O. (1996) Different genetic characteristics of *Plasmodium falciparum* isolates collected during successive clinical malaria episodes in Senegalese children. *Am. J. Trop. Med. Hyg.* **54,** 632-643.

Cox D.R. & Miller H.D. (1965) *The Theory of Stochastic Processes*. Methuen and Co. Ltd, London.

Daubersies P., Sallenave-Sales S., Magne S., Trape J.F., Contamin H., Fandeur T., Rogier C., Mercereau-Puijalon O., & Druilhe P. (1996) Rapid turnover of *Plasmodium falciparum* populations in asymptomatic individuals living in a high transmission area. *Am. J. Trop. Med. Hyg.* **54,** 18-26.

Davison A.C. (2003) *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Dietz K., Molineaux L., & Thomas A. (1974) A malaria model tested in the African savannah. *Bull. WHO* **50,** 347-357.

Dietz K. (1988) Mathematical models for transmission and control of malaria. In *Malaria, Principles and Practice of Malariology* (Wernsdorfer W.H. & McGregor I., eds) Churchill Livingstone, Edinburgh, pp 1091-1133.

Draper C.C., Voller A., & Carpenter R.G. (1972) The epidemiologic interpretation of serologic data in malaria. *Am. J. Trop. Med. Hyg.* **21,** 696-703.

Druilhe P., Daubersies P., Patarapotikul J., Gentil C., Chene L., Chongsuphajaisiddhi T., Mellouk S., & Langsley G. (1998) A primary malarial infection is composed of a very wide range of genetically diverse but related parasites. *J. Clin. Invest.* **101,** 2008-2016.

Earle W.C. (1962) The course of naturally acquired malaria. *WHO/Mal/333/mimeographed*.

Earle W.C., Pérez M., del Rio J., & Arzola C. (1938) Observations on the course of naturally acquired malaria in Puerto Rico. *Puerto Rico Journal of Public Health and Tropical Medicine* **14,** 391-406.

East Africa High Commision (1960). Report on the Pare-Taveta Malaria Scheme 1954-1959. Dar es Salaam.East African Institute of Malaria and Vector-Borne Diseases.

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.*, 3rd edition. John Wiley and Sons, Inc., New York.

Eyles D. & Young M. (1951) The duration of untreated or inadequately treated *Plasmodium falciparum* infections in the human host. *Journal of the National Malaria Society* **10,** 327-336.

Farnert A., Snounou G., Rooth I., & Bjorkman A. (1997) Daily dynamics of *Plasmodium falciparum* subpopulations in asymptomatic children in a holoendemic area. *Am. J. Trop. Med. Hyg.* **56,** 538-547.

Felger I. & Beck H.P. (2002) Genotyping of *Plasmodium falciparum*. PCR-RFLP analysis. *Methods Mol. Med.* **72,** 117-129.

Felger I., Genton B., Smith T., Tanner M., & Beck H.P. (2003) Molecular monitoring in malaria vaccine trials. *Trends Parasitol.* **19,** 60-63.

Felger I., Irion A., Steiger S., & Beck H.P. (1999a) Genotypes of merozoite surface protein 2 of *Plasmodium falciparum* in Tanzania. *Trans. R. Soc. Trop. Med. Hyg.* **93 Suppl 1,** 3-9.

Felger I., Marshal V.M., Reeder J.C., Hunt J.A., Mgone C.S., & Beck H.P. (1997) Sequence diversity and molecular evolution of the merozoite surface antigen 2 of *Plasmodium falciparum*. *J. Mol. Evol.* **45,** 154-160.

Felger I., Smith T., Edoh D., Kitua A., Alonso P., Tanner M., & Beck H.P. (1999b) Multiple *Plasmodium falciparum* infections in Tanzanian infants. *Trans. R. Soc. Trop. Med. Hyg.* **93 Suppl 1,** 29-34.

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., Tavul L., Kabintik S., Marshall V., Genton B., Alpers M., & Beck H.P. (1994) *Plasmodium falciparum*: extensive polymorphism in merozoite surface antigen 2 alleles in an area with endemic malaria in Papua New Guinea. *Exp. Parasitol.* **79,** 106-116.

Fine P.E.M. (1975) Superinfection-a problem in formulating a problem. *Tropical Diseases Bulletin* **72,** 475-488.

Franks S., Koram K.A., Wagner G.E., Tetteh K., McGuinness D., Wheeler J.G., Nkrumah F., Ranford-Cartwright L., & Riley E.M. (2001) Frequent and persistent, asymptomatic *Plasmodium falciparum* infections in African infants, characterized by multilocus genotyping. *J. Infect. Dis.* **183,** 796-804.

Fraser-Hurt N., Felger I., Edoh D., Steiger S., Mashaka M., Masanja H., Smith T., Mbena F., & Beck H.P. (1999) Effect of insecticide-treated bed nets on haemoglobin values, prevalence and multiplicity of infection with *Plasmodium falciparum* in a randomized controlled trial in Tanzania. *Trans. R. Soc. Trop. Med. Hyg.* **93 Suppl 1,** 47-51.

Gazin P., Robert V., Cot M., & Carnevale P. (1988) *Plasmodium falciparum* incidence and patency in a high seasonal transmission area of Burkina Faso. *Trans. R. Soc. Trop. Med. Hyg.* **82,** 50-55.

Genton B., Al Yaman F., Beck H.P., Hii J., Mellor S., Narara A., Gibson N., Smith T., & Alpers M.P. (1995) The epidemiology of malaria in the Wosera area, East Sepik Province, Papua New Guinea, in preparation for vaccine trials. I. Malariometric indices and immunity. *Ann. Trop. Med. Parasitol.* **89,** 359-376.

Genton B., Betuela I., Felger I., Al Yaman F., Anders R.F., Saul A., Rare L., Baisor M., Lorry K., Brown G.V., Pye D., Irving D.O., Smith T.A., Beck H.P., & Alpers M.P. (2002) A recombinant blood-stage malaria vaccine reduces *Plasmodium falciparum* density and exerts selective pressure on parasite populations in a phase 1-2b trial in Papua New Guinea. *J. Infect. Dis.* **185,** 820-827.

Gill P.E. & Murray W. Minimization subject to bounds on the variables. *NPL Report NAC 72*. 1976. National Physical Library.

Gilles H.M. & Warell D.A. (1993) *Bruce-Chwatt's Essential Malariology*, 2 edn. Edward Arnold.

Hastings I.M., Watkins W.M., & White N.J. (2002) The evolution of drug-resistant malaria: the role of drug elimination half-life. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357,** 505-519.

Heidelberger P. & Welch P. (1983) Simulation run length control in the presence of an initial transient. *Operations Research* **31,** 1109-1144.

Hill W.G. & Babiker H.A. (1995) Estimation of numbers of malaria clones in blood samples. *Proc. R. Soc. Lond. B. Biol. Sci.* **262,** 249-257.

Irion A., Felger I., Abdulla S., Smith T., Mull R., Tanner M., Hatz C., & Beck H.P. (1998) Distinction of recrudescences from new infections by PCR-RFLP analysis

in a comparative trial of CGP 56 697 and chloroquine in Tanzanian children. *Trop. Med. Int. Health* **3,** 490-497.

Jafari S., Le Bras J., Bouchaud O., & Durand R. (2004) *Plasmodium falciparum* clonal population dynamics during malaria treatment. *J. Infect. Dis.* **189,** 195-203.

James P., Nicol W.D., & Shute P.G. (1932) A study of induced malignant tertian malaria. *Proc. Roy. Soc. of Med.* **25,** 1153-1186.

James P., Nicol W.D., & Shute P.G. (1936) Clinical and parasitological observations on induced malaria. *Proc. Roy. Soc. of Med.* **29,** 879-894.

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.* 781-789.

Jeffrey G.M. & Eyles D.E. (1954) The duration in the human host of infections with a Panama strain of *Plasmodium falciparum*. *Am. J. Trop. Med. Hyg.* **3,** 219-224.

Johnson N.L., Kotz S., & Balakrishnan N. (1995) *Continuous Univariate Distributions*, 2nd edition. John Wiley and Sons, Inc., New York.

Kitua A.Y., Smith T., Alonso P.L., Masanja H., Urassa H., Menendez C., Kimario J., & Tanner M. (1996) *Plasmodium falciparum* malaria in the first year of life in an area of intense and perennial transmission. *Trop. Med. Int. Health* **1,** 475-484.

Klein J.P. & Moeschberger M.L. (1997) *Survival Analysis: Techniques for Censored and Truncated Data.* Springer Verlag, New York.

Koram K.A., Owusu-Agyei S., Fryauff D.J., Anto F., Atuguba F., Hodgson A., Hoffman S.L., & Nkrumah F.K. (2003) Seasonal profiles of malaria infection, anaemia, and bednet use among age groups and communities in northern Ghana. *Trop. Med. Int. Health* **8,** 793-802.

Lengeler C. (2004) Insecticide treated nets and curtains for preventing malaria. *Cochrane database Syst Rev* CD000363.

Macdonald G. (1950a) The analysis of infection rates in diseases in which superinfection occurs. *Tropical Diseases Bulletin* **47,** 907-915.

Macdonald G. (1950b) The analysis of malaria parasite rates in infants. *Tropical Diseases Bulletin* **47,** 915-938.

Macdonald G. (1952) The analysis of the sporozoite rate. *Tropical Diseases Bulletin* **49,** 569-585.

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.

Macdonald I.L. & Zucchini W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall/CRC, Boca Raton.

McGregor I.A. (1984) Epidemiology, malaria and pregnancy. *Am. J. Trop. Med. Hyg.* **33,** 517-525.

Metselaar D. (1957) A pilot project of residual insecticide spraying in Netherlands New Guinea: contribution to the knowledge of holo-endemic malaria. *Ph.D. dissertation. Leiden University.*

Molineaux L. (1988) The epidemiology of human malaria as an explanation of its distribution, including some implications for its control. In *Malaria, Principles and Practice of Malariology* (Wernsdorfer W.H. & McGregor I., eds) Churchill Livingstone, Edinburgh, pp 913-998.

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.

Molineaux L. & Gramiccia G. (1980) *The Garki Project*. World Health Organisation, Geneva.

Molineaux L., Muir D.A., Spencer H.C., & Wernsdorfer W.H. (1988) The epidemiology of malaria and its measurement. In *Malaria, Principles and Practice of Malariology* (Wernsdorfer W.H. & McGregor I., eds) Churchill Livingstone, Edinburgh, pp 999-1089.

Msuya F.H. & Curtis C.F. (1991) Trial of pyrethroid impregnated bednets in an area of Tanzania holoendemic for malaria. Part 4. Effects on incidence of malaria infection. *Acta Trop.* **49,** 165-171.

Nedelman J. (1984) Inoculation and recovery rates in the malaria model of Dietz, Molineaux and Thomas. *Mathematical Biosciences* **69,** 209-233.

Nedelman J. (1985) Estimation for a model of multiple malaria infections. *Biometrics* **41,** 447-453.

Ntoumi F., Contamin H., Rogier C., Bonnefoy S., Trape J.F., & Mercereau-Puijalon O. (1995) Age-dependent carriage of multiple *Plasmodium falciparum* merozoite surface antigen-2 alleles in asymptomatic malaria infections. *Am. J. Trop. Med. Hyg.* **52,** 81-88.

Nyarko P., Wontou P., Nazzar A., Phillips J., Ngom P., & Binka F. (2002) Navrongo DSS, Ghana. In *INDEPTH Network. Population and Health in Developing Countries. Volume 1. Population, Health, and Survival at INDEPTH Sites* pp 247-256.

Owusu-Agyei S., Awini E., Anto F., Afful T.M., Adjuik M., Hodgson A., Afari E.A., & Binka F.N. (2006) Monitoring and evaluating Roll Back Malaria activities at the district level in Africa: a situation analysis of the Kassena-Nankana district in Northern Ghana. *Health Policy and Planning* (submitted).

Owusu-Agyei S., Smith T., Beck H.P., 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.

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.

Peters W. (1990) The prevention of antimalarial drug resistance. *Pharmacol. Ther.* **47,** 499-508.

Port G.R., Boreham P.F.L., & Bryan J.H. (1980) The relationship of host size to feeding by mosquitos of the *Anopheles-gambiae* giles complex (Diptera, Culicidae). *Bulletin of Entomological Research* **70,** 133-144.

Ranford-Cartwright L.C., Balfe P., Carter R., & Walliker D. (1993) Frequency of cross-fertilization in the human malaria parasite *Plasmodium falciparum*. *Parasitology* **107 (Pt 1),** 11-18.

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.

Richard A., Richardson S., & Maccario J. (1993) A three-state Markov model of *Plasmodium falciparum* parasitaemia. *Mathematical Biosciences* **117,** 283-300.

Ridley R.G. (2002) Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature* **415,** 686-693.

Rogier C. & Trape J.F. (1995) [Study of premunition development in holo- and meso-endemic malaria areas in Dielmo and Ndiop (Senegal): preliminary results, 1990-1994[]. *Med. Trop.(Mars.)* **55,** 71-76.

Ross R. (1911) *The prevention of malaria*, 2nd edn. London.

Ross R. (1916) An application of the theory of probabilities to the study of a priori pathometry, Part I. *Proc. Roy. Soc. A.* **92,** 204-230.

Saiki R.K., Scharf S., Faloona F., Mullis K.B., Horn G.T., Erlich H.A., & Arnheim N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230,** 1350-1354.

Sama W., Killeen G., & Smith T. (2004) Estimating the duration of Plasmodium falciparum infection from trials of indoor residual spraying. *Am. J. Trop. Med. Hyg.* **70,** 625-634.

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. *Statistics in Medicine* **24,** 3269-3288

Sergent E., Donatien A., Parrot L., Lestoquard F., Plantureux E., & Rougebief H. (1924) Études Expérimentales Sur Les Prioplamoses Bovines d'Algérie. *Annales de l'Institut Pasteur* **38,** 273-343.

Shampine L.F. (1994) *Numerical solution of ordinary differential equations*. Chapman and Hall.

Simon F., Le Bras J., Gaudebout C., & Girard P.M. (1988) Reduced sensitivity of *Plasmodium falciparum* to mefloquine in West Africa. *Lancet* **1,** 467-468.

Smith B.J. (2003) *Bayesian Output Analysis Program (BOA), version 1.0 Users Manual*.

Smith T., Beck H.P., Kitua A., Mwankusye S., Felger I., Fraser-Hurt N., Irion A., Alonso P., Teuscher T., & Tanner M. (1999a) Age dependence of the multiplicity of *Plasmodium falciparum* infections and of other malariological indices in an area of high endemicity. *Trans. R. Soc. Trop. Med. Hyg.* **93 Suppl 1,** 15-20.

Smith T., Felger I., Fraser-Hurt N., & Beck H.P. (1999b) 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., Felger I., Kitua A., Tanner M., & Beck H.P. (1999c) Dynamics of multiple *Plasmodium falciparum* infections in infants in a highly endemic area of Tanzania. *Trans. R. Soc. Trop. Med. Hyg.* **93 Suppl 1,** 35-39.

Smith T., Felger I., Tanner M., & Beck H.P. (1999d) Premunition in *Plasmodium falciparum* infection: insights from the epidemiology of multiple infections. *Trans. R. Soc. Trop. Med. Hyg.* **93 Suppl 1,** 59-64.

Smith T., Killeen G., Lengeler C., & Tanner M. (2004) Relationships between the outcome of *Plasmodium falciparum* infection and the intensity of transmission in Africa. *Am. J. of Trop. Med. Hyg.* **70 (Suppl 2),** 80-86.

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.

Snounou G., Viriyakosol S., Zhu X.P., Jarra W., Pinheiro L., do R., V, Thaithong S., & Brown K.N. (1993) High sensitivity of detection of human malaria parasites by

the use of nested polymerase chain reaction. *Mol. Biochem. Parasitol.* **61,** 315-320.

Snounou G., Zhu X., Siripoon N., Jarra W., Thaithong S., Brown K.N., & Viriyakosol S. (1999) Biased distribution of msp1 and msp2 allelic variants in *Plasmodium falciparum* populations in Thailand. *Trans. R. Soc. Trop. Med. Hyg.* **93,** 369-374.

Soper F.L. & Wilson D.B. (1943) *Anopheles gambiae in Brazil, 1930-1940*. The Rockerfeller Foundation, New York.

Spiegelhalter D.J., Thomas A., & Best N. (2000) WinBUGS User Manual, version 1.3. *Medical Research Council, Cambridge, U.K.*

Spiegelhalter D.J., Thomas A., Best N., & Lunn D. (2003) *WinBUGS User Manual, version 1.4. Medical Research Council, Cambridge, U.K.*

Steketee R.W., Nahlen B.L., Parise M.E., & Menendez C. (2001) The burden of malaria in pregnancy in malaria-endemic areas. *Am. J. Trop. Med. Hyg.* **64,** 28-35.

Stich A.H., Maxwell C.A., Haji A.A., Haji D.M., Machano A.Y., Mussa J.K., Matteelli A., Haji H., & Curtis C.F. (1994) Insecticide-impregnated bed nets reduce malaria transmission in rural Zanzibar. *Trans. R. Soc. Trop. Med. Hyg.* **88,** 150-154.

Trape J.F. (2001) The public health impact of chloroquine resistance in Africa. *Am. J. Trop. Med. Hyg.* **64,**12-17.

Walton G.A. (1947) On the control of malaria in Freetown, Sierra Leone. 1. *Plasmodium falciparum* and *Anopheles gambiae* in relation to malaria occurring in infants. *Ann. Trop. Med. Parasitol.* **41,** 380-407.

White N.J. (1992) Antimalarial drug resistance: the pace quickens. *J. Antimicrob. Chemother.* **30,** 571-585.

White N.J. (1999a) Antimalarial drug resistance and combination therapy. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **354,** 739-749.

White N.J. (1999b) Antimalarial drug resistance and mortality in *falciparum* malaria. *Trop. Med. Int. Health* **4,** 469-470.

WHO (1991) The urban crisis. *World Health Statistics Quarterly* **44,** 189-197.

WHO (1995) World Health Organisation Report. *Division and control of tropical diseases.*

WHO (1997). Guidelines for the evaluation of Plasmodium falciparum vaccines in populations exposed to natural infections. TDR/MAL/VAC/97. World Health Organisation.

WHO (1999a) Malaria, 1982-1997. *Weekly Epidemiological Record* **74,** 265-270.

WHO (1999b) The World Health Report 1999-making a difference. Geneva. *World Health Organisation*.

Wilk M.B. & Gnanadesikan R. (1968) Probability Plotting Methods for Analysis of Data. *Biometrika* **55,** 1-17.

Curriculum Vitae

Name: Wilson Bigina Sama-Titanji

Date of Birth: 20 May 1974

Place of Birth Cameroon

Nationality/Residence: Cameroonian, resident in Switzerland (B permit holder)

Marital Status: Married to Marion Enow-mba Tabe

Address: C/o Swiss Tropical Institute, Socinstrasse 57
CH-4002, Basel, Switzerland.
Email: wilson.sama@unibas.ch / samawnb@yahoo.com

Higher Education:

2002-2006: Ph.D. in Epidemiology, Swiss Tropical Institute, Basel, Switzerland

2002-2004: M.Sc. in Applied Statistics, University of Neuchâtel, Switzerland

1997-1999: M.Sc. Mathematics, University of Buea, Cameroon.

1994-1997: B.Sc. Mathematics and a minor in Computer Science, University of Buea Cameroon.

Job Experience.

2002-2005: Doctoral student in the Biostatistics unit at the Swiss Tropical Institute, Basel, Switzerland. Teaching of Biostatistics to post-graduate students with a biomedical background from the University of Basel.

1999-2001: Mathematics Instructor, Department of Mathematics and Computer Science, University of Buea, Cameroon.

1997-2001. Mathematics and Physics Instructor, Buea Bilingual High School, Buea, Cameroon.

Position of responsibility

Student leader, Department of Public Health and Epidemiology, Swiss Tropical Institute. Organizing and chairing student weekly meetings to discuss research topics and other student activities (October 2004 to June 2005).

Professional Societies:

Member of Swiss Statistical Society .

Publications.

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. *Statistics in Medicine* **24**: 3269-3288.

Sama W., Owusu-Agyei S., Felger I., Dietz K., Smith T. (2005). Age and seasonal variation in the transition rates and detectability of *Plasmodium falciparum* malaria. *Parasitology* (in press).

Sama W., Killeen G., Smith T. (2004). Estimating the duration of *Plasmodium falciparum* infection from trials of indoor residual spraying. *American Journal of Tropical Medicine and Hygiene* **70(6)**: 625-634

Smith T., **Sama W.** Estimating the duration of common persistent infections. Pp 52-53 in Workshop: Design and Analysis of Infectious Disease Studies. Report No. 49/2004, Mathematisches Forschungsinstitut Oberwolfach, Germany.

Ngwa G.A., Ngeh, N.C., **Sama, W.** (2001). A model for endemic malaria with delay and variable population. *Journal of the Cameroon Academy of Sciences*; 1 (3), 169-185.

Sama W., Dietz K., Smith T. The distribution of survival times of deliberate *Plasmodium falciparum* infections in tertiary syphilis patients. *Transactions of the Royal Society of Tropical Medicine and Hygiene* (in press).

Falk N., Maire N., **Sama W.**, Owusu-Agyei S., Smith T., Beck H.-P., Felger I. Comparison of PCR-RFLP and GeneScan–based genotyping for analyzing infection dynamics of *Plasmodium falciparum*. *American Journal of Tropical Medicine and Hygiene* (in press).