Statistical analysis of *Plasmodium falciparum* infection dynamics

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Michael Bretscher
aus Winterthur (ZH)

Basel, 2012

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Dr. Thomas A. Smith und Prof. Azra Ghani

Basel, 21.9.2010

Prof. Dr. Martin Spiess
Dekan

Man will occasionally stumble over the truth,
but most of the time he will pick himself up and continue on.

Winston Churchill

For my wife Fridah,
and in memory of my beloved father

Zusammenfassung

Malaria ist global gesehen einer der grössten Versursacher von Krankheit und Tod unter den Infektionskrankheiten. Weltweit erkrankten geschätzte 200 Millionen Menschen an Malaria im Jahr 2008, grösstenteils (zu 85%) in Afrika. Dies führte zu einer geschätzten Million Todesfälle, mit einem ähnlich grossen Anteil (89%) in Afrika. Es gibt verschiedene Species, welche Malaria verursachen, aber die meisten Todesfälle gehen auf das Konto von *Plasmodium falciparum*.

Malaria bleibt eine grosse Herausforderung für die wissenschaftliche Forschung: der Parasit evolviert fortlaufend Resistenzen gegen bestehende Medikamente, so dass immer neue Substanzen gefunden werden müssen um Malaria zu heilen. Eine Impfung gegen *Plasmodium falciparum* zu finden erweist sich als äusserst schwierig weil der Parasit Wege gefunden hat, sich der menschlichen Immunantwort zu entziehen. Wie genau dies passiert ist wenig genau verstanden. Zusätzlich sind viele betroffene Länder arm und verfügen über unzureichende Gesundheitssysteme um effektiv gegen die Seuche vorzugehen.

In den 1950er Jahren fasste die Weltgesundheitsorganisation (**WHO**) die weltweite Ausrottung der Malaria ins Auge: das neu entdeckte Insektizit DDT schien sehr gut geeignet dafür, die Anzahl der Fälle zu reduzieren. Dies durch Abtöten der *Anopheles* Mücken, durch welche die Malaria übertragen wird. Zusätzlich sagten die mathematischen Modelle jener Zeit voraus, dass es im Prinzip möglich wäre, die Krankheit vollkommen auszurotten. Trotz grosser Erfolge in der Karibik, in Teilen Asiens, Süd- und Zentralamerikas, und erfolgreicher Ausrottung in Europa und Nordamerika während der folgenden Jahrzehnte, blieb der Erfolg in Afrika und Teilen Asiens aus.

Nach diesem Rückschlag wurde Malaria für lange Zeit vernachlässigt. Erst seit kurzem steht Malaria wieder ganz oben auf der globalen Gesundheitsagenda. Nach den massiven Fortschritten, welche die Biologie in den letzten Jahrzehnten gemacht hat, verfügt man heute über neue Werkzeuge um den Lebenszyklus des Parasiten besser zu verstehen und ihn möglicherweise wirkungsvoller zu bekämpfen.

Ein wichtiger Faktor, welcher in der Vergangenheit einem besseren Verständnis der Epidemiologie des Parasiten im Wege stand ist die Tatsache, dass Mikroskopie als Diagnosemethode nicht in der Lage ist, einzelne Infektionen zu unterscheiden: In Gebieten, wo Malaria endemisch vorkommt, tragen die Menschen oft gleich mehrere Infektionen in sich, häufig asymptomatisch. DNS-basierte Diagnosemethoden benützen gezielt genetische Loci, bei welchen zahlreiche Varianten innerhalb der Parasitenpopulation vorkommen, um die einzelnen Infektionen zu unterscheiden. Ein solcher Locus heisst “merozoite surface protein 2” (**msp2**).

Diese Dissertation entwickelt statistische Modelle um solche Daten über das Vorhandensein von (hauptsächlich) msp2 Genotypen zu analysieren. Im Speziellen wird der Datensatz aus einer Kohortenstudie, welche in Navrongo, im Norden Ghanas, durchgeführt wurde analysiert. Zusätzlich werden in Kapitel 6 Daten aus Papua Neu Guinea verwendet. Eine grosse Herausforderung bei der Analyse solcher Daten ist, dass der Parasit nicht immer detektierbar ist: er versteckt sich in den Kapillargefässen durch Anheften an die Gefässwände und ist im peripheren Blut nicht immer vorhanden.

Die drei Parameter welche durch unsere statistischen Modelle aus den erwähnten Zeitreihen-

daten geschätzt werden sind i) die “force of infection” (die Anzahl neuer Infektionen, welche pro Person und Jahr erworben werden), ii) die Dauer einer einzelnen Infektion, und iii) die “detectability” (Wahrscheinlichkeit eine vorhandene Infektion im peripheren Blut nachweisen zu können).
Frühere Ansätze zur Analyse von longitudinalen genetischen Daten waren limitiert in dem Sinne, dass nur Exponentialverteilungen benutzt werden konnten, um die Infektionsdauer zu modellieren: Dies ist gleichbedeutend mit der Annahme einer konstanten Eliminierungsrate (pro Zeit). Die Grund hierfür war einzig mathematische Einfachheit: Unter Annahme einer Exponentialverteilung für die “Überlebenzeiten” der Infektionen kann die Altersstruktur der Infektionspopulation innerhalb eines Menschen vernachlässigt werden, da die Eliminierungsrate ja konstant und somit unabhängig vom Alter einer Infektion ist. Anders ausgedrückt: Man verwendete man für den “Zerfall” von Infektionen mathematisch dasselbe Modell wie für den radioaktiven Zerfall von Atomen. Biologisch macht dies wenig Sinn, und wenn man mehr über die Dynamik der Anzahl Parasiten innerhalb eines Menschen oder die Auswirkungen von Immunität verstehen möchte, sollte man zwischen alten und jungen Infektionen unterscheiden können.

Die Dissertation entwickelt eine Erweiterung zu Bestehenden statistischen Methoden und benützt parametrische Verteilungen aus der Ereigniszeitanalyse um zu Beschreiben, wie die Eliminierung von Infektionen von deren Alter abhängt. Zusätzlich wird der Einfluss des Alters des infizierten Menschen untersucht, da dieses als Indikator für Immunität[1] interpretiert werden kann: Je älter ein Mensch, desto mehr Infektionen hat er durchgemacht und desto mehr Immunität hat er erworben. Änderungen in der Eliminierung von Infektionen in Abhängigkeit des Alters des Trägers kann deshalb als Immunitätseffekt interpretiert werden.

Es wurde ein Unterschied zwischen der Verteilung der Infektionsdauer im Datensatz aus Ghana im Quervergleich zu künstlich herbeigeführten Infektionen[2] festgestellt: Ein grosser Teil der Infektionen in der Ghanaischen Population dauern nur sehr kurz. Da dies in allen Altersklassen der Fall ist, kann ein Effekt von über längerer Zeit erworbener Immunität ausgeschlossen werden. Dies ist das erste Mal, dass dies nachgewiesen werden konnte. Zudem konnte das Resultat mit einer verschiedenen statistischen Methodik und anderen Daten bestätigt werden. Als wahrscheinlichste Erklärung dafür kommt eine Interaktion zwischen Infektionen, welche den gleichen Menschen befallen, in Frage. Die Konsequenzen dieser Entdeckung für das Verständnis der Prozesse im innern des Menschlichen Körpers bei Infektion mit *Plasmodium falciparum* werden diskutiert.

[1] Immunität gegen Malaria schützt nicht vor Neuinfektionen, sondern gegen Fieber und schwere Krankheitsverläufe.

[2] Bevor geeignete Antibiotika zur Verfügung standen, mussten Syphilispatienten durch Infektion mit Malaria behandelt werden. Vermutlich heilten die Fieberschübe die Syphilis.

Summary

Malaria is one of the major contributors to the global burden of disease. Worldwide, there were an estimated number of 200 million malaria cases in the year 2008, with a vast majority (85%) of those being in the African Region. This has lead to an estimated number of up to one million deaths, with a similar majority (89%) happening in the African region. There are several parasite species causing malaria, but most deaths are caused by *Plasmodium falciparum*.

Malaria remains a major challenge for scientific research: constantly the parasite evolves resistance against existing drugs, and ever new substances to cure malaria need to be found. Creating a vaccine against *Plasmodium falciparum* proves exceptionally difficult, because the parasite has found ways to escape the human immune response. How exactly, is poorly understood. In addition, many countries affected by the disease suffer from poverty and ineffective health infrastructure.

In the 1950's the final eradication of malaria was envisioned by the WHO: the newly discovered insecticide DDT showed very promising results in reducing the malaria burden by killing the *Anopheles* mosquitoes, through which malaria is transmitted, and mathematical models of malaria transmission predicted that eradication of the disease would be possible. Despite great successes in the Carribean, parts of Asia and South-Central America, and elimination in Europe and North America during the following decades, the efforts did not succeed in tropical Africa and many parts of Asia. After that failure, malaria was a "neglected" disease for a long period. Only since recent times malaria is again high on the global health agenda. Now, the enormous progress in the life sciences during the last decades provides new tools to better understand the parasite's natural history, and perhaps will reveal new ways of attacking it.

One factor which limited the understanding of the epidemiology of the parasite was that microscopy as diagnostic tool is not able to distinguish multiple concurrent infections within one human host: people in endemic areas often harbour several infecting clones in parallel. DNA-based methods make use of genetic loci of which many different variants exist in the parasite population, e.g. merozoite surface protein 2 (**msp2**), to distinguish co-infecting clones. This thesis develops statistical models to analyse such data on the presence of (mostly) msp2 genotypes. In particular, data from a longitudinal study in Navrongo, Northern Ghana is used in all chapters except chapter 6, where data from Papua New Guinea is analysed. A major challenge in the analysis of this type of data is the phenomenon of imperfect detection: the parasite hides in the deep blood vessels by attaching to the capillary walls, and it can therefore not be always detected in the peripheral blood.

The three parameters which are estimated by our statistical models from time-series on presence or absence of genotypes are i) the force of infection (the number of infections acquired per person and year), ii) the duration of infection for one parasite clone, and iii) the detectability (the probability of detecting a parasite, given it is present).

Previous statistical methods for the analysis of longitudinal genotyping data were restricted to exponential distributions of infection duration: this means that a constant rate (per time) is assumed at which infections are cleared. The reason for this was mathematical simplicity: the age structure of the infection population within a host can be neglected because

the clearance rate is constant and does not depend on the age of an infection. In other words, the same mathematical model as for radioactive decay was used. Biologically, this is a very unrealistic assumption, and to understand more about within-host dynamics of *P. falciparum* or immunity against it one would like to distinguish between young and old infections.

This thesis develops an extension to previous statistical analysis methods and makes use of parametric survival distributions to describe infection clearance and how it depends on the age of an infection. In addition to the age of infection, the effect of host age on infection clearance is investigated: older persons have experienced more infections and are therefore more immune[3]. Changes in infection clearance with host age can therefore be interpreted as effects of immunity.
An difference between the distribution of infection durations in the Ghanaian dataset compared to artificial infections[4] emerged: a large proportion of infections in the Ghanaian population are cleared quickly after inoculation. It is the first time this could be measured from field data, and the result was confirmed using a different statistical method and study design. The difference between artificial infections and the field data cannot be attributed to acquired immunity in the Ghanaian population because all age groups show a similar abundance of very short infection durations. An interaction between the multiple infections within one host in Northern Ghana appears to be the most likely explanation. The implications of this finding for our understanding of the within-host processes in *falciparum* malaria are discussed.

[3]Immunity against malaria protects from severe disease or fever, not re-infection.
[4]Before suitable antibiotics were available, infection with malaria was a method to treat syphilis.

Acknowledgements

It appears nearly impossible to thank all the people who have contributed to this thesis in one way or another, and accidentally omitting somebody appears almost inevitable. I therefore apologize in advance and solemnly promise and vow to acknowledge those forgotten with certainty in the next PhD thesis.

Foremost I would like to thank my wife Fridah for being so great. I thank her also for the strong support and friendship throughout the years of study, when no end was in sight. Its done!
I thank my supervisor, Tom Smith, for sharing his vast knowledge and experience with me, and for always being supportive of the project, even in times of doubt. And for making jokes about frogs. I thank all the members of the "Biostatistics and Computational Sciences" unit of the Swiss TPH, for their friendship and for interesting discussions, especially Nakul Chitnis for teaching me mathematics and some french on a bike ride through New Orleans after midnight with the great Joshua Yukich, Konstantina Boutsika for remembering all the birthdays, Aurelio Di Pasquale for the wonderful time in Palermo, Guillaume Gnaegi, Maria Laura and Gabriel Dominic Gosoniu, Leticia Grize, Diggory Hardy, Michael Hegnauer, Dirk Keidel, Nicolas Maire for friendship and teaching me a lot about programming, Melissa Penny, Amanda Ross, Emmanuel Schaffner, Christian Schindler, Christoph Schmid, Ronaldo Scholte, Michael Tarantino, Penelope Vounatsou, Amek Ombek, Valerie Crowell, Simon Kasasa, Federica Giardina, Bernadette Huho, Angelina Lutambi, Nadine Riedel, and Susan Rumisha.
Furthermore, I would like to thank all the collaborators outside the group, particularly Ingrid Felger, Sonja Schöpflin, Francesca Valsiangiacomo, and Cristian Köpfli for friendship and an interesting exchange of knowledge. And for making this thesis possible by generating data of great value. I thank all the people who participated in the field studies. The insights from the obtained datasets will continue to advance malaria research.
Special thanks deserve the past and present members of the "harmonic oscillators", a hypothetical band with ever-changing name and line-up, for creating an inspiring counterweight to science. We all regret that our music never made us famous - yet. Among the members are Joshua Jukich, Nicolas Maire, Daniel Weibel, Daniel Dröschel, and Thomas Brunner.
I thank all my family members, especially my mother Marie-Therese Bretscher, my sister Simone Bretscher and Josef Seiler for friendship and for providing a base in the countryside, where all sorts of theories could be discussed. I thank my late father Hans-Jörg Bretscher for teaching me how to think in a way that should turn out to be quite useful when doing research. Special thanks are extended to Rosemarie, Alain and Murielle Pastier, and Therese Bretscher for significantly contributing to this thesis by letting me spend a week of writing in seclusion at her place in Brunnen, right at the beach. Sincere thanks to the Kenyan part of the familiy for the cordial reception of a stranger. These are Julia Karimi N'thaka, James Mwenda Murega, Irene Ntinyari Murega, Glory Kathambi Murega, and many others.
I thank all my friends, some of them office mates in the student-office, for creating an atmosphere of mutual support and for sharing with me countless unforgettable moments

during the PhD times. Among those are Benedict and Irene Notter, Stephan Dongus, Irene Küpfer, Bettina Ley, Borna Müller, Christoph Dumelin, Nicole Joller, Patrik Zimmerli, Andrea Kümmerle, Bianca Plüss, Salomé Dürr, Silvia Alonso-Alvarez, Karin Gross, Martin Maire, Rea Tschopp, Sandra Alba, Amanda Brosius, Benjamin Dahl, Boris Bear, Christian Flück, Constanze Pfeiffer, Eric Diboulo, Eelco Jacobs, Eva Maria Hodel, Julie Balen, Manuel Hetzel, Matthias Bischoff, Musa Mabaso, Paola Favuzza, Patricia and Anu Lannen, Peter Steinmann, Raffael Ayé, Ricarda Windisch, Serej Ley, Tanja Jaeggi, Thomas Ephraim Erlanger, Wilson Sama, Yvonne Geissbühler, Mercy Ackumey, Angel Dillip, Eveline Hürlimann, Steffi Knopp, Barbara Matthys, Pie Müller, Claudia Sauerborn, Christian Schätti, Mirko Winkler, Don de Savigny, Christian Lengeler, and others.

Last but definitely not least, I would like to thank all the people who keep Swiss TPH running, notably the director, Marcel Tanner, who is always available when problems occur, as well as the members of the IT team, Marco Clementi, Mike Schur, Lukas Camenzind, Steven Paul, Simon Schlumpf, and Dominique Forster for excellent support throughout the thesis, Heidi Immler and the library team as well as the countless good souls in the administration and in the “Technischer Dienst”, and Yvette Endriss.

Contents

1 Introduction **1**

1.1 Malaria . . . 1
1.1.1 Overview . . . 1
1.1.2 The history of malaria . . . 1
1.1.3 The biology of *Plasmodium falciparum* . . . 3
1.1.4 The epidemiology of *Plasmodium falciparum* . . . 4
1.2 Mathematical modeling of malaria . . . 5
1.3 Infection dynamics in the human population . . . 7
1.3.1 The parameters of infection dynamics . . . 7
1.3.2 The force of infection . . . 7
1.3.3 The duration of infection . . . 8
1.3.4 Detectability . . . 10
1.4 Molecular data . . . 11
1.4.1 The analysis of molecular data . . . 11
1.5 Objectives of the thesis . . . 14
1.5.1 The distribution of infection durations . . . 14

2 Detectability of *Plasmodium falciparum* clones **17**

2.1 Abstract . . . 17
2.2 Background . . . 18
2.3 Methods . . . 19
2.3.1 Study site and sample collection . . . 19
2.3.2 Genotyping . . . 19
2.3.3 Data analysis . . . 20
2.3.4 Models of detection . . . 21
2.3.5 Bias correction of detectability estimates . . . 23
2.4 Results . . . 24
2.4.1 Tests of proportion and correlation . . . 24
2.4.2 Model comparison . . . 26
2.4.3 Estimates of q . . . 26
2.5 Discussion . . . 27
2.5.1 Within-host dynamics . . . 28
2.5.2 Measurement of detectability . . . 29
2.5.3 Epidemiological significance of detectability . . . 30
2.6 Conclusions . . . 30
2.7 Appendix . . . 32
2.7.1 Maximum likelihood estimation of q . . . 33

3 The dynamics of natural *Plasmodium falciparum* infections **37**
3.1 Abstract . . . 37
3.2 Introduction . . . 38
3.3 Materials and Methods . . . 38
3.3.1 Field methodology . . . 38
3.3.2 DNA isolation and genotyping . . . 40
3.3.3 Data analysis . . . 41
3.4 Results . . . 42
3.5 Discussion . . . 47

4 The distribution of *Plasmodium falciparum* infection durations **53**
4.1 Abstract . . . 53
4.2 Background . . . 54
4.3 Methods . . . 55
4.3.1 Study design and sample collection . . . 55
4.3.2 Genotyping . . . 55
4.3.3 Data preparation . . . 56
4.3.4 Models of infection dynamics . . . 56
4.3.5 Model equations . . . 58
4.3.6 Model implementation and parameter estimation . . . 62
4.4 Results . . . 63
4.4.1 Simulated data . . . 63
4.4.2 Estimates from the Ghanaian dataset . . . 63
4.5 Discussion . . . 66
4.5.1 Distribution estimates . . . 66
4.5.2 Validation . . . 69
4.5.3 The difference to malariatherapy data . . . 69
4.5.4 Limitations of the method . . . 70
4.6 Conclusions . . . 70
4.7 Appendix . . . 72
4.7.1 Exponential survival of infections . . . 72
4.7.2 Non-exponential survival of infections . . . 73

5 Effects of host age on clearance of malaria infections **75**
5.1 Abstract . . . 75
5.2 Background . . . 76
5.3 Methods . . . 77
5.3.1 Study design and sample collection . . . 77
5.3.2 Genotyping . . . 77
5.3.3 Data preparation . . . 77
5.3.4 Models of infection dynamics . . . 78
5.4 Results . . . 80
5.4.1 Force of infection . . . 82
5.4.2 Clearance of infections . . . 82

5.4.3 Detectability . . . 84
5.5 Discussion . . . 91
5.5.1 Force of infection . . . 91
5.5.2 Clearance of infections . . . 92
5.5.3 Detectability . . . 94
5.5.4 Time to near-elimination . . . 94
5.6 Conclusions . . . 95

6 Are all malaria parasites equal? Human *Plasmodia* compared. 97
6.1 Background . . . 98
6.2 Methods . . . 99
6.2.1 Field survey and patients . . . 99
6.2.2 Laboratory procedures . . . 99
6.2.3 Data analysis . . . 100
6.3 Results . . . 101
6.3.1 Effect of repeated sampling on prevalence . . . 101
6.3.2 Effect of repeated sampling on detection of individual clones . . . 101
6.4 Discussion . . . 103

7 General discussion 107
7.1 Review of chapters . . . 107
7.2 Results of the thesis . . . 109
7.2.1 Interpretation . . . 109
7.2.2 Virulence of *P. falciparum* . . . 110
7.2.3 Consequences for control and elimination . . . 112
7.3 Methodology . . . 113
7.3.1 Short history of methods . . . 113
7.3.2 Directions of future development . . . 114

Bibliography 117

List of Figures

1.1 The global distribution of malaria . . . 2
1.2 The life cycle of *Plasmodium falciparum* . . . 3
1.3 Prevalence of *P. falciparum* in the Garki Project . . . 4
1.4 The Ross-Macdonald model of malaria transmission . . . 6
1.5 Seasonality in the force of infection in Navrongo, Northern Ghana . . . 9
1.6 Age dependence of detectability . . . 12
1.7 The distribution of infection durations . . . 15

2.1 Study design . . . 19
2.2 Expected and actual frequencies of sequence types . . . 25
2.3 The error in prevalence measurements becomes more important at low MOI 31

3.1 Sampling intervals and rainfall in Navrongo/Northern Ghana during the study period . . . 40
3.2 Multiplicity of infection and prevalence . . . 44
3.3 Transitions . . . 45
3.4 Detectability and parasite density . . . 46
3.5 Duration of infection by host age . . . 48
3.6 Multiplicity of infection . . . 49
3.7 List of all 100 msp2 genotypes detected in cohort from Northern Ghana (349 individuals) . . . 52

4.1 Validation using simulated data . . . 59
4.2 Results from the Ghanaian dataset compared to malariatherapy data . . . 64
4.3 Estimates of the force of infection . . . 67
4.4 Estimates of detectability . . . 68

5.1 The force of infection . . . 81
5.2 Clearance of infections . . . 83
5.3 Average duration of infection . . . 85
5.4 Variation in the duration of infection . . . 86
5.5 Mean residual lifetime . . . 87
5.6 Persistence of infections . . . 88
5.7 Time until most infections are cleared . . . 89
5.8 Detectability . . . 90

7.1 Early clearance of infections confirmed . . . 111

List of Tables

2.1 Data coding 20
2.2 Comparison of Models M1-M3 26
2.3 Direct estimation of q on all survey pairs, using M0. 27

3.1 Characteristics of the cohort studied 39
3.2 Parameter estimates of statistical models 43

4.1 Survival distributions 57
4.2 Parameter estimates from simulated datasets 58
4.3 Parameter estimates from the Ghanaian dataset 65
4.4 Correlation matrix 66

5.1 Parametric survival distributions 79
5.2 Force of infection and detectability 81
5.3 Parameters related to clearance of infections 82
5.4 Mean durations 84

6.1 Effect of repeated sampling on prevalence as determined by microscopy . . 101
6.2 Effect of repeated sampling on detection of parasites and alleles by PCR and on multiplicity if infection 102
6.3 Detectability of parasite clones by PCR in different age groups 103

Chapter 1

Introduction

1.1 Malaria

1.1.1 Overview

Malaria is one of the major contributors to the global burden of disease [1]. Worldwide, there were an estimated number of roughly 200 million malaria cases in the year 2008, with a vast majority (85%) of those being in the African Region. This has lead to an estimated number of up to one million deaths, with a similar majority (89%) happening in the African region [2]. Sub-Saharan Africa in particular, where around 70% of clinical attacks occur [3], has such a high incidence of malaria because ideal climatic conditions for transmission coincide with the presence of efficient malaria vector mosquitoes [4].
Malaria is caused by protozoan parasites of the genus *Plasmodium*. There are five different parasite species of said genus which are known to infect humans, namely *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* and, as recently confirmed, *P. knowlesi* [5–7]. Of these, *P. falciparum* is responsible for the majority of severe disease and death [8].
Human malaria is exclusively transmitted by mosquitoes of the genus *Anopheles*. The female *Anopheles* require blood as protein source for egg production, and thereby create the opportunity for human-to-mosquito and mosquito-to-human transmission. Anophelines, albeit present worldwide, are most common in tropical and subtropical regions, and are only found at altitudes below 2500m. There are approximately 430 *Anopheles* species, of which around 70 are malaria vectors, but only 40 of these are thought to be of major public health importance [5]. In Sub-Saharan Africa the two major malaria vectors are *Anopheles funestus* and the members of the *Anopheles gambiae* complex [9].
A number of different strategies are available today in order to prevent or cure malaria infections. For treatment of malaria, a range of different drugs are available today, and many of these can be taken preventively. Despite these good news, the enormous number of infections acquired every year, the logistical and financial constraints developing countries are facing, and the rapid evolution of drug resistance when a drug is used on a large scale make the sheer existence of such drugs insufficient to tackle the problem properly. The distribution of insecticide-treated bednets (ITNs) has proven to be an effective and affordable tool of malaria prevention [11], as has the spraying of the inside walls of houses with residual insecticide (Indoor Residual Spraying, **IRS**) [12–14].

1.1.2 The history of malaria

Malaria was recognised as a disease in China almost 5000 years ago. It's characteristic symptoms have been described in many other parts of the world since then, such as in an-

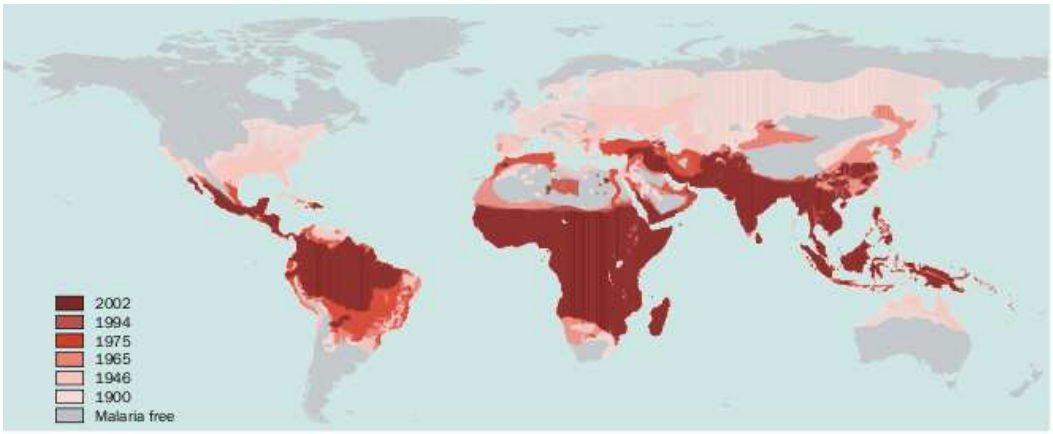


Figure 1.1: **The global distribution of malaria since preintervention (1900-2002)** - Illustration from [10].

cient Indian scriptures (3500 to 2000 years ago) as well as in Greek and Roman medical works [15], and the curative effects of quinine have been known for hundreds of years. Malaria parasites in human blood were first described in 1880 by Laveran, a french army surgeon in Algeria. However, it was not until 1897 that Ronald Ross discovered in India that malaria is transmitted by mosquitoes [5]. This discovery sparked a new era of research related to the control of malaria, which at that time still affected large areas of North America and Europe, including parts of Switzerland [15, 16]. With the disease vector now known, larvicides to prevent the breeding of mosquito larvae in water were developed. At the beginning of the second World War, the strong insecticidal action of Dichlordiphenyltrichlorethan (**DDT**) was discovered, and the subsequent spraying of insecticides against adult mosquitoes revolutionised malaria control. The effects of this can clearly be seen in Figure 1.1. Moreover, treatment options for malaria improved substantially after chloroquine was developed in 1934. The final eradication of malaria was envisioned in the 1950s, when DDT application showed very promising results in reducing the malaria burden [5]. Despite great successes in the Carribean, parts of Asia and South-Central America, and elimination in Europe and North America during the follwing decades, the efforts did not succeed in tropical Africa and many parts of Asia [15, 17]. This failure can partly be attributed to emerging drug and insecticide resistance, but also to the fact that in sub-Saharan Africa ideal climatic conditions for transmission coincide with the ranges of the most efficient vector mosquitoes in the world [18]. After that failure, malaria was a “neglected” tropical disease until recently, and malaria control was emphasized rather than eradication. It was almost considered a taboo to earnestly discuss local elimination or even eradication (the elimination of malaria from the globe). The Bill&Melinda Gates Foundation broke this taboo, by - quiet boldly - declaring malaria eradication as feasible and making it their primary long-term goal. Today, with unprecedented political will and financial support, malaria eradication is back on the global health agenda [19, 20].

1.1.3 The biology of *Plasmodium falciparum*

The malaria parasite life cycle [21], as shown in Figure 1.2, involves two hosts. During a blood meal, a malaria-infected female Anopheles mosquito inoculates sporozoites into the human host (1). Sporozoites infect liver cells (2) and mature into schizonts (3), which rupture and release merozoites (4). (Of note, in P. vivax and P. ovale a dormant stage [hypnozoites] can persist in the liver and cause relapses by invading the bloodstream weeks, or even years later.) After this initial replication in the liver (exo-erythrocytic schizogony A), the parasites undergo asexual multiplication in the erythrocytes (erythrocytic schizogony B). Merozoites infect red blood cells (5). The ring stage trophozoites mature into schizonts, which rupture, releasing merozoites (6). Some parasites differentiate into sexual erythrocytic stages (gametocytes) (7). Blood stage parasites are responsible for the clinical manifestations of the disease. The gametocytes, male (microgametocytes) and female

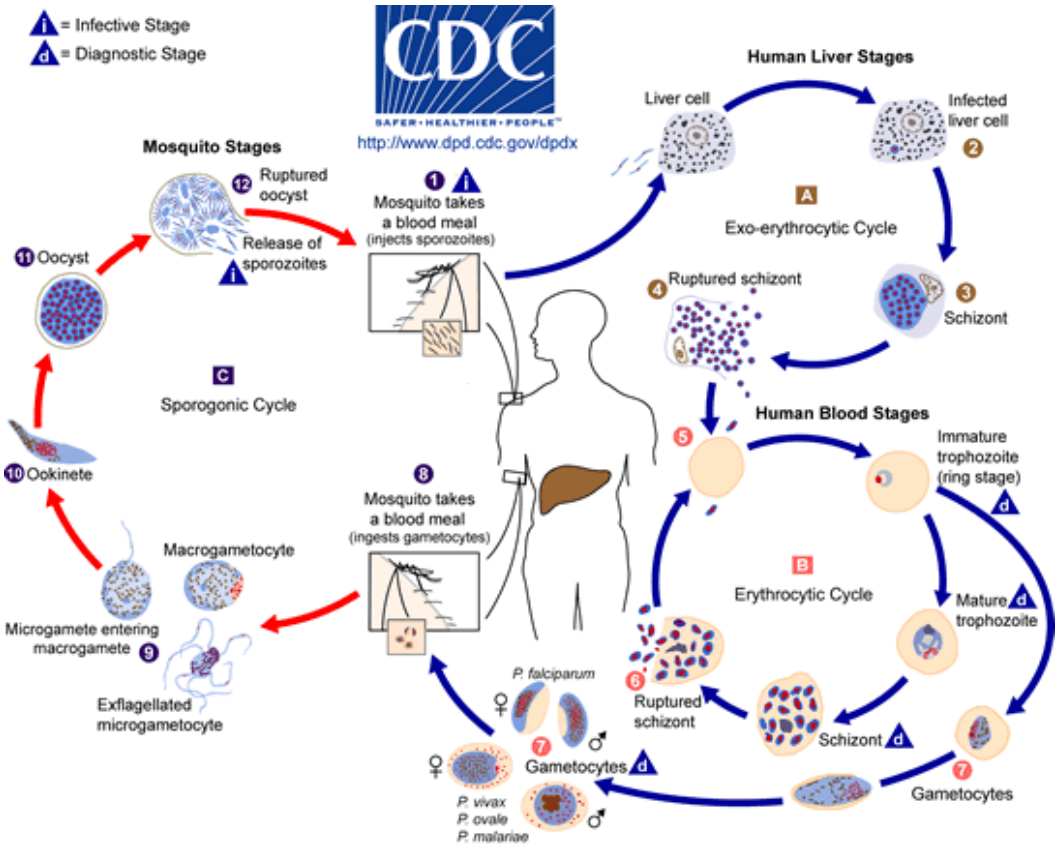


Figure 1.2: **The life cycle of *Plasmodium falciparum*** - Illustration from [21]

(macrogametocytes), are ingested by an *Anopheles* mosquito during a blood meal (8). The parasites multiplication in the mosquito is known as the sporogonic cycle C. While in the mosquito's stomach, the microgametes penetrate the macrogametes, generating zygotes (9). The zygotes in turn become motile and elongated (ookinetes) (10) which invade the midgut wall of the mosquito where they develop into oocysts (11). The oocysts grow, rup-

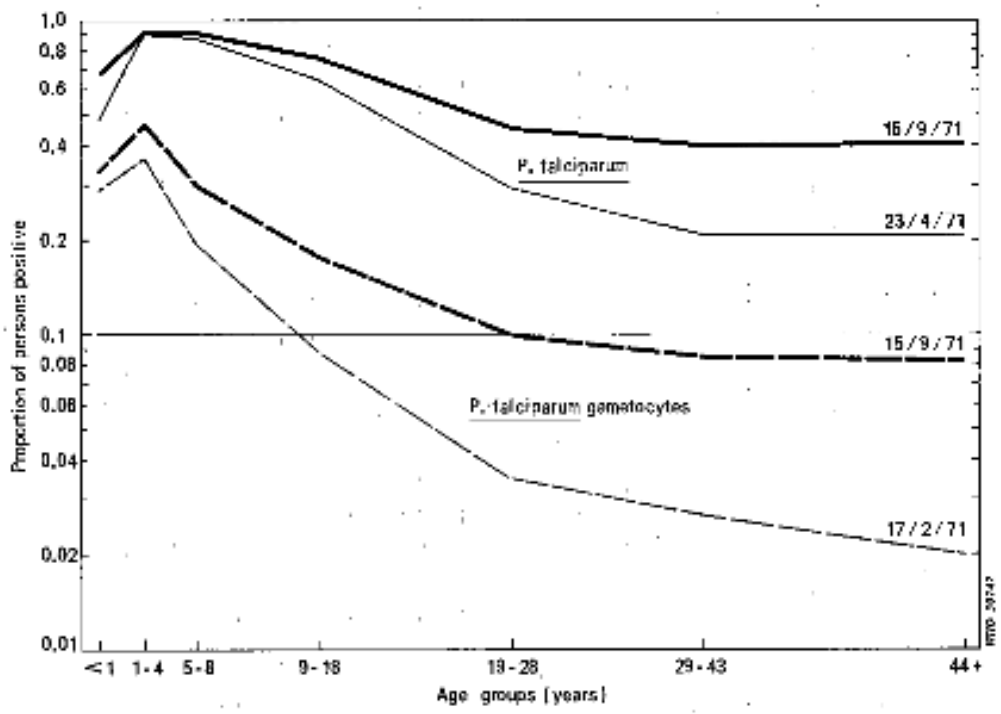


Figure 1.3: **Prevalence of *P. falciparum*** (trophozoites and/or gametocytes) and of gametocytes only, by age and season. Graph from [22]. It characterizes the malaria situation in Garki, Nigeria, in the year 1971.

ture, and release sporozoites (12), which make their way to the mosquito's salivary glands. Inoculation of the sporozoites (1) into a new human host perpetuates the malaria life cycle.

1.1.4 The epidemiology of *Plasmodium falciparum*

Measures of malaria transmission quantify malaria risk and endemicity levels and they are the basis of rational decision making in malaria control. These measures include parameters related to malaria transmission from mosquito to humans (i.e. entomological inoculation rates, force of infection, incidence rates, parasite prevalence) and parameters related to malaria vectors (i.e. mosquito survival, infection probability). On the human side, parasite prevalence is the most commonly used measure of malaria endemicity. Especially age-prevalence curves (Figure 1.3) provide insight into age-related aspects of disease prevalence. Malaria epidemiology is mainly dependent on the occurrence of efficient malaria vectors, climatic favourability for mosquito breeding as well as for parasite development, and presence of the human host. Given the presence of a mosquito population capable of transmitting the disease, transmission intensity heavily depends on the longevity of the adult anopheline vector. This is because the mosquito, infected after biting a human, has to survive the sporogonic development cycle of the *Plasmodia*, and after that survive another

few days in order to infect human hosts [23, 24]. Given the seasonal nature of environmental factors influencing mosquito emergence and survival, it is evident that also malaria transmission can be highly seasonal.

In the tropics, routine malaria diagnosis is carried out mainly by microscopic methods, but rapid diagnostic tests (**RDT**'s) are being evaluated as possible replacement. RDT testing results are positive when antigens of the parasite are present in a blood sample, and the test has therefore a higher sensitivity than microscopy [25]. However, it is also more expensive, and more difficult to store (cooling), and it's use may therefore not be practical in all circumstances. For epidemiological studies, there are also DNA-based diagnostic methods (see Section 1.4).

The clinical presentation of a malaria infection is with - sometimes periodic - fever, nausea and headache. These symptoms are rather general, and therefore there is a high danger of confusing malaria with other febrile diseases. Only diagnosis by microscopy or RDT's can securely identify a malaria infection as cause. Additional symptoms of severe infection are anaemia and acidosis. Severe cases may progress to losing consciousness, and death may occur. Particularly children are vulnerable to one of the most serious clinical complications, cerebral malaria [26]. The pathogenic mechanisms underlying cerebral malaria and why a small percentage of patients develop it are not fully understood, but the accumulation of large numbers of parasites in specific sites such as the brain or placenta appears to be important [27, 28]. In areas of intense transmission, new born children are relatively protected against malaria infection for the first three months due to passive immunity acquired from the mother through breastfeeding [29]. After that period, infants and children become highly susceptible to severe clinical manifestations of malaria and the overwhelming burden of morbidity and mortality falls upon this age group [26, 30–32]. If children survive past the age of five years, after being repeatedly inoculated with sporozoites and therefore exposed to pathogenic asexual blood-stages, they acquire a state of semi-immunity which protects them from the severest outcomes of malaria. This occurs primarily through the suppression of parasite densities without necessarily shortening the duration of infection [33–37]. For this reason, malaria prevalence in adults in highly endemic areas is often relatively low whereas the majority of young children are patently infected (Figure 1.3) [26, 38]. However, prevalence in semi-immune adults and older children is probably underestimated because low-density infections are harder to detect by microscopy [39].

1.2 Mathematical modeling of malaria

Mathematical models are used in many areas of scientific research as a tool for computing the consequences of one's assumptions. Devising such models is essentially just a more rigorous and formal way of "thinking" about a specific natural process of interest. By forcing the modeler to accurately state underlying assumptions in the form of mathematical equations, and by providing an exact framework which allows for deduction of the consequences of these assumptions, mathematical modeling helps mitigate the scientific "weaknesses" of the human mind, such as wishful thinking and the ability and willingness to overlook contradictions as well as the inability to grasp complex nonlinear interactions.

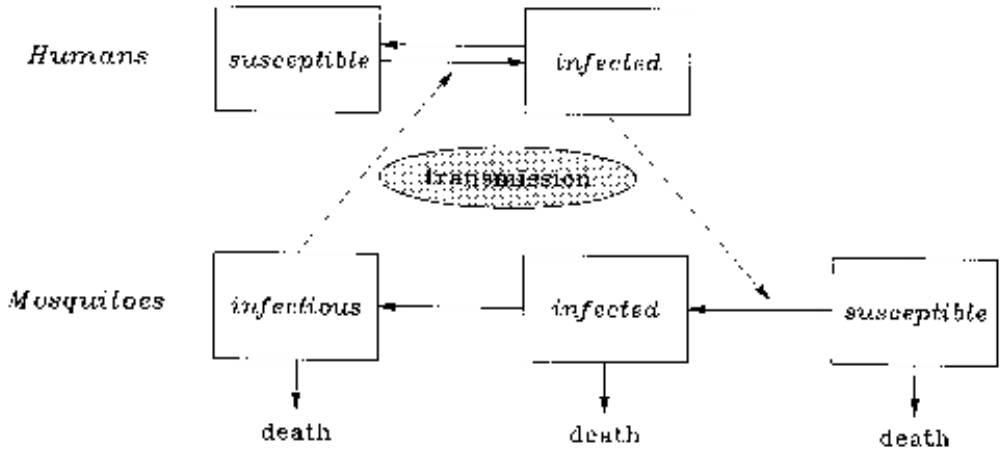


Figure 1.4: **The Ross-Macdonald model of malaria transmission** - The flow of humans from a susceptible class to an infected class and, through recovery from infection, the reverse, are shown in the upper part of the figure. The flow of mosquitoes from a susceptible class to an infected class, and finally to an infectious class are shown on the bottom. The human and mosquito populations are linked through the transmission process. Illustration from [42]

.

The importance of a quantitative description of malaria transmission in order to explain the observed differences in malaria patterns in different areas of the world was recognised very early. The first mathematical models of malaria were devised by Ronald Ross [40,41], who previously discovered that malaria is transmitted by mosquitoes.

George MacDonald later refined and extended Ross' modeling work, and mathematical modelling played a role in the planning of the malaria eradication campaign of the 1950s and 60s. A simple example of a transmission model, termed the Ross-MacDonald-Model, is given in Figure 1.4. The list of features of malaria which are not "accurately" represented by this model is long: the possibility that a host can harbour more than one infection (superinfection), the effects of immunity in the host, aspects of the natural history of the vector, the possible development of insecticide- and drug resistance, clinical manifestations of the disease, etc. Nevertheless its analysis has led to such fundamental insights as "...to counteract malaria anywhere we need not banish *Anopheles* there entirely...we need only to reduce their numbers below a certain figure.", or that the weakest link in the chain of malaria transmission was the survivorship of adult *Anopheles*. Given that MacDonald's models were published at about the same time the global eradication campaign based on DDT began, with DDT targeted at adult female *Anopheles*, it is not surprising that this conclusion was recruited to the cause [43]. The subsequent failure of the global eradication campaign is a reminder of the fact that conclusions from mathematical models must be interpreted with great care, or, in other words: malaria is easiest eradicated in models. Later modeling publications sought to identify and include other aspects of malaria transmission,

such as host immunity [22].
Today there exists a multitude of different approaches, modelling different aspects of the disease, such as whithin-host parasite dynamics, mosquito life cycle, migration of people, health systems, etc. The "Biostatistics and Computational Sciences" group at Swiss TPH is currently working on an integrated framework, where different model types can be integrated and tested with respect to their goodness of fit [44]. This platform is intended to help answer questions of interest in current malaria research and control, such as what the quantitative public health impact of mosquito net distribution would be, or under what conditions a vaccine would be effective for the control or elimination of malaria.

1.3 Infection dynamics in the human population

Modelling malaria transmission presents a number of challenges additional to those encountered when modelling diseases with completely infection-blocking immunity, like measles or rubella. An infection with *P. falciparum* does not protect from being re-infected, which inevitably leads to superinfection, i.e. some individuals harbouring several infections at a time. With microscopy alone, which is not able to distinguish between individual infection clones, it was impossible to tell how many infections a person is harbouring and to distinguish new from old infections. This made it difficult to validate existing mathematical models of infection dynamics in humans with field data.

1.3.1 The parameters of infection dynamics

Models of malaria infection dynamics in the host typically require three parameters to capture the processes of infection, clearance, and detection of an infection. Effects of immunity can then be expressed as changes in those parameters as a consequence of previous exposure. Since previous exposure is not readily accessible to measurement, host age is commonly used as a proxy, since an older age usually means that more infections were experienced. Other, "observable" epidemiological measures, such as prevalence, may subsequently be derived from the models, e.g. in order to allow for statistical fitting. Here we introduce the three parameters by highlighting the possibilities of measuring them using microscopy data. As examples of actual measurements of these parameters, graphs are shown which were reproduced after [37], using molecular data as information source. The workings of analysis methods using molecular data will be explained further below in section 1.4.1.

1.3.2 The force of infection

The acquisition of new infections is measured by the force of infection (**FOI**), denoted by the parameter λ. It is defined as the number of infections an individual acquires on average per unit of time. In many situations, λ is the measure of choice for measuring malaria transmission, e.g. in areas of very high endemicity, where the prevalence may not be informative because it approaches saturation. The force of infection is also an important determinant of the incidence of clinical disease, as it is believed that clinical attacks are

due to "new" infections and chronic infections tend to be asymptomatic. In clinical trials, knowledge of the force of infection assists in the design and in the estimation of required sample size and observation periods. When transmission rates are very high, children can be followed from birth until they become infected, in order to provide an estimate of λ as originally proposed by MacDonald [45]. This estimate is called the infant conversion rate. Another method recommended by [46] for determining λ is to use a drug to clear parasitaemia from a representative cohort of people. The subjects are then bled at regular intervals, perhaps once a week to once a month (depending on the level of malaria transmission) and blood films are collected for examination by microscopy. In this way, the force of infection can be calculated as the total number of (first) re-infection events divided by the total time at risk in the cohort. Alternatively, the presence of malaria-specific antibodies in blood serum (seropositivity) can be used to measure the force of infection. People of different ages are sampled in cross-sectional survey, their immune responses determined and sero-positivity is then a reliable measure of cumulative prevalence. [47] used indirect fluorescent antibody tests (**IFAT**) to estimate the force of infection for *P. falciparum* in this way, but were forced to omit children under one year of life from their surveys, because maternal antibodies would complicate the picture. This method cannot obviously not be used at high levels of transmission where almost the whole population is seropositive, but it is particularly useful in areas of very low transmission, where one is unlikely to find any ongoing infections.

1.3.3 The duration of infection

The duration of a *P. falciparum* infection is also the duration of potential infectivity of the host and is therefore of high importance in any transmission model. Not only does it affect the magnitude of transmission from humans to the mosquito population, but it also gains special significance in settings where malaria transmission is seasonal: the fraction of infections surviving a hypothetical dry season constitutes the founder population for the new transmission season. Any rational planning of a malaria control or elimination therefore profits from accurate measurements of infection duration. In analogy to the strategy of estimating the force of infection by following cohorts of uninfected people, until they become infected, the average duration of infection can be estimated by following naturally infected individuals and recording when the infection dissapears. In practise, however, this seemingly simple problem is marred by a number of difficulties: in naturally exposed populations, people tend to have multiple infections. This can be accounted for in the mathematical models used to analyse the data, yet because microscopy cannot distinguish the individual infections, untestable assumptions about the degree and nature of superinfection have to be made. In addition, imperfect detection is expected to bias the measurements, as “not detecting” an infection does not necessarily mean that it was cleared. A number of different approaches to measuring the infection duration from such data have been described, and the consensus appears to be a mean duration of approximately 200 days. For further information on measuring infection durations from microscopy data, the interested reader is referred to [48]. A second source of information on the duration of infection is malariatherapy data. These data were obtained from neurosyphilis patients, who were

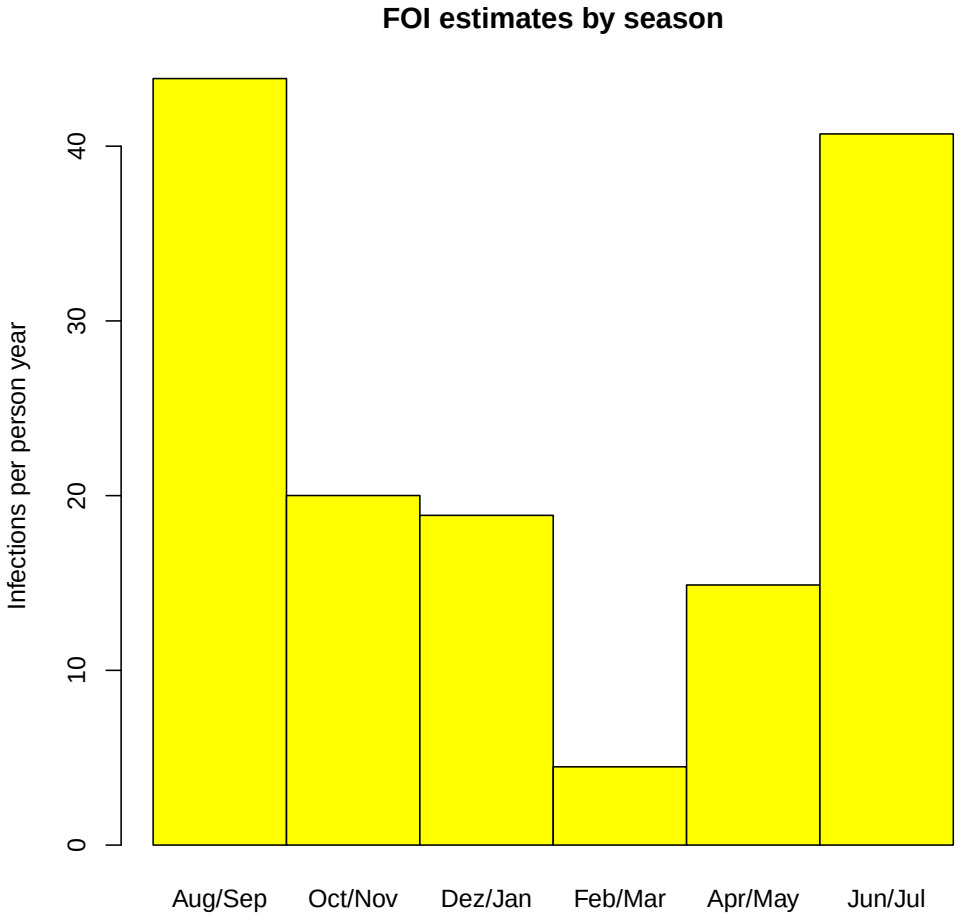


Figure 1.5: **Seasonality in the force of infection** - These FOI estimates from Navrongo, Northern Ghana, show both the pronounced seasonality of transmission in this area and the generally high number of infections a person acquires per year. Estimates were obtained by use of the method developed in chapter 4.

intentionally infected with *P. falciparum* strains. Infection with malaria was, before the discovery of modern antibiotics against syphilis, the only way of curing the disease. It is thought, that the recurrent fevers caused by malaria eventually cleared the syphilis. Malariatherapy data from the 1940's to the 1960's is up to this day an important source of information on the whithin-host dynamics of malaria. Sama *et al* [49] analysed these data and measured a mean infection duration of approximately 210 days. There are, however, several problems associated with malariatherapy data: firstly, it is unknown whether syphilis changes anything about the dynamics of malaria infections, and secondly, the data is obtained from individuals who never experienced malaria before, i.e. are immunologically naïve. It is therefore unclear how acquired immunity changes the duration of infection as measured from these data. Apart from that, many of the patients were given sub-curative treatments during the course of the infection, and it is also not known whether this affects the estimates. Often one uses a clearance rate μ as parameter to quantify the duration of infection: assuming an exponential distribution of infection duration, this corresponds to an average duration of $1/\mu$. The implications of making assumptions about the distribution of infection durations will be explained in more detail in section 1.5.

1.3.4 Detectability

Diagnostic methods in epidemiology have their limitations. It is therefore often of interest to compare different methods with respect to their sensitivity, i.e. their probability of correctly identifying an infected host. This is common practice in epidemiology, as one wants to find the "best" diagnostic method, yet it implicitly assumes that failure to detect an infection is solely due to properties of the diagnostic test being used. In addition, there may be properties of the host-parasite system, such as changes in parasite densities, which have an influence on the probability of detection. The term "detectability" is used, when the interest is not in comparing different diagnostic methods, but rather in understanding the factors of the host-parasite system influencing detection. Because a discussion of detectability is only meaningful for a given diagnostic method at a time, we denote it with "Q" when using microscopy, and with "q" when using molecular methods. The difference is, that Q is the probability of detecting at least one infection in a potentially superinfected individual, and q denotes the probability of detecting a single clonal infection.
Failure to account for imperfect detection biases several standard epidemiological measures such as prevalence and multiplicity of infection (when using molecular methods), and also affects estimates of the force and duration of infection, as mentioned above. In a single cross-sectional survey it is not possible obtain information about the numerical value of detectability because non-detection due to absence of a parasite is not distinguishable from a failure of detection. Longitudinal data, on the other hand, contains information on the value of detecability, provided there is some knowledge about duration and force of infection: e.g, a negative sample between two positive samples can be regarded as detection failure, if clearance and infection rates are known to be sufficiently well known in order to exclude clearance and subsequent re-infection during the time interval. Several related methods have been described in order to estimate the numerical value of detectability from longitudinal data. For microscopy data, where individual parasite clones cannot be distin-

guished, the methods of [50] and [51] are applicable. Both assume infections are neither acquired nor cleared during the study.

1.4 Molecular data

Molecular diagnostic methods make use of the desoxyribonucleic acid (**DNA**) of highly variable regions of the parasite genome to distinguish different infection clones within one individual.In *P. falciparum*, established methdods are, e.g., the use of restriction fragment length polymorphism (**RFLP**) genotyping, where DNA is amplified using polymerase chain reaction (**PCR**, a method to produce numerous copies of a DNA template) and subsequently digested using sequence-specific bacterial restriction enzymes. This yields DNA-fragments with a characteristic length-distribution [52] for each parasite clone. A different approach uses genes which contain repetitive sequences of variable length within the parasite population. Two genes of this class are the "Merozoite Surface Proteins", msp1 and msp2. PCR Amplification of a region delimited by primers up- and downstream of the variable region then yields DNA fragments of different length, which can be classified using gel or capillary electrophoresis. Most datasets analysed in this thesis were obtained using the msp2 marker gene. Msp2 is located on the merozoite surface and appears to be essential for the parasite, yet it's exact function is not clear [53]. Despite it's essentiality, no homologues are found in other human malaria species, and the closest homologue is found in *P. reichenowi*, a close relative of *P. falciparum* infecting chimpanzees [54].

1.4.1 The analysis of molecular data

Molecular diagnostic methods have the advantage that individual infections whithin one host can be distinguished. This is a considerable improvement over microscopy as far as epidemiological research is concerned, because it has the potential to yield a much more detailed picture of infection dynamics in the human host. However, microscopy still retains a value for diagnosis as a cheap and easily maintainable diagnostic method in potentially resource-constrained settings. Moreover, microscopy allows for measurement of actual parasite densities, while genotyping data rather indicates presence or absence of a particular strain. Molecular diagnostic methods are able to detect very small amounts of DNA in a sample and are therefore much more sensitive than microscopy. Yet, at least two diagnostic problems remain. Firstly, it is possible that two parasite clones share the same genotype at the locus which is used for diagnosis. As far as analysis of longitudinal molecular data is concerned, there is to date no reliable statistical method to correct for this, and one generally has to assume that re-infection with the same genotype is a rare event and does not occur. This assumption will also be made throughout this thesis, but most analyses here use data from northern Ghana, where genetic diversity in the parasite population is high. There are, however, methods to assess whether it is safe to make said assumption [55–57]. The second problem which complicates the analysis of molecular data is that sometimes the peripheral blood of infected persons does not contain any parasites at all - an effect which is summarised in the detectability parameter q, as explained in section 1.3.4. The biological reasons for the complete absence of parasites in the peripheral blood are com-

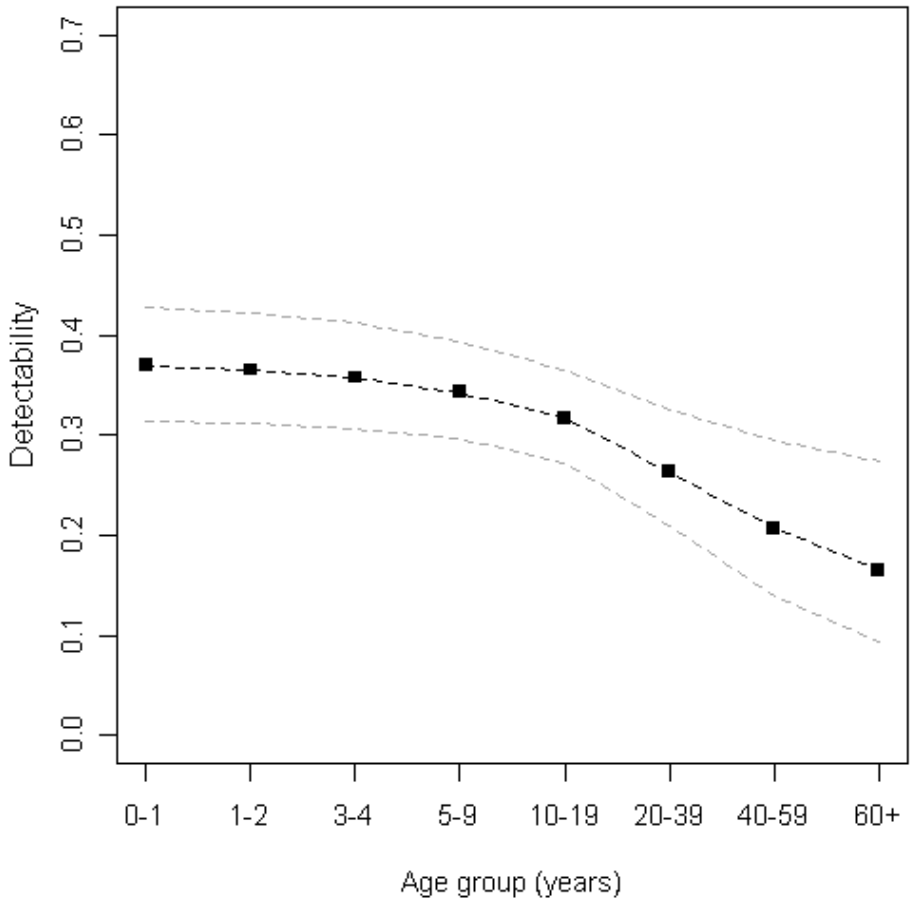


Figure 1.6: **Age dependence of detectability** - These measurements from Navrongo, Northern Ghana, suggest a decrease of detectability with the age of the host. This is interpreted as an effect of acquired immunity, which acts to reduce parasite densities after repeated exposure. Estimates and confidence intervals were obtained by use of the method developed in chapter 4.

monly attributed to attachment of the late intra-erythrocytic stages of *P. falciparum* to the endothelium of the blood vessels, and a synchronization of the parasite population with respect to stages of development. This would then have the effect that, at certain times, a majority of the parasite population is attached to the walls of the blood vessels. However, this is not very well understood, and chapter 2 will further discuss this phenomenon. A typical longitudinal dataset, produced using molecular diagnostics, will yield a sequence of presence and absence for each genotype in every human. For convenience, one uses a binary notation, i.e. "1010" would then mean, that the particular genotype under consideration was detected on the first survey, not detected on the second survey, and so on. Every individual in the dataset would "contain" zero or more such binary sequences, depending on the degree of superinfection. A very straightforward method of analysing such data would be to just assume perfect detection, and count the number of appearances and dissapearances of a genotype. The number of appearances per time unit would then be the force of infection, and the time until a genotype dissapears would be an estimate for the duration of infection. Since some of the times the infection was present, but not detected, this obviously leads to very bad estimates of both of these parameters: the force of infection should be greatly overestimated, and the duration underestimated. A slightly better method is assuming that re-infection with the same genetic marker is very rare event and does not happen at all - an assumption which is justified if the marker diversity in the parasite population is high enough. Using this method one would obtain a better estimate of λ as the number of "new" appearances of a genotype, and an estimate of the average duration $1/\mu$ would be the average time between the first and last detections of infections in the data. This method would also yield a rough estimate of the detectability q, namely in form of the proportion of times an infection was not detected between first and last "sighting". These methods are not very good, because especially at low values of q infections are acquired much before they are first and terminate much after they are last seen, or they may not be detected at all. Nevertheless, such considerations highlight the fact that a central challenge in the analysis of molecular data stems from the fact that detection is imperfect. Clearly, a statistical model in the form of an expression for the likelihood of a particular dataset as a function of parameters λ, μ and q is needed. One of the early developments in this direction is proposed by [58], who use a multinomial likelihood model for the frequency distribution of short binary sequences, termed triplets. A later, generalized variant of the method applied the markov chain formalism in order to calculate the likelihood of the complete dataset, rather than only looking at certain triplets. The basic idea of those methods was subsequently worked out using a Poisson likelihood model (instead of a multinomial likelihood model) by [37, 57]. Their method allows for estimation of all three parameters of infection dynamics simultaneously from longitudinal typing data. Through modelling seasonality by using a different parameter λ_i for every season i in the dataset (Figure 1.5), and by allowing the detectability parameter to decrease as a function of age of the host (Figure 1.6), the model fit could be substantially improved. The duration of infection, $1/\mu$, showed no dependence on the age of the host, and remained constant at approximately 140 days [37]. This is an interesting finding, because it questions the hypothesis that immunity shortens the duration of a *P. falciparum* infection. Whether or not this conclusion will be confirmed by future analyses, it is clear that the method of [57] offers the possibility to

learn more about the infection dynamics of malaria by comparing different statistical models with respect to their goodness of fit. Molecular typing data holds the key to a wealth of information on malaria infection dynamics, which can be harvested by advancing suitable methods of analysis.

1.5 Objectives of the thesis

The work presented in this thesis aims to extend the existing methods of analysis of molecular data. The focus shall be on increasing the knowledge about *P. falciparum* within-host dynamics, as far as it is observable by methods similar to the one of [57]. The plan is to successively drop or change different key assumptions and compare the different models with respect to their goodness of fit to data.
Chapter 2 investigates the short term within-host dynamics using statistical methods. The motivation for this comes from the need to model imperfect detection as a Bernoulli process. This essentially implies that detections at two points in time are independent of each other. As there are reports of complicated dynamics of parasite densities, the assumption of a constant probability of detection must be investigated.
Chapter 3 reports for the first time the complete molecular dataset from a one-year longitudinal study in Navrongo, Northern Ghana. Analyses are performed using the method developed in chapter 4, among others.
Chapter 4 develops an extension of the method of [57] where a range of parametric survival distributions can be used to model infection durations. This is of interest because the current models only permit an exponential distribution of infection durations, which is biologically very unrealistic.
Chapter 5 uses the method developed in chapter 4 in order to explore the effects of acquired immunity on infection dynamics, especially on clearance of infections.
Chapter 6 analyses data on *P. falciparum* and *P. vivax* from a study in Papua New Guinea (**PNG**). Estimates of detectability obtained using a formula introduced in chapter 2 are compared between species.

1.5.1 The distribution of infection durations

The main assumption in [57] that this thesis aims to overcome, is the assumption of an exponential distribution of infection durations. This is a very common assumption in many areas of mathematical modelling, and it is implied by having a constant clearance rate μ. If the clearance rate is constant, the durations of all malaria infections are then distributed according to an exponential distribution, with mean duration $1/\mu$. One of the reasons for this is a purely mathematical one: in an exponential distribution, the age structure of the infection population can be neglected because the probability that an infection is cleared within a certain time interval is entirely independent of the current age of the infection. This assumption closely reflects reality within the context of e.g. radioactive decay (the time until a particle decays), or chemical reactions (the time a molecule exists before reacting on). Yet, with respect to the distribution of biological infection durations, an exponential model is arguably not very realistic.

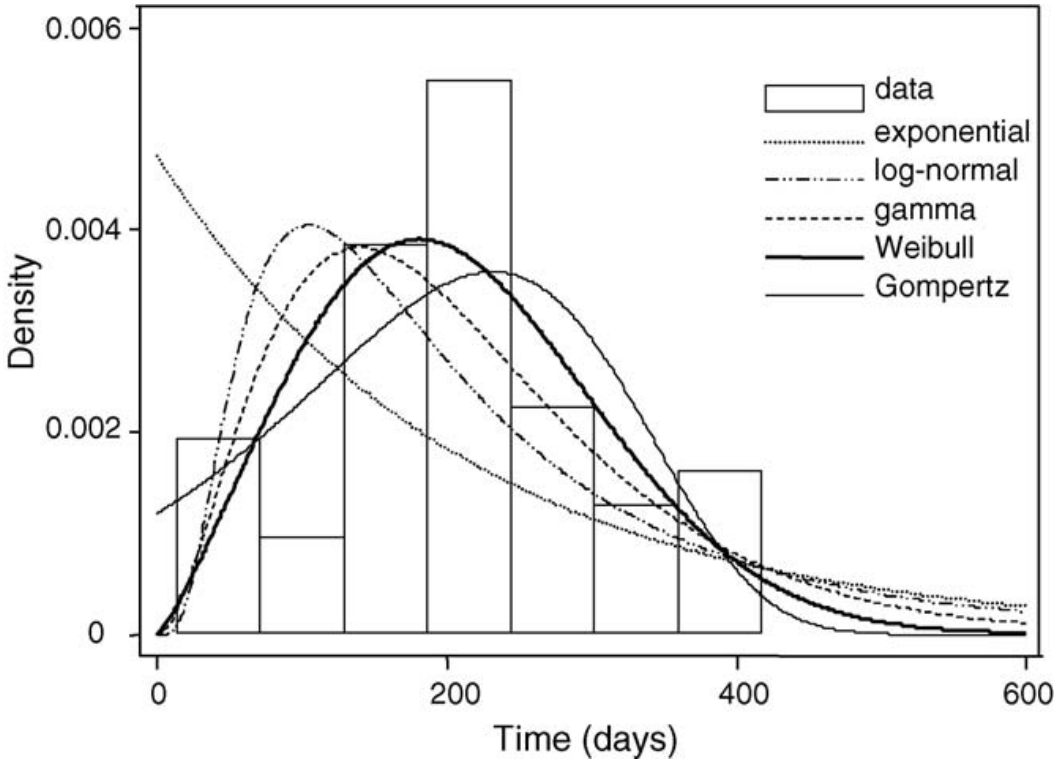


Figure 1.7: **The distribution of infection durations** - Sama *et al* [49] fitted different survival distributions to the durations measured in malariatherapy data. The data is comprised of patients which were deliberately infected with *P. falciparum* in order to cure an existing neurosyphilis - a common treatment for syphilis before the advent of suitable antibiotics. Since the start and, with less precision, the end points of all infections are known in this type of data, standard survival analysis methodology can be used for estimation of the distributional parameters. Yet, it is not known how acquired immunity would change this picture, since inference of similar information from naturally exposed individuals is complicated by imperfect detection.

To date, the only information on how durations of *P. falciparum* infections are distributed, comes from malariatherapy data [49]. The analyis of such data is facilitated by the fact that the dates of infection are exactly known, and patients were only infected with one strain at a time. The duration of every single infection is therefore approximately known, and the methods of statistical survival analysis can be applied. The authors have compared a number of parametric survival distributions and concluded that the Gompertz and Weibull distributions gave the best fit to the data, followed by the gamma, lognormal, and exponential distributions (Figure 1.7). However, their analysis was solely based on infections in immunologically naïve patients, and it is not clear what the picture would look like in people who have experienced high malaria transmission throughout their lives.
The main target of the work presented in this thesis is to fit survival distributions to genotyping data from exposed individuals, by extending the method of [57], and to investigate the effects of acquired immunity on infection durations.

CHAPTER 2

Detectability of *Plasmodium falciparum* clones

Authors: Michael T Bretscher, Francesca Valsangiacomo, Seth Owusu-Agyei, Melissa A Penny, Ingrid Felger, Tom Smith

2.1 Abstract

Background: In areas of high transmission people often harbour multiple clones of *Plasmodium falciparum*, but even PCR-based diagnostic methods can only detect a fraction (the detectability, q) of all clones present in a host. Accurate measurements of detectability are desirable since it affects estimates of multiplicity of infection, prevalence, and frequency of breakthrough infections in clinical drug trials. Detectability can be estimated by typing repeated samples from the same host but it has been unclear what should be the time interval between the samples and how the data should be analysed.
Methods: A longitudinal molecular study was conducted in the Kassena-Nankana district in northern Ghana. From each of the 80 participants, four finger prick samples were collected over a period of 8 days, and tested for presence of different Merozoite Surface Protein (msp) 2 genotypes. Implications for estimating q were derived from these data by comparing the fit of statistical models of serial dependence and over-dispersion.
Results: The distribution of the frequencies of detection for msp2 genotypes was close to binomial if the time span between consecutive blood samples was at least 7 days. For shorter intervals the probabilities of detection were positively correlated, i.e. the shorter the interval between two blood collections, the more likely the diagnostic results matched for a particular genotype. Estimates of q were rather insensitive to the statistical model fitted.
Conclusions: A simple algorithm based on analysing blood samples collected 7 days apart is justified for generating robust estimates of detectability. The finding of positive correlation of detection probabilities for short time intervals argues against imperfect detection being directly linked to the 48-hour periodicity of *P. falciparum*. The results suggest that the detectability of a given parasite clone changes over time, at an unknown rate, but fast enough to regard blood samples taken one week apart as statistically independent.

2.2 Background

In areas of high endemicity of *Plasmodium falciparum*, human hosts are often superinfected with multiple clones of the parasite [58]. Identification of these concurrent infections is important for understanding patterns of drug resistance [59] and of the transmission of the parasite. PCR-based methods for detecting parasites not only have lower detection limits than blood smear microscopy, but also make it possible to distinguish genetically distinct clones, and hence to compute multiplicity of infection. But at least two diagnostic problems remain: i) the same host might be infected with more than one parasite clone of the same genotype, which can introduce bias into estimates of multiplicity of infection [55]. ii) PCR detection can be negative because the sample taken does not contain any parasites. This may happen due to effects of acquired immunity or synchronization of the parasite population. Failure to account for imperfect detection biases several standard epidemiological measures, such as prevalence and multiplicity of infection. Most critically, analysis of drug failure rates using molecular typing may overlook breakthrough parasite clones or conversely misclassify them as new infections after treatment.
Repeated blood samples from the same host can be analysed to estimate the probability that a clone is detected in any given sample (the detectability, q). For microscopy data, where individual parasite clones cannot be distinguished, the statistical methods of [50, 51] are applicable. Both assume infections are neither acquired nor cleared during the study. For molecular data, several pieces of work aiming at estimating infection duration and force of infection also yielded measurements of detectability and its dependence on age of the host [37, 57, 58, 60]. These methods make use of data collected over longer time periods (several months up to a year), with surveys every 1 to 2 months, the kind of which may not be easily available in practice. Moreover, the obtained estimates of detectability depend on simultaneous estimates of infection- and recovery rates as well as on assumptions concerning these processes. A simple method is therefore presented, to estimate the detectability of infecting clones from molecular data with short inter-survey intervals. It makes use of pairs of surveys sufficiently close in time, such that reinfection with the same parasite genotype can be safely excluded. The method is similar to the one presented in [50], but adapted for the context of molecular diagnostic methods. This implies that the maximal number of "infections" is not limited by the number of hosts in the study, but rather represents individual parasite clones. The methods of [37, 57, 58, 60] as well as the one presented here assume that the detections of an infecting clone at different time points are independent from each other. While it seems reasonable to make such an assumption, provided intervals between surveys are long enough, it is not clear how long these intervals need to be. Numerous publications report complicated periodic behaviour of fevers or parasitaemia [61], or detection events [62], which creates a need to establish the circumstances under which the methods mentioned above can be applied.

In order to evaluate the effect of possible "non-random" behaviour of clonal infections on estimates of detectability, a longitudinal study comprising 80 individuals was conducted in northern Ghana. From each participant, four blood samples were collected over a period of 8 days. Using these data, various statistical models are compared with respect to their

goodness of fit, and a series of hypothesis tests is performed. The resulting statistical description of the within-host dynamics of *P. falciparum* clones, as observed by molecular typing methods, allows us to justify a simple algorithm for obtaining reasonably robust estimates of q and specify the circumstances under which this method is applicable.

2.3 Methods

2.3.1 Study site and sample collection

The present survey was conducted following a one year longitudinal study on malaria epidemiology [37,52,57,63] in the Kassena-Nankana district (KND), in the Upper East Region of Ghana. The malariological situation in this area is characterized by very high prevalence and multiplicity of infection [63, 64], and year-round transmission with seasonal variation in transmission intensity [37]. From the participants of the mentioned main study, 80 individuals below 20 years of age were randomly selected for this follow-up. From these, a total of four blood samples were taken on the last survey of the main study as well as 1, 6 and 7 days later (Figure 2.1). The present analysis was restricted to these four samples within eight days. Study participants were visited in the early mornings of each day and houses were visited in the approximately the same order, to ensure sample collection at roughly the same time of day for each individual. Whole blood was collected on "ISOCode^{TM} Stix" PCR template preparation dipsticks (Schleicher & Schuell, Dassel, Germany).
Study participants who were sick at the time of the survey were referred to the routine health services. No anti-malarial treatments were administered by the research team.

2.3.2 Genotyping

DNA was eluted from "ISOCode^{TM} Stix" filter paper and screened for presence of *P. falciparum* by polymerase chain reaction (PCR). Sample processing and PCR conditions have been described in detail [65]. In brief, all samples were subjected to PCR using primers specific for the merozoite surface protein (msp) 2 locus. Genotypes were distinguished on

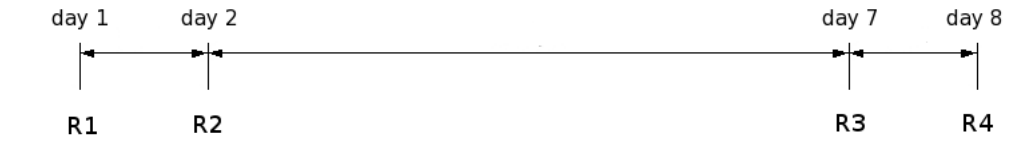


Figure 2.1: **Study design** - Blood samples were collected in four survey rounds (R1-R4), on day 1, 2, 7 and 8. The result of this study design are two sampling intervals of 1 day, one of 5, two of 6, and one of 7 days. A 48-hour periodicity of *P. falciparum* detectability could therefore be identified, as it should show positive correlation of detection outcomes between surveys with even-numbered interval length, and negative correlation between surveys with odd-numbered interval length (in days).

R1	R2	R3	R4	sequence no.	count
0	0	0	0	0	-
0	0	0	1	1	43
0	0	1	0	2	42
0	0	1	1	3	26
0	1	0	0	4	54
0	1	0	1	5	10
0	1	1	0	6	13
0	1	1	1	7	19
1	0	0	0	8	28
1	0	0	1	9	21
1	0	1	0	10	41
1	0	1	1	11	34
1	1	0	0	12	64
1	1	0	1	13	22
1	1	1	0	14	41
1	1	1	1	15	61

Table 2.1: **Data coding** - Failure or success to detect a clone at any given survey round was coded using binary notation. This yielded 519 sequences of length four. Sequences were numbered according to the binary value they encode. Sequence no. 0 is invisible.

the basis of length polymorphism and PCR fragments were precisely sized by automated capillary electrophoresis and GeneMapper® software. An in-house generated software identified all genotypes per sample and transformed the data into different formats suitable for data management and statistical analysis. Given the high number of msp2 genotypes in the population, re-infection with the same genotype was assumed to be a rare event. As a consequence of this, for any given host, msp2 genotype is assumed to be synonymous with "infecting clone" in all analyses.

2.3.3 Data analysis

Only data of those participants who were present at all four survey rounds, and where at least one genotype was found, were included in the analysis. This reduced the number of individuals in the data set to 69. Patterns of appearance and disappearance of specific parasite genotypes depend on rates of infection and clearance as well as on detectability. However, for the purpose of the present analysis, acquisition and loss of infections were neglected. It was assumed that there are no false positive results and that an infecting clone is present throughout all four surveys if detected at least once. This is justified by the comparatively short time interval between the first and the last survey, and by previously published estimates of infection- and clearance rates from the dataset of the main study [37]:
According to the authors, a person experienced an estimated 0.6 new infections during the time of the study (31 new infections per annum in the corresponding season). This implies that around $0.6 * 69 \approx 41$ or approximately 8% of the 519 clones in the data set may have

been acquired during the study. Similarly, assuming an average (clonal) infection duration of 150 days and that infections were acquired at random times relative to the time of the study leads to an estimate of $7/150 \approx 5\%$ of clones being cleared during the study period of seven days.
Failure or success to detect a strain was denoted by 0 or 1, respectively, yielding 519 binary sequences of length four. The 15 possible sequence types containing at least one positive test result are referred to by the binary number they encode (Table 2.1). The resulting pool of sequences was either analysed as a whole, or split into the following age-groups (age in years): 0-2, 3-5, 6-10, 11-15, 16-20. This mode of analysis implies that clones infecting the same host are assumed independent of each other. Further, the present analysis is only concerned with variation in detectability among clones, not among hosts.
A series of χ^2 tests and Spearman's rank correlation analysis yielded qualitative information on the temporal behavior of detectability. Further, a series of models for the dynamics of detectability were fit to the data using Bayesian MCMC. These are described in detail below. The models and their estimates of detectability are compared using Deviance Information Criterion (DIC) as measure of goodness of fit [66]. That only sequences with at least one positive result were included in the data, and therefore the data are biased, was accounted for in all analyses. The software Winbugs [67] was used for all Bayesian model fitting, whereas for all other analyses the software package R was used [68].

2.3.4 Models of detection

In order to explore the short term dynamics of detectability, three statistical models are compared with regard to their goodness of fit (M1 to M3 below). These models are in the form of an expression for the detectability of clone i at time point t. This allows for fitting of the models by Bayesian Markov Chain Monte Carlo (MCMC), assuming individual detections are Bernoulli-distributed as

$$X_{i,t} \sim Bern(q_{i,t}^{obs}).$$

In addition, a simple method of directly measuring detectability from pairs of surveys (M0) is used. Applying this method to all available survey pairs in the data set and comparing the estimates of q with the model results allows us to develop criteria for the circumstances under which the method may be used.

2.3.4.1 M0: Direct estimation of detectability

Following [50], a method is proposed for direct estimation of the detectability q from pairs of observations. The estimate is a function of the number of infecting clones that were detected in only one of two survey rounds (n_1), and the number which was detected in both (n_2). Assuming a binomial distribution of the number of times a clone is detected and correcting for the detection bias leads to the following expression for the estimated detectability $\hat{q}$ so given by

$$\hat{q} = \frac{2n_2}{n_1 + 2n_2}. \tag{2.1}$$

The binomial likelihood model implies statistical independence of detections at different time points. This assumption might be violated if detectability exhibits temporally structured behavior. The detectability model underlying this method is identical to M1 below, but this method only uses two observations, and model fitting in order to estimate q is done analytically.
This simple method is compared to models M1 to M3 (below) in order to justify it's use, and to establish the conditions where it can be applied. For a formal derivation of (2.1) and confidence intervals for $\hat{q}$ please refer to the appendix (additional file 1).

2.3.4.2 M1: Binomial model

Model 1 follows M0 in assuming that the detectability $q_{i,t}$ is a constant for all clones i and time points t, namely

$$q_{i,t} = \bar{q}.$$

This implies independence of detecting a clone at time t from whether it was detected at other time points, and homogeneity of the infection population with respect to detectability.

2.3.4.3 M2: Beta-binomial model

Model 2 allows for variation in detectability among clones, but requires every clone to have the same detectability throughout the study. Variation in detectability is modeled using a beta distribution:

$$q_{i,t} = q_i \sim Beta(a, b),$$

where a and b are the shape parameters of the beta distribution.

2.3.4.4 M3: First order Markov Chain

Model 3 uses a two-state, first order Markov chain to represent the time evolution of detectability. In a first order Markov chain, the probability of detecting a clone at time t depends on whether it was detected at time $t-1$. This is achieved by defining $q_{i,t}$ as one of two detectabilities q_1 or q_0, depending on whether the clone i was detected at time $t-1$, or not.

$$q_{i,t} = \begin{cases} q_1 & \text{, if clone detected at } t-1 \\ q_0 & \text{, otherwise} \end{cases}$$

This is equivalent to a two-state Markov chain defined by the transition matrix

$$T = \begin{pmatrix} t_{00} & t_{01} \\ t_{10} & t_{11} \end{pmatrix} = \begin{pmatrix} 1 - q_0 & q_0 \\ 1 - q_1 & q_1 \end{pmatrix},$$

where $t_{i,j}$ is the probability that a transition from state i to state j occurs in the data, when the two observations are one time unit apart (24h in this case).
If the two observations are n days apart, the the transition matrix is raised to the power of n and becomes

$$T_n = T^n.$$

The probability of detecting a clone at the first survey, $q_{i,1}$, was assumed to be equal to the expected detectability $\bar{q}$, which follows from the stationary distribution of the Markov chain defined by T. In equilibrium, the the number of transitions from 0 to 1 and from 1 to 0, respectively, must be equal. Therefore $\bar{q}t_{10} = (1 - \bar{q})t_{01}$, which leads to the expression for $\bar{q}$ as given by

$$\bar{q} = \frac{t_{01}}{t_{01} + t_{10}} = \frac{q_0}{1 + q_0 - q_1}.$$

An important feature of this simple model is that it not only represents a random walk in "detection space" (i.e. switching between being detected and not being detected), but that it can also be interpreted as a random walk in "detectability space" (switching between the two detectabilites). This can be illustrated as follows: The probability that a clone changes its internal state from q_0 to q_1 is equal to the probability that it is detected while being in state q_0, which is equal to q_0. Likewise, the probability of a transition from q_1 to q_0 is $1 - q_1$. The resulting transition matrix for such a process is identical to T. However, since this model is fitted to a population of sequences, as opposed to just fitting it to one single time-series, one has to be careful in interpreting a possible best fit of this model: simple heterogeneity in detectability among clonal infections would also result in different estimates for q_1 and q_0, even if there is no random walk in detectability within a single clone. Therefore, only in combination with the results of M2, which is also able to capture such heterogeneity, is one able to interpret t_{10} and t_{01} as transition probabilities between q_0 and q_1.

2.3.5 Bias correction of detectability estimates

An "observed" detectability estimated by fitting these models does not correspond to the underlying "true" detectability because clones only appear in the data if detected at least once. Therefore, a bias correction is required in order to estimate the true detectability q^{true}. This was achieved as follows: by considering only the (time independent) mean detectability $q^{obs} = E\left[q_{i,t}^{obs}\right]$, the corresponding mean true detectability q^{true} is approximated by q^{obs} times the probability that a clone is included in the data, so given by

$$q^{true} \approx q^{obs}(1 - (1 - q^{true})^4).$$

This expression can be solved numerically by starting with the approximation

$$q^{true} \approx q^{obs}(1 - (1 - q^{obs})^4),$$

and iteratively approaching q^{true}. The magnitude of the bias in detectability estimates can thus be examined numerically. It amounts to approximately 10^{-2} for the values of q in the present data. The approach was used in all models to correct the measured detectabilities for detection bias. Similarly, the true number of clones present, N_{true}, is approximated as

$$N_{true} \approx \frac{N_{obs}}{(1 - (1 - q^{true})^4)},$$

with $N_{obs} = 519$.

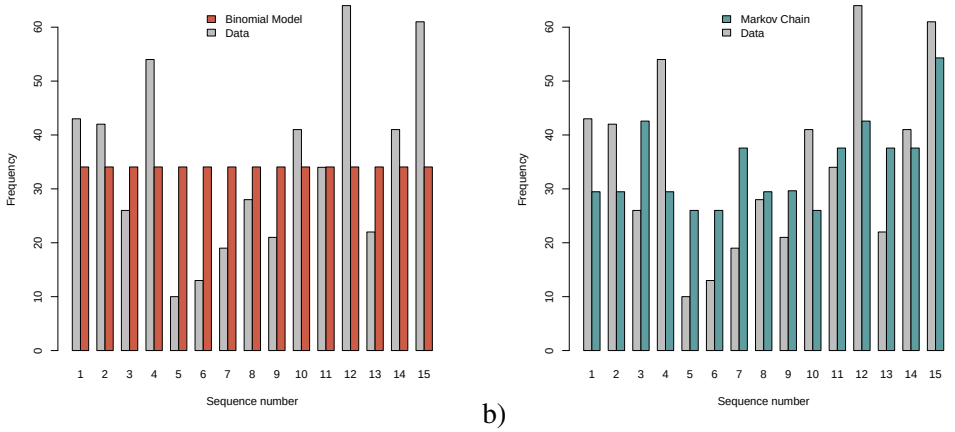
2.4 Results

In the complete study population (80 individuals), the average prevalence across all four survey rounds was 46 % by microscopy and 69 % by PCR. The dataset used for statistical analyses comprised 69 parasite-positive individuals between 6 months and 20 years of age, with a median age of 5.2 years (inter-quartile range 3.5-9.7). The median multiplicity of infection (MOI) among these was 10 (inter-quartile range 7-13), when pooling all four observations from each individual. This differs from standard practice when reporting MOI, but was justified given the very short interval between the surveys. The obtained value is expected to be a better estimate of the true MOI. When only considering single survey rounds, the median MOI among PCR-positives was 4. The 519 detected clones belonged to 77 different msp2 genotypes, with the most common allele reaching a frequency of 9.2 %.

2.4.1 Tests of proportion and correlation

A series of hypothesis tests was performed in order to gain insight into the statistical properties of the data-generating process. These do not relate to the models M1-M3 directly, but rather aim to look at similar questions using a completely different methodology. Any conclusions would need to be consistent with both approaches. In this analysis, the detection bias is accounted for by adding a total of 26 all-zero sequences to the data set, such that the total number of sequences equals 545. This is the "true" number of clonal infections, as estimated robustly by models M1 to M3. In the following list, $H_1 - H_5$ indicate the hypotheses tested, and the corresponding p-values obtained using χ^2 tests (with the exception of the Spearman's Rank Correlation analysis) are given:

H_1: All 4 surveys have an equal proportion of positive results, i.e. $\sum d_{i,1} = \sum d_{i,2} = \sum d_{i,3} = \sum d_{i,4}$ (Data: 312, 284, 277, 236). This hypothesis of stationarity is rejected by a χ^2 test with 3 degrees of freedom: P-value <0.0001.



a) b)

Figure 2.2: **Expected and actual frequencies of sequence types** - Comparison of sequence type frequencies in the data with their expectations from a) the binomial model (M1), and b) the Markov Chain model (M3). M3 fitted the data better, yet did not fully explain it. The beta-binomial model (M2) is not shown since it measured almost no variation in detectability among clones, and therefore effectively reduced to M1.

H_2: The frequencies s_i of the 16 binary sequences (including the added all-zero sequence) are multinomially distributed with expectations $s_i = 545q^{o_i}(1-q)^{4-o_i}$, where o_i is the number of positive testing results in sequence i). This hypothesis, effectively proposing that a Bernoulli-process is able to perfectly describe the data, is rejected by a χ^2 test with 14 degrees of freedom: P-Value <0.0001.

H_3: The number of sequences with $i = 0, 1, .., 4$ detections are multinomially distributed with expectations $s_i = 545\binom{4}{o_i}q^{o_i}(1-q)^{4-o_i}$. This is a slightly relaxed version of H_2, such that the time order of detections is neglected, and sequences with a certain number of detections are pooled. However, this hypothesis is rejected by a χ^2 test with 3 degrees of freedom: P-Value <0.0001.

H_4: The frequencies of all four possible results of a survey pair (i.e. "00","01","10" and "11") are multinomially distributed with expectations $s_i = 545q^{o_i}(1-q)^{2-o_i}$. This is a special case of H_2, only applied to a pair of surveys, instead of the whole dataset. Except for survey pair 2-3 (p-value 0.06) this hypothesis is rejected on all survey pairs by χ^2 tests with 2 degrees of freedom: P-values <0.0001.

H_5: The distribution of the number of successful detections in pairs of surveys is binomially distributed. This is very similar to H_3, except that only pairs of surveys are considered. The duration between the observations turns out to be important, as the hypothesis is rejected by χ^2 tests with 1 degree of freedom on all survey pairs (p-value <0.0001), except on the longest interval with a duration of 7 days (p-value: 0.83). The p-values of all pairs

are listed in Table 2.3. This result is particularly interesting as it could be interpreted as test of an ergodic hypothesis, which implies that after enough time has passed, the system "forgets" where it started and its state at the second observation is independent from the first observation.

A Spearman's Rank Correlation analysis showed significant positive correlation between all pairs of surveys which are 24h apart, and no correlation for all other pairs, with the exception of the pair formed of surveys 1 and 3.

2.4.2 Model comparison

A comparison of models M1 to M3 with respect to their goodness of fit - as indicated by lower values of DIC - reveals that M3 fits the data best (Table 2.2). This was the case both when splitting the data by age group (DIC 2860.0) and without doing so (DIC 2856.4). The model estimating a separate parameter set for each age group indicates a decreasing trend in detectability with age, as observed by others [37, 58, 69]. However, the model with only one parameter set for all ages has a lower value of DIC, and therefore this trend is not significant. There is almost no difference in goodness of fit between the binomial (M1) and beta binomial models (M2), as indicated by the corresponding DIC values (all between 2870.2 and 2872.3). This is in line with the finding that the beta-binomial models estimated almost no variation in q and therefore effectively reduced to the corresponding binomial model. A graphical comparison of M1 and M3 is presented in Figure 2.2.

model	age groups	$\bar{q}$			DIC
M1 (binomial)	1	0.50			2870.2
M1 (binomial)	5	0.55, 0.52, 0.49, 0.53, 0.41			2872.3
model	age groups	$\bar{q}$	var(q)		DIC
M2 (beta-bin.)	1	0.51	0.003		2871.5
M2 (beta-bin.)	5	0.51, 0.46, 0.46, 0.47, 0.46	0.004,0.005,0.005,0.005,0.005		2871.3
model	age groups	$\bar{q}$	q_0	q_1	DIC
M3 (Markov)	1	0.50	0.47	0.59	**2856.4**
M3 (Markov)	5	0.55, 0.53,0.49, 0.53, 0.41	0.45, 0.44, 0.47, 0.56, 0.46	0.65, 0.64, 0.56, 0.55, 0.45	2860.0

Table 2.2: **Comparison of Models M1-M3** - Model results: M3 without age groups fitted the data best, indicated by it's lowest value of DIC. M2 effectively reduced to M1, as it estimated very low variance of detectability. All models estimated a true number of clones of approx. 546 (not shown), and similar values for the mean detectability $\bar{q}$. Values of $\bar{q}$ represent the bias-corrected mean detectability, whereas q_0 and q_1 are not bias corrected, but represent "observed" detectabilities (see section "bias correction").

2.4.3 Estimates of q

The estimates of detectability (Table 2.2) were found to be similar for models M1 through M3, especially when common parameters for all age groups were estimated. All values were approximately 0.5, and showed little prior sensitivity. Although measuring separate detectabilities for every age group did not improve model fit, a decreasing trend of detectability with age was observed. Estimates of q for the youngest group are between 0.51

survey pair	interval (days)	n_1	n_2	$\hat{q}$	95% CI	p-value
1-2	1	220	188	0.63	0.59-0.68	9.4e-12
3-4	1	233	140	0.55	0.49-0.60	9.1e-7
2-3	4	293	134	0.48	0.43-0.53	1.1e-2
1-3	5	235	177	0.60	0.55-0.65	2.4e-7
2-4	5	276	112	0.45	0.39-0.50	1.7e-3
1-4	**7**	**272**	**138**	**0.50**	**0.45-0.55**	**0.83**

Table 2.3: **Direct estimation of q on all survey pairs, using M0.** -Direct estimation of q on all pairs of surveys gave heterogeneous results. However, only in pair 1-4 are the proportions of single and double positives compatible with the binomial assumption of the direct estimation method (p-value:0.83). The estimate of q from this pair is close to the estimate of M3.

and 0.55, and decrease to values between 0.41 an 0.46 for the oldest age group. This is consistent with the findings of other authors [37]. Estimates of q obtained using M0, however, show some variation, with values ranging from 0.45 to 0.63. Table 2.3 shows the corresponding estimates obtained from all available pairs of surveys. Only for the measurement using survey pair 1-4 were the criteria for using the method fulfilled, as the corresponding p-value of 0.83 indicates, and the value of q estimated from this pair matches the estimates from the models very well.

2.5 Discussion

The short term dynamics of asymptomatic *P. falciparum* clonal infections *in vivo* were characterized in order to find a simple way of measuring detectability in the field. A series of statistical tests as well as a progression through three simple models provided insight into some statistical properties of within-host dynamics monitored by molecular typing. Classical PCR ignores absolute parasite densities, but length polymorphic amplicons make it possible to distinguish between co-infecting parasite clones. Detectability, however, can be used as a proxy for parasite densities, as the two must be correlated. Since key epidemiological measures such as prevalence and multiplicity of infection (MOI) depend on the numerical value of detectability, planning and monitoring of malaria interventions rely on accurate measurements of detectability. It is important to note that detectability may not only depend on host or parasite factors, but also on the methods for generation of genotyping data. These include the method for collecting blood samples, the actual volumes of blood collected, storage conditions, PCR conditions, and competition within the PCR assay limiting the detection of minority clones. The present analysis did not consider any of these factors, but rather assumed that their impact is more or less identical for all samples and clones.

2.5.1 Within-host dynamics

Measurement of detectability from longitudinal data will for practical reasons rely on binomial models of detection. This creates a need to establish under what conditions such binomial models are applicable. Different hypotheses tested on the typing data showed complicated dynamics of clonal infections for short timescales. These dynamics could not be described by a binomial model (rejection of hypotheses H_1 to H_3). A hypothesis which could not always be rejected was H_5. This hypothesis stated that the number of successful detections in sample pairs were binomially distributed. It was not rejected for the sample pair collected at the most distant dates, i.e. from surveys 1 and 4 with interval of 7 days (Table 2.3). This finding could indicate that the processes governing detectability on short time scales are prone to stochastic variation such that the effect of the initial state of a clone vanishes after some time, and the two observations become independent. In other words, the corresponding test could be interpreted as test of an ergodic hypothesis, which implies that the system under investigation "forgets" it's initial state after enough time has passed. That the frequencies of "01" and "10" sequences are not equal, and therefore H_4 (stricter than H_5) is rejected on all survey pairs except pair 2-3, questions this interpretation, and can not be explained it in a satisfactory way. Since there are consistently more "10" pairs than '01", one could presume that the detectability of clones simply decreases with time, which would also explain the decreasing trend in the number of detections per survey (H_1). However, this is mere speculation and can hardly be shown from this dataset. Nevertheless, the method of directly measuring detectability is presumably little affected by this phenomenon, as it's estimates of q only depend on the sum of single positive pairs and the obtained numerical values of q agree very well with the results of the other models.

Intuitively, one would expect a certain amount of variation in detectability among clonal infections, especially since these were pooled across individuals. It is therefore surprising that M2, which would allow for such variation, measured zero variance of q and effectively reduced to the binomial model M1. The best fitting model M3 offers a possible interpretation, as it is capable of capturing change in detectability over time. M3 models the time evolution of detectability as a Markov chain, which is equivalent to assume that a clone has detectability q_0 if it was not detected on the preceding survey, and detectability q_1 if it was. The obtained estimates of q_0 and q_1 as 0.47 and 0.59, respectively, could either indicate variation in the dataset with respect to detectability or that the detectability of a clone performs a random walk in detectability space, alternating between the two states q_0 and q_1. Since M2 reduced to M1, and estimated practically no variation in detectability, we have to assume the latter.
One might expect parasite densities to fluctuate with a period of approximately 48 hours, as observed in malariatherapy-data [61], and in good agreement with in-vitro measurements of a 48 hour erythrocytic cycle. In fact, such periodic behaviour of asymptomatic infections has been reported [62, 70]. The present analysis does not find a 48-hour periodicity, rather the opposite: both the best fitting model as well as the results of the Spearman's rank correlation analysis indicate positive autocorrelation between time points which are 24 hours apart. A process with a periodicity of 48 hours, on the contrary, should show nega-

tive correlation. A possible explanation for the difference between malariatherapy data and the data presented here could be that malariatherapy patients were not immune and therefore had fever more often. The question of periodicity in symptomatic malaria should be considered separately, and it's causes are thought to be well explained [71, 72]: high temperature (fever) differentially affects the intra-erythrocytic stages of parasite development, and nearly stops development in some of these. This leads to "queuing" of the parasite population, and when the fever goes down, all parasites continue their development in a synchronized way. Fever can by definition not be operating in asymptomatic individuals, but at least in simian and avian malaria an effect of normal diurnal changes in body temperature on synchronization has been demonstrated, alongside with the observation that sometimes the parasite population is split into "two broods [..], coming to schizogony on alternate days" [73]. Two broods, synchronized within themselves, appearing in the peripheral blood with a 48 hour periodicity, yet with a 24 hour phase-shift, would appear in the data as having a 24 hour periodicity. This would be consistent with the finding that detection results one day apart are positively correlated. As the data does not contain smaller time intervals, however, any such periodicity cannot be distinguished from a simple gradual change in detectability.

Analysis of periodicity of clonal infections would ideally make use of long series of parasitological observations of untreated infections with short intervals, but few studies have collected such data, partly for ethical reasons. Exceptions include the malariatherapy datasets [35], the studies of Farnert *et al* [62, 74] and Magesa *et al* [75] in Tanzania, and Bruce *et al* [70, 76] from Papua New Guinea. Bruce *et al* aggregated data for paired observations with identical interval length and calculated the probability of detecting an infection at the second occasion, conditional on it being detected at the first occasion. This analysis suggests values of detectability similar to the estimates in the present study, with a six day periodicity. This periodicity was interpreted as signal of a 48-hour underlying cycle because the sampling interval was three days, which meant that six-day and two-day periodicity could not be distinguished. Similar analyses of the other available datasets would be of value.

2.5.2 Measurement of detectability

A comparison of different approaches for estimating detectability found remarkably good agreement of the obtained numerical values. Of practical interest is the use of a direct method of estimating the detectability q from pairs of surveys by using the number of clones which were detected once (n_1), or twice (n_2):

$$q \approx \frac{2n_2}{n_1 + 2n_2}.$$

This approach was found to give very similar results as the more sophisticated methods, provided the underlying assumption is met: the number of successful detections must follow a binomial distribution. The statistical properties of the data, as assessed by a series of tests, suggest that if there is an interval of at least 7 days between consecutive surveys, it is safe to make these assumptions (see H_5). Alternative methods [37, 57, 58, 60] rely on

the same two assumptions, yet are further incorporating models for the processes of acquisition and loss of infections. While those may themselves be of interest, the associated measurements of q may be affected by assumptions about acquisition and loss of infections. Direct estimation using M0 is therefore recommended as a simple and practical alternative, if only detectability is of interest, and if the interval between two surveys is short enough so acquisition and loss of infection clones can safely be excluded.

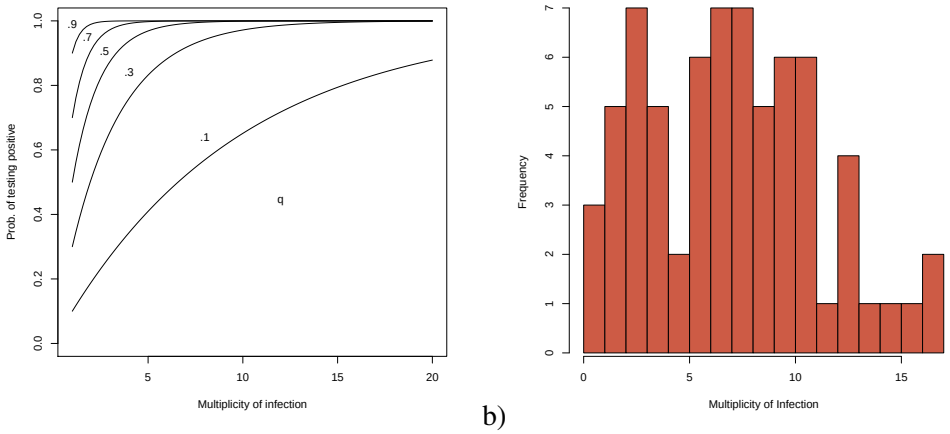
2.5.3 Epidemiological significance of detectability

Prevalence and multiplicity of infection (MOI) are key epidemiological parameters, which characterize the malariological situation in a given area, and are routinely being reported. Quantities like these are ultimately important for rational planning of interventions. Both mentioned quantities are, however, affected by the value of detectability, which in comparison receives little attention. It seems plausible, that on average the "true" MOI should be the "observed" MOI divided by the detectability, which implies - given values of q around 0.5 - that true MOI's are roughly double of what is being reported. However, this ignores, that detectability itself might depend on MOI, and is merely an approximation. What about estimates of prevalence? It seems plausible that the extent to which measurements of prevalence are influenced by the value of q should vary with the multiplicity of infection, as the probability to miss every single one of n clones in a host (and obtain a false negative result) could be stated as $(1-q)^n$ (Figure 2.3).

This implies that the measurement error for prevalence, when neglecting detectability, should be highest at the lowest multiplicities of infection - a situation to be expected when approaching local elimination. It is therefore desirable to routinely report q together with other epidemiological measures, if possible.
In drug efficacy trials, the phenomenon of imperfect detection complicates the task of distinguishing new from breakthrough infections, and therefore must have an influence on drug efficacy estimates. In addition, residual drug levels may keep parasite densities at undetectable levels for some time, which is usually taken into account when designing drug efficacy trials. No satisfactory statistical methodology for analysis of such trials appears to exist, taking into account both imperfect detection and residual drug levels. It is suspected that many recrudescent infections, i.e. infection clones which survive treatment and are detected several days or weeks later, might in fact often be detected earlier if multiple testing took place. This is strongly supported by the findings of [77], who note that consecutive-day blood sampling changes the results of a drug efficacy trial compared to single-day blood sampling.

2.6 Conclusions

The presented work demonstrates the importance of paying attention to the phenomenon of imperfect detection not only in the sense of assessing the sensitivity of diagnostic tests, but also looking at it as a property of infections or individuals. Various epidemiological measures, such as prevalence or MOI, are affected by imperfect detection. Failure to account



a) b)

Figure 2.3: **The error in prevalence measurements becomes more important at low MOI** - a) Prevalence estimates are biased due to imperfect detection. Assuming that infecting clones within a particular host are independent from each other, the probability of missing all of them and therefore falsely classify an individual as negative, is highest for low multiplicity of infection. This graph shows - for different values of q - how the number of clonal infections in a host affects the estimates of prevalence. The probability of correctly recognizing a positive individual with n infections is calculated as $1 - (1 - q)^n$. It follows that the effect of detectability on prevalence estimates is highest at low multiplicity of infection and therefore low transmission, for example when being close to local elimination. However, low transmission intensity might prevent acquisition of immunity and therefore raise the value of detectability. It is therefore desirable to report estimates of q and multiplicity of infection together with prevalence estimates. b) The distribution of MOI. Contrary to common practice, observations from all four surveys are pooled for the calculation of MOI. This corresponds to the assumption that clones are present throughout all surveys if detected once. With the help of subfigure a, the bias on prevalence estimates in this population, as introduced by imperfect detection, can be roughly estimated.

for it may severely distort the outcome of measurements and may even lead to wrong conclusions. As an example, the decrease in prevalence with age, as it is frequently observed in malaria, might actually mean that detectability decreases with age, while prevalence remains constant or even increases. Indeed, some publications suggest that this might be the case [37]. In addition, it is likely that underestimation of prevalence may be substantial in situations where the multiplicity of infection is low. This is because the chance of missing every single clone in a host is highest when there are only few.

A simple method of estimating detectability from molecular data, using pairs of surveys, was presented. It is a modification of existing methods, which can deal with data on multiple infections within one host. The numerical estimates of detectability obtained using said formula appeared remarkably robust. Through comparison of the detectability estimates with estimates from different models, and through a series of statistical tests, the conditions under which the underlying assumptions of the method are fulfilled could be established. It's use is recommended when the time interval between the two surveys is one week or more, but discouraged on data with shorter time intervals, if possible. Both the method itself as well as the way of addressing it's applicability are not restricted to malaria, but may in a similar way be used for other infectious diseases where molecular data on individual clones is available.

The restrictions on applicability stem from the complicated dynamics of detectability on short time scales. These were investigated and it was found that treating individual detections as statistically independent is only an acceptable approximation for time intervals longer than one week. Contrary to expectation, however, no changes in detectability indicative of a 48 hour cycle were found, as is reported from malariatherapy data. This suggests that not the 48 hour erythrocytic cycle of *P. falciparum* is dominating detectability in vivo, but that other factors, such as e.g. the dynamics of the immune system, may be important. As the participants of the study must be considered partly immune, it is presumed that the within-host dynamics of infections differ between immune and non-immune individuals. This questions the use of malariatherapy data for fitting of within-host models for the immune host, and encourages further collection of relevant data as well as development of analysis methods in order to gain better insight into the within-host dynamics of *P. falciparum* in immune individuals.

2.7 Appendix

A formula to estimate the detectability q from pairs of surveys has to take into account the fact, that double-negative pairs are never detected, and that therefore the data is biased. Such a formula was given by [50] for the diagnosis of onchocerciasis by microscopy. Here, the corresponding formula is derived for the context of genotyping data, where a human host can harbor multiple infections. The obtained algebraic expression in the case of two survey rounds will turn out to be identical to the one given in [50], despite the slightly

different assumptions of the two approaches [1].

Assuming a binomial distribution of the number of successful detections, the probability that an infection produces a pair of observations with k positive results, p_k is

$$p_k = \frac{2}{k!(2-k)!} q^k (1-q)^{2-k}$$

A heuristic way of arriving at an estimator $\hat{q}$ of q is as follows: it is assumed that the actual data equal their expectations, i.e. $n_k = p_k n_{tot}$, of which n_1 and n_2 are known. Algebraically dividing n_2 by n_1,

$$\frac{n_2}{n_1} = \frac{q^2 n_{tot}}{2q(1-q)n_{tot}} = \frac{q}{2-2q},$$

and solving for q yields equation (2.1) as the desired result [2]:

$$\hat{q} = \frac{2n_2}{n_1 + 2n_2}.$$

2.7.1 Maximum likelihood estimation of q

A formal derivation of equation (2.1) makes use of a multinomial likelihood model and uses the probabilities of getting k successes conditional on the probability that an infection appears in the data[3], i.e. $\frac{p_k}{1-p_0}$. The likelihood of having n_1 single and n_2 double detections in $n_1 + n_2$ trials can then be written as follows:

$$\begin{aligned} L(q) &= \underbrace{\frac{(n_1+n_2)!}{n_1!n_2!}}_{=konst.} \left(\frac{p_1}{1-p_0}\right)^{n_1} \left(\frac{p_2}{1-p_0}\right)^{n_2} \\ &\propto \left(\frac{2q(1-q)}{1-(1-q)^2}\right)^{n_1} \left(\frac{q^2}{1-(1-q)^2}\right)^{n_2} \\ &\propto \left(\frac{2-2q}{2-q}\right)^{n_1} \left(\frac{q}{2-q}\right)^{n_2}. \end{aligned}$$

Omitting constant factors and taking the logarithm yields the log-likelihood function:

$$l(q) = n_1 \log(2-2q) - n_1 \log(2-q) + n_2 \log q - n_2 \log(2-q).$$

[1] In the method of [50], n_k signifies the number of individuals testing positive k times in a study, and the prevalence is unknown. For the context of genotyping data, n_k denotes the total number of infections found in the study population which were detected k times, and the total number of infections in the study population is unknown.

[2] Through division by q the solution $q = 0$ is lost. It is not plausible for physical reasons, as one would then not observe any data.

[3] A note on conditional probabilities: The probability that an event A occurs, given that an independent event B has already occurred, equals the probability that both events occur divided by the probability that B occurs, namely $P(A|B) = P(A,B)/P(B)$. This may at first not be obvious, but follows through simple rearrangement of the more familiar expression $P(A|B)P(B) = P(A,B)$.

We obtain the score function $S(q)$ by taking the derivative of $l(q)$ with respect to q:

$$S(q) = \frac{d}{dq} l(q) = \frac{2n_2 - q(n_1 + 2n_2)}{q(1-q)(2-q)}$$

The maximum likelihood estimator $\hat{q}$ of q can then be determined by finding the root of the score function

$$\frac{2n_2 - \hat{q}(n_1 + 2n_2)}{\hat{q}(1-\hat{q})(2-\hat{q})} = 0.$$

This expression can only be zero, if the numerator is zero, and therefore it simplifies to

$$2n_2 - \hat{q}(n_1 + 2n_2) = 0.$$

Solving for $\hat{q}$ leads to equation (2.1)

$$\hat{q} = \frac{2n_2}{n_1 + 2n_2},$$

which confirms that it is a maximum likelihood estimator of q.

Confidence interval

Construction of a confidence interval requires the Fisher information $I(q)$, which is the negative derivative of the score function, namely

$$I(q) = -\frac{d}{dq} S(q) = -\frac{d}{dq}\left(\frac{2n_2 - qn_1 - 2qn_2}{q(1-q)(2-q)}\right),$$

which leads to

$$I(q) = -\frac{n_1 + n_2}{(q-2)^2} + \frac{n_1}{(q-1)^2} + \frac{n_2}{q^2}. \tag{2.2}$$

The observed fisher information I_{obs} is $I(q)$ evaluated at $q = \hat{q}$, so

$$I_{obs} = -\frac{n_1 + n_2}{(\hat{q}-2)^2} + \frac{n_1}{(\hat{q}-1)^2} + \frac{n_2}{\hat{q}^2},$$

which simplifies to

$$I_{obs} = \frac{(n_1 + 2n_2)^4}{4n_1 n_2 (n_1 + n_2)}$$

This allows us to calculate the standard error of $\hat{q}$ as

$$SE(\hat{q}) = \frac{1}{\sqrt{I_{obs}}} = \frac{2\sqrt{n_1 n_2 (n_1 + n_2)}}{(n_1 + 2n_2)^2}.$$

A confidence 95% confidence interval for $\hat{q}$ can then be constructed using Wald's approximation:

$$[\hat{q} \pm 1.96 S\,E(\hat{q})]$$

CHAPTER 3

The dynamics of natural *Plasmodium falciparum* infections

Authors: Ingrid Felger, Martin Maire, Michael T. Bretscher, Nicole Falk, André Tiaden, Wilson Sama, Hans-Peter Beck, Seth Owusu-Agyei, Thomas Smith

3.1 Abstract

Background: The force of *Plasmodium falciparum* infection and the duration of untreated infections are key variables in understanding the temporal dynamics of effects of interventions. Estimation of these parameters from the field requires molecular typing techniques for longitudinal tracking of co-infecting clones. Statistical methods need to allow for imperfect detection of infections. Our previous ability to analyze age-dependency in average duration of infection was restricted by limited availability of genotyping data. Recent improvements in mathematical modeling and in the precision and throughput of our genotyping system enable us to overcome this limitation.

Methods and Findings: In a longitudinal study conducted in a highly exposed population in Northern Ghana over a period of one year, up to six consecutive blood samples were analyzed from each of 349 individuals of an age-stratified cohort. An immigration-death model was fitted to detection time-series of merozoite surface protein 2 (**msp2**) genotypes. The *P. falciparum* clone acquisition rate was highest about two months after the onset of the rainy season. The model gave estimates of the duration of clonal infections against a background of high multiplicity of infection (**MOI**) and imperfect detection. Detectability was 38% in children and decreased with age, reaching 17% in adults >60 yrs. The estimated duration of infection peaked in 5-9 year old children, in whom on average an infection lasted 319 days (CI 318-320). In adults duration decreased to about 131 days (CI 114-148). We corrected MOI for the effects of imperfect detection. The modeled MOI was considerably higher than the observed number of clones, with the greatest effect in the older age groups where a mean of only 2.8 clones per parasite positive carrier was observed, compared with an estimated average of 17 distinct clones present.

Conclusions: Persistence and chronicity of individual parasites is the hallmark of malaria infections. *P. falciparum* infections persisted in all age groups for longer than 100 days and duration of infection was highest in 5-9 year old children. The semi-immune status achieved with adolescence seems to contribute to a somewhat increased elimination rate as well as to a reduction of parasite load. The true age pattern of MOI is profoundly affected by lower detectability in adults as a consequence of the decreasing parasite load leading to higher number of sub-patent infections.

3.2 Introduction

The rates at which new infections of the malaria parasite *Plasmodium falciparum* are acquired and existing infections are lost may be affected by the host's acquired immunity, and thus by the age of the host. Although many mathematical models of malaria transmission assume that acquired immunity reduces infection duration there are few experimental data available on clearance and acquisition rates of natural infections observed over extended periods of time.

The duration of a clonal infection is difficult to determine in field samples from endemic areas due to the abundance of superinfection. A background of ongoing infections precludes the identification of new infections by microscopy, thus the duration of a natural infection can only be determined using molecular typing. Sequestration of late stage parasites by cytoadherence further obscures the study of infection dynamics because parasites disappear from the peripheral blood circulation at periodic intervals, so that when parasites are well synchronised entire parasite clones can escape detection. Failure to detect a sub-patent parasite clone can also depend on its density and the detection limit, even when using nested PCR (**nPCR**). Longitudinal studies with short term sampling (daily or weekly) in asymptomatic individuals have found that individual parasite clones identified by restriction fragment length polymorphisms (**RFLP**) frequently appear to be lost only to reappear again [62, 78]. As a consequence, temporary absence from the peripheral blood has to be distinguished from parasite clearance.

We have developed several statistical methods to estimate infection duration allowing for this imperfect detectability of individual clones [37, 57, 58, 60]. These models assume that the failure to detect a particular genotype at a date between two positive samples need not reflect loss, but may also result from presence below the detection limit. However, in our previous analyses uncertainties were caused by subjectivity in the visual analysis of RFLP gels and the small number of samples that could be processed using this laborious genotyping approach. In addition, we were not able to fit separate models to different age groups, and as a consequence, did not consider non-monotonic patterns of age dependence.

To reduce these uncertainties we recently established and validated a highly accurate genotyping technique of the *P. falciparum* merozoite surface protein 2 (**msp2**) gene, that is based on precise sizing of PCR products by capillary electrophoresis. This approach facilitates longitudinal tracking of multiple infections using an automatic readout [52]. We have now expanded our previous pilot analyses of parasite typing and dynamics in Navrongo, northern Ghana [37, 52, 57, 63], to include data from all 349 individuals of all ages who were followed up over one year in two monthly surveys. We now present age-specific estimates of infection duration and detectability based on these data.

3.3 Materials and Methods

3.3.1 Field methodology

The study was carried out in Kassena-Nankana District (**KND**) in northern Ghana where malaria transmission is intense with strong seasonality [79]. For a molecular epidemio-

No. individuals enrolled	347
No. of samples collected	1902
No. of individuals with a complete set of 6 samples	269
No. of individuals with >1 parasite-positive blood samples	216
Female to male ratio (%)	53/47
P. falciparum positivity by microscopy (proportion)	
<1yr	0.549
1-2 yrs	0.838
3-4 yrs	0.803
5-9 yrs	0.826
10-19 yr	0.633
20-39 yrs	0.448
40-59 yrs	0.371
>60 yrs	0.363
overall	0.515
P. falciparum positivity by PCR, all ages (proportion)	0.898
Geometric mean parasite density by age group (95% CL)	
<1yr	770.6 (500.1; 1187.6)
1-2 yrs	1053.6 (766.7; 1447.9)
3-4 yrs	1064.5 (805.3; 1407.2)
5-9 yrs	420.5 (344.5; 513.2)
10-19 yr	291.0 (233.4; 362.9)
20-39 yrs	111.4 (84.5; 146.9)
40-59 yrs	119.8 (84.2; 170.4)
>60 yrs	116.8 (89.9; 151.8)

Table 3.1: **Characteristics of the cohort studied**

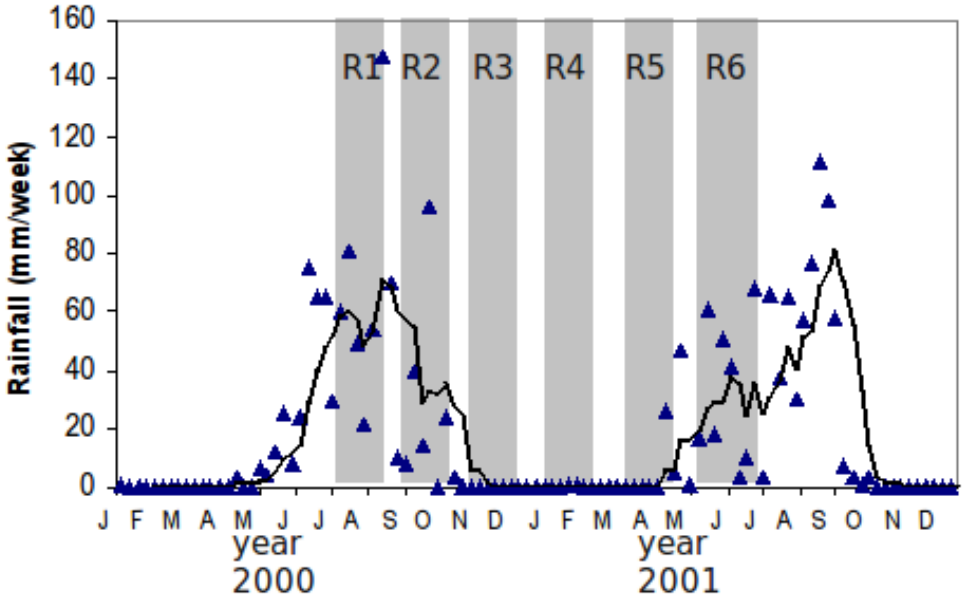


Figure 3.1: **Sampling intervals and rainfall in Navrongo/Northern Ghana during the study period** - Blue triangles represent weekly rainfall, the black line shows the 5 weeks moving average of rainfall. Grey bars represent the six sampling periods.

logical 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 database of the Navrongo Demographic Surveillance Site, which covers the whole district. From the area centered on 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+ [63]. Blood samples were collected on DNA ISOCode^{TM} Stix (Schleicher & Schuell) in intervals of two months, starting in mid 2000, resulting in a total of 6 samples per participant (R1-R2-R3-R4-R5-R6). Study participants who were sick at the time of the survey were referred to the routine health services. No antimalarial treatments were administered by the research team. 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.

3.3.2 DNA isolation and genotyping

Processing of DNA and PCR amplifications were described previously [52, 65]. The capillary electrophoresis (**CE**)-based genotyping technique used differentially labeled fluorescent dyes for each of the two allelic families of the polymorphic marker gene msp2 (VIC for 3D7-types and 6-FAM for FC27-types). The GeneMapper® Software version 3.7 (Applied Biosystems) was used for calculating fragment sizes. This typing technique had been

validated with a subset of our samples [52]. Direct comparison of CE-based results with those obtained by the previously used PCR-RFLP technique showed that the CE technique is more precise and more sensitive.
For fragment sizing by CE, 0.25 μl of nPCR product was combined with 10 μl ROX-labeled size standard (diluted 1:40 with water 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, mixed well, and after denaturation, samples were analyzed on an ABI PRISM 3700 genetic analyzer.

3.3.3 Data analysis

An in-house generated computer program was used to classify peaks sized by GeneMapper® software. To allow for the inaccuracy in size determination, peaks were assigned to size bins with a length of 3 base pairs since the genotyping applied to a coding region of DNA. A clone with a given length was considered to be present when the peak height exceeded a well- and plate-specific cutoff. The cutoff height was a product of an allelic family-specific constant, the average height for the size standard, and the relative height of the size standard for the plate. The allelic family-specific constant was chosen to give optimal detection of peaks using samples with known DNA content. Post-processing procedures applied to the raw data consisted in genotype calling and elimination of so called “bleeding” peaks, “plus-A” artifacts, and other allele-specific PCR artifacts. These procedures have been described previously in greater detail [52].
The genotyping data were first analyzed by calculating the frequencies of gain, loss and persistence of infecting clones. An infection present in the survey at time t, but not detected in the subsequent survey t+1 was considered a “loss” (+ -), whereas a “gain” (- +) was counted 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”. Two different methods were used to simultaneously estimate detectability and duration of infection by allowing for undetected infections: For the first so-called “triplet” approach the data were arranged as records of the presence or absence of individual parasite genotypes in sequences of up to three successive samples from the same individual. A sliding window of three consecutive samples was used so that each sample could appear at any position in the triplet. Four types of sequences were counted (++x; +-+; +-- ;+-?; where + represents positive, - negative, ? missing, and x can be any of the three). [58] give formulae for the relative frequencies of these patterns assuming parasite persistence to be a homogeneous first order Markov process, and parasite detection to be described by a constant detectability parameter. Assuming a multinomial distribution for the frequencies of the different sequence types, separate estimates for each age group of both detectability and of clearance rates were made using the software Winbugs version 1.4 [67].
The second approach, an immigration-death model, used sequences of all six observations and allowed for imperfect detection of an infection at some of the time points [37, 57]. This approach made a more efficient use of the data and adjusted for seasonality in the infection process. Detectability, force of infection and duration of an infection were simultaneously estimated. This entailed maximum likelihood fit to the frequency distributions

of the full sequences of six observations. We present the results for a minor modification of the best fitting model among those fitted previously to the pilot database as determined using the Akaike Information Criterion (**AIC**) [37]. The revised model corresponds to an immigration-death process with force of infection $\lambda(t)$ and clearance rate μ(a). Parameter estimates are given in Table 3.2. $\lambda(t)$ is the season-specific infection rate, taking distinct values for each of the 6 two-monthly periods corresponding to the inter-survey intervals, assuming the same seasonal pattern to have recurred annually since birth. In the best fitting model of Sama *et al* [37] the clearance rate, μ, was modeled as an exponential function of age, whereas in our updated model, separate clearance rates μ_1, μ_2 $\ldots\mu_6$ were estimated, depending on the age group of the host upon acquisition of an infection. This allows for more flexibility in representing the dependence of infection duration on the age of the host and was necessary to capture a peak of infection duration at intermediate ages. We parameterize the detectability s as a logit-linear function of age, i.e.:

$$\text{logit}(s(a)) = s_0 + s_1(a - \bar{a}) \,,$$

where $\bar{a}$ is the mean age and s_0 and s_1 are parameters to be estimated. Further details of the general approach were given previously [37, 57].

3.4 Results

From the 349 individuals enrolled at baseline, 1978 blood samples were collected during the one year follow-up in two-monthly intervals. All 1386 PCR positive blood samples were genotyped by our highly accurate technique. This amounted to a total of 6386 detected PCR fragments. The GeneMapper® analysis distinguished 103 different msp2 genotypes. 28 belonged to the FC27 allelic family and 75 were of 3D7-type. Genotypes belonging to the FC27 family generally reached higher allelic frequencies than 3D7 genotypes. The most frequent genotype represented 10.2% of all fragments detected (651/6386) and was of FC27 type. The most frequent 3D7 genotype represented 3.6% of all fragments. In the overall data set *P. falciparum* prevalence was 48% by microscopy and 75% by PCR [57]. The age distribution of PCR positivity shows a peak in the 5-9 year old children with 93% of these children being parasite positive (Figure 3.2a). Mean multiplicity measured in the PCR positive samples also peaked in the age group of 5-9 years. Both prevalence and the mean number of concurrent infections detected in a blood sample (multiplicity of infection, **MOI**) were lowest in the 60+ age group.

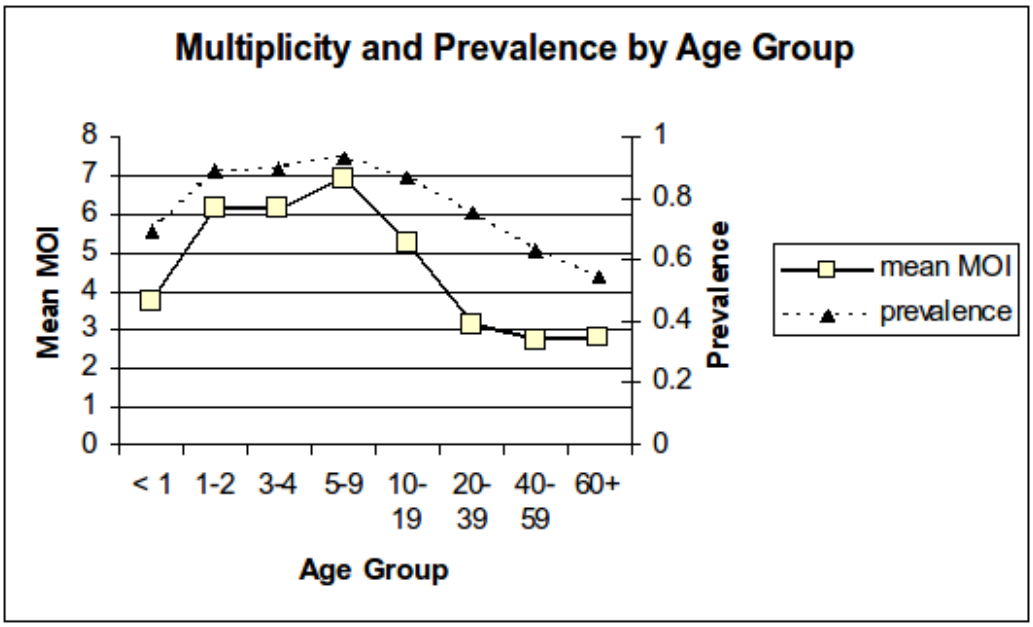
The highest prevalence by PCR (88%) and highest MOI (a mean of 5.4 concurrent clones per individual) was observed in Round 2 (Figure 3.2b). These peaks coincided with the rainy season as sampling for Round 2 started just after the rainfall had reached a maximum (Figure 3.1). Prevalence was lowest (60%) in Round 6, which marked the beginning of the next rainy season.

The apparent dynamics of *P. falciparum* infections can be described by the observed rates of clone appearance and disappearance (Figure 3.3). Rates of gain were higher in children than in adults, with a peak in the 5-9 years age group that showed an average of about 5 clones gained per interval. The observed rate of clone disappearance showed an opposite

Parameter		Unit	Value	
			ID model	Triplet model
λ_1	Force of Infection Aug/Sep	per yr	44.8 (CI 44.7-44.9)	n.a.*)
λ_2	Force of Infection Oct/Nov	per yr	19.2 (CI 19.1-19.3)	n.a.
λ_3	Force of Infection Dec/Jan	per yr	18 (CI 17.8-18.2)	n.a.
λ_4	Force of Infection Feb/Mar	per yr	7.5 (CI 7.4-7.6)	n.a.
λ_5	Force of Infection Apr/May	per yr	12.3 (CI 12.2-12.3)	n.a.
λ_6	Force of Infection Jun/Jul	per yr	39.6 (CI 39.3-39.8)	n.a.
$1/\mu_1$	Duration of infection (<1 yr)	days	155 (CI 128-183)	129 (CI 108-160)
$1/\mu_2$	Duration of infection (1-2 yrs)	days	256 (CI 254-259)	198 (CI 143-322)
$1/\mu_3$	Duration of infection (3-4 yrs)	days	257 (CI 254-259)	196 (CI 162-247)
$1/\mu_4$	Duration of infection (5-9 yrs)	days	319 (CI 318-320)	216 (CI 169-297)
$1/\mu_5$	Duration of infection (10-19 yrs)	days	176 (CI 156-196)	189 (CI 143-279)
$1/\mu_6$	Duration of infection (20-39 yrs)	days	129 (CI 124-134)	124 (CI 93-186)
$1/\mu_7$	Duration of infection (40-59 yrs)	days	126 (CI 113-139)	158 (CI 95-472)
$1/\mu_8$	Duration of infection (>60 yrs)	days	131 (CI 114-148)	190 (CI 97-5924)
s_0	Logit detectability	at 20 yrs	-0.84 (CI -0.91 - -0.78)	n.a.
s_1	Change in logit detectability	per 10 yrs	-0.17 (CI -0.21 - -0.13)	n.a.
q_1	Detectability (<1 yr)	percent	n.a.	61 (CI 48-74)
q_2	Detectability (1-2 yrs)	percent	n.a.	46 (CI 34-58)
q_3	Detectability (3-4 yrs)	percent	n.a.	46 (CI 39-52)
q_4	Detectability (5-9 yrs)	percent	n.a.	40 (CI 34-47)
q_5	Detectability (10-19 yrs)	percent	n.a.	33 (CI 26-41)
q_6	Detectability (20-39 yrs)	percent	n.a.	33 (CI 19-48)
q_7	Detectability (40-59 yrs)	percent	n.a.	22 (CI 09-35)
q_8	Detectability (>60 yrs)	percent	n.a.	13 (CI 03-23)

Table 3.2: **Parameter estimates of statistical models** - *) n.a.: not applicable; the estimation method did not provide these estimates.

a)



b)

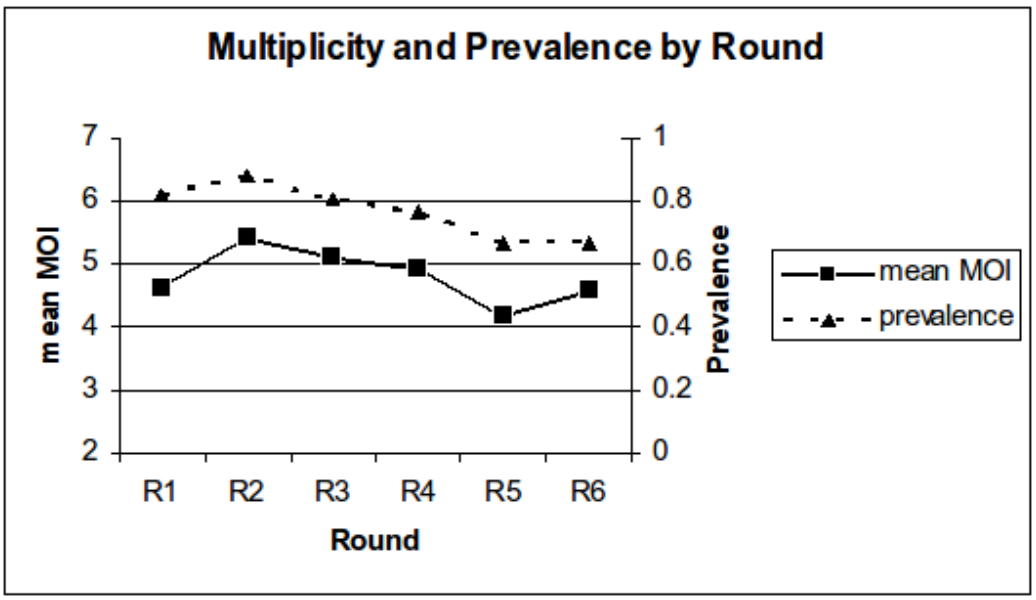


Figure 3.2: **Multiplicity of infection and prevalence** - Mean multiplicity of infection (MOI) and prevalence by PCR (a) by age and (b) by survey. Mean multiplicity is calculated from PCR positive samples only.

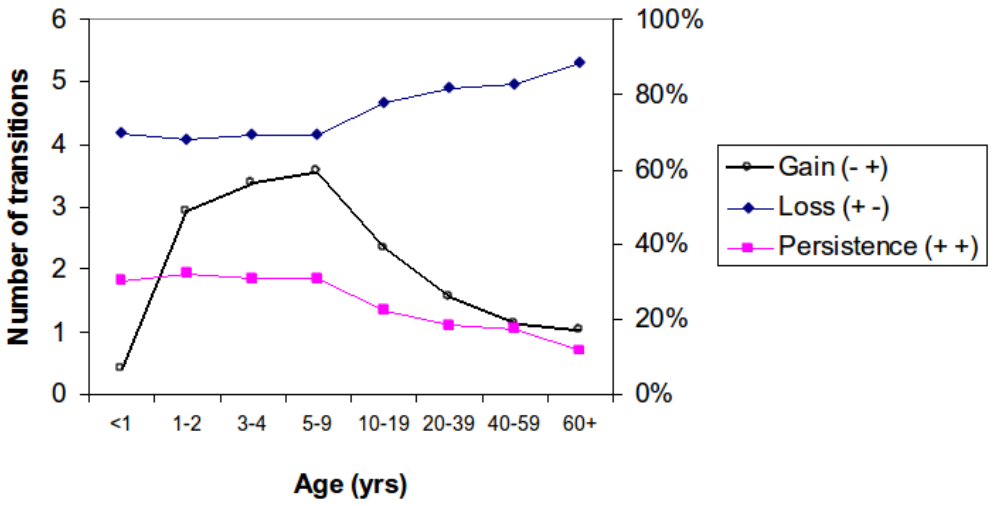


Figure 3.3: **Transitions** - Transition types observed between two consecutive survey rounds (gains or losses of parasite clones and persistent clones) by age group. Transitions were monitored in 1978 blood samples collected longitudinally from 349 individuals from Northern Ghana.

trend. More clones are lost between two surveys in older individuals than in young children. The apparent persistence of clones thus decreased with age.

The data set used in the present analysis to estimate detectability and duration of infection from longitudinal genotyping data contained 216 individuals with complete records for each round. Both statistical methods for estimating detectability found that even highly sensitive nPCR combined with precise capillary electrophoresis and GeneMapper® based fragment sizing detected less than half of the clones present at a given time in an individual (Figure 3.4). Both methods also agreed that in the younger age groups <10 years, detectability was higher than in older age groups, reaching a minimum of 17% in the oldest individuals. This age-dependent decline of detectability was in line with decreasing microscopic parasite density with age (Figure 3.4). Our implementation of the immigration-death model constrained the relationship of detectability with age to be monotonic, leading to a decrease in detectability with age. Overall mean detectability in all six rounds was estimated to be 30% by the immigration death (**ID**) model.

The duration of infection was also estimated by both models (Figure 3.5). The "triplet" model gave estimates of duration much longer than those based on direct observations in blood samples, averaging 168 days (95% CI 144-202) over the whole age range. The immigration-death (ID) model led to even longer estimates of mean duration of 194 days (95% CI 191- 196). Both models considered each age group separately and thus allowed non-monotonic age effects. Figure 3.5 shows a peak in duration of infection in children 5-9 years of age. Infection duration for infants and older ages were shorter. Table 3.2 provides

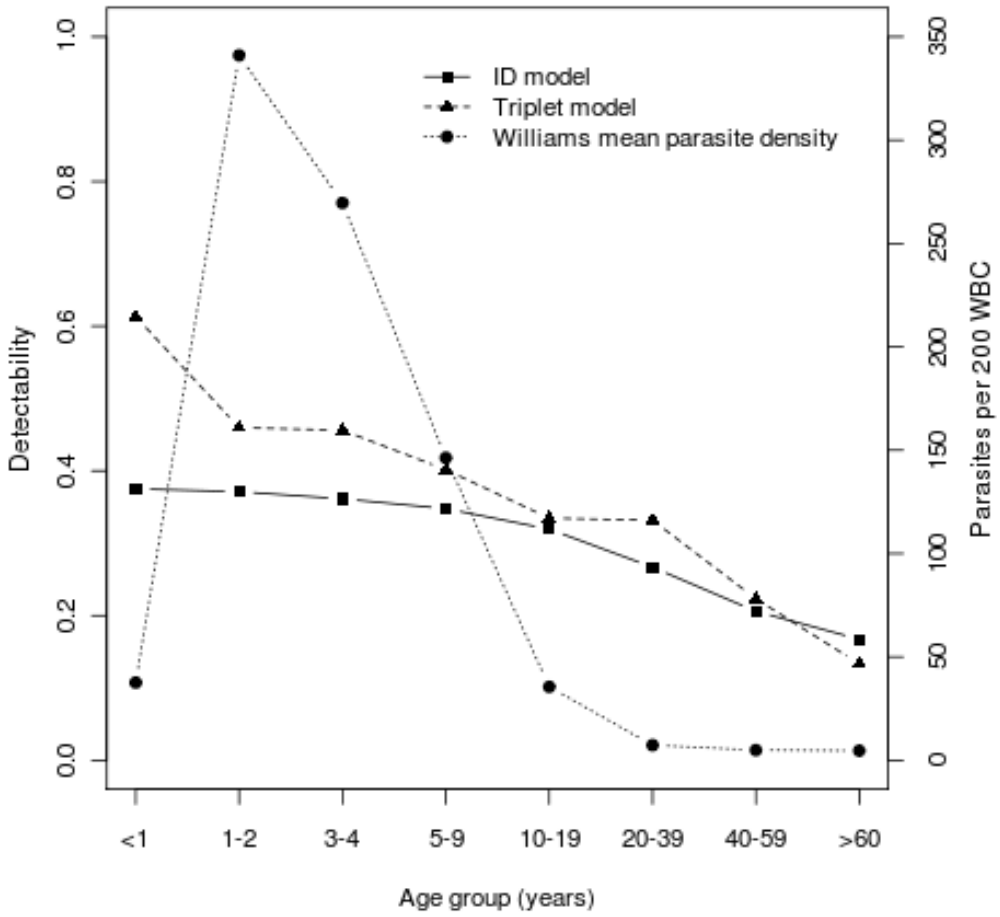


Figure 3.4: **Detectability and parasite density** - Detectability by age group and Williams mean parasite densities by age assessed by microscopy.

a summary of the parameter estimates. Compared to the previously published version of our ID model (model 6 in [37]), the current model results were superior according to AIC value (8005.1).
Estimates of the age profile of MOI need to be corrected for the effects of imperfect detection of parasite clones. Figure 3.6 shows the observed versus the true MOI corrected for imperfect detection by both models.
The age pattern of multiplicity, adjusted for detectability was similar, whichever model was used to estimate detectability. Estimated MOI in infants is lowest at roughly 10 infecting clones per child. There is a peak in 5-9 year old children with about 19 concurrent infections per child. All ages above 20 years show a lower MOI when adjusted for detectability by the ID model. Overall, correction with the immigration death model provides higher estimates of MOI and gives estimates of overall mean MOI reaching a maximum of 18 at the end of the wet season and a minimum of 14 at the end of the dry season, assuming an average detectability of 0.3.

3.5 Discussion

Although the duration of malaria infections is a key parameter in mathematical models of transmission, few studies have attempted to estimate this quantity from field data. Analyses of microscopically determined parasitaemia are complicated by the complex patterns of disappearance and reappearance of parasites from the peripheral blood, and this leads to difficulties in estimating infection duration from repeated survey data [48]. In situations where repeated superinfection is occurring, it is in any case the persistence of individual parasite clones that needs to be assessed, and this requires the use of parasite typing data.
Some authors have interpreted the complex patterns of intermittent appearance of individual clones as rapid turnover [78]. Our studies strongly suggest these patterns result from fluctuations in density of persisting infections, and that new inoculations are relatively infrequent. We find a remarkable agreement between the average duration of 210 days of artificial *P. falciparum* infections in malariatherapy patients [49, 80] and estimates from our field studies. This agrees with observations that infections from the transmission season in the Sudan are retained as sub-patent asymptomatic infections throughout the long transmission free dry season [81].
Quantitative analysis of longitudinal studies of malaria genotypes requires a typing system with high resolution and discrimination power. Traditional side-by-side runs of PCR products or RFLP fragments on polyacrylamide gels were laborious, and allowed comparison of only a limited number of samples. Capillary electrophoresis for sizing of PCR fragments, combined with fluorescence-labels for the discrimination of msp2 allelic families [52] makes it possible to accrue much larger and more accurate datasets for analyzing parasite dynamics. A further advance would be to test the reproducibility of the results using a second marker gene in addition to msp2.
There are limits to possible improvements in field and the laboratory systems for detecting parasites. In previous studies we have estimated detectability to be 0.51 in young Tanzanian children [69], and 0.35 and 0.47 in the pilot study in Ghana, that include 100 individuals

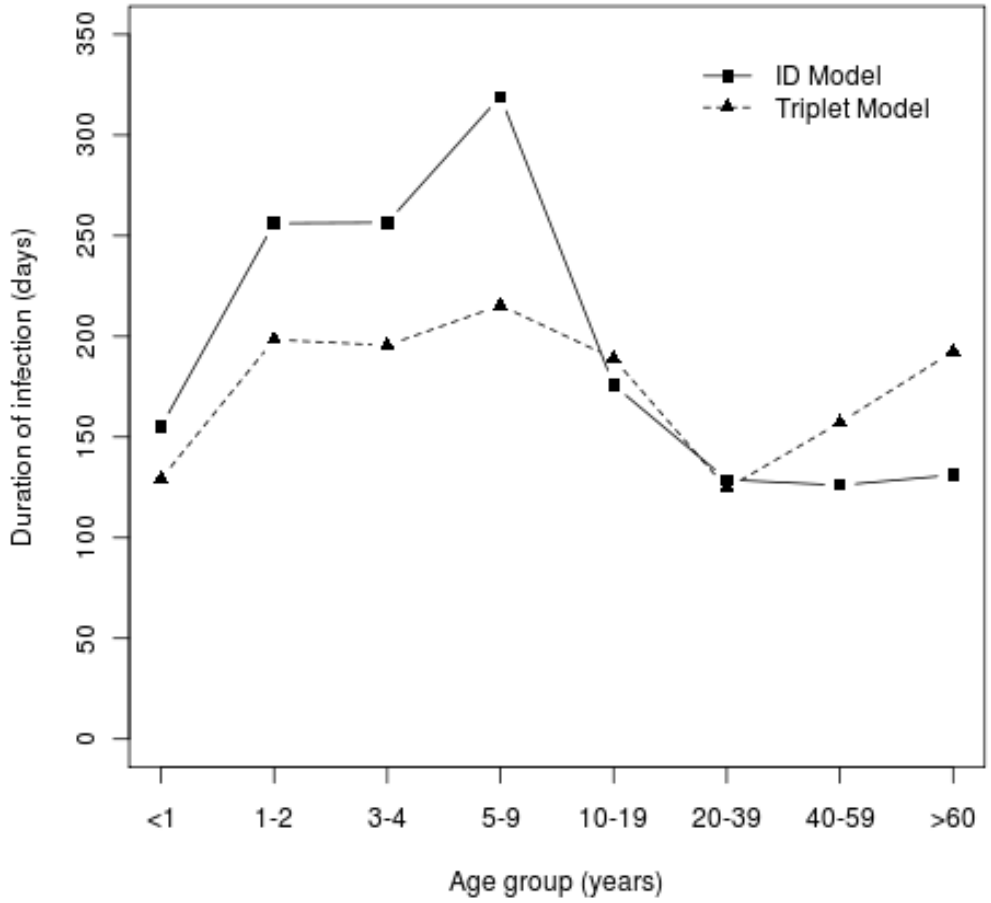


Figure 3.5: **Duration of infection by host age**

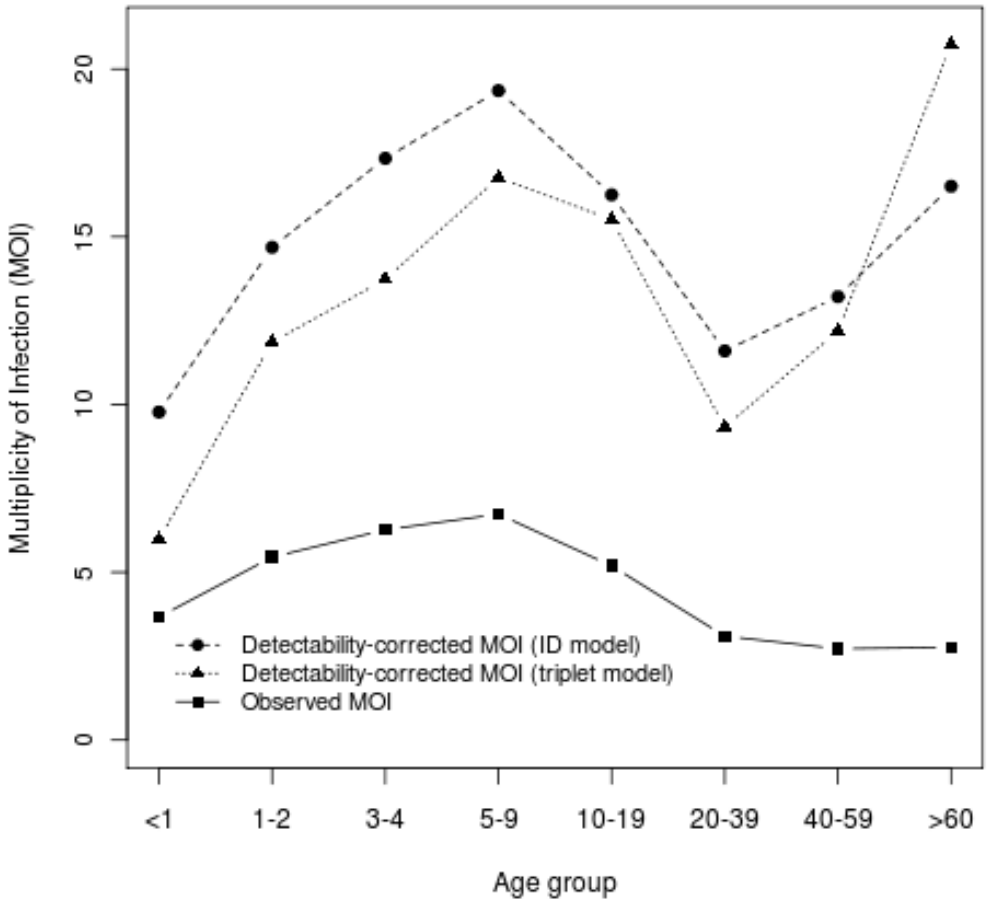


Figure 3.6: **Multiplicity of infection** - Observed mean multiplicity of infection (MOI) by age group in Northern Ghana and MOI. MOI was determined from PCR positive blood samples only.

of all ages, with the values in the latter study depending on the genotyping method applied [52]. Standard molecular techniques thus detect at most about half of all clones and even nPCR misses many low density infections. Increasing the volume of blood sample per DNA extraction and thus the amount of target DNA per PCR augments detectability, but this will never reach 100% because of sequestration of some parasite clones. Nor will collection of samples at very short intervals overcome the problem of imperfect detection [82].

We will therefore continue to need statistical methods that allow for imperfect detection, with the immigration-death model [37, 57] representing our preferred approach for estimating *P. falciparum* infection duration from typing data. This makes use of all the data while allowing for variation in effects of age on both detection and clearance rate, as well as for seasonal variation in transmission intensity. It could also allow for other factors known to complicate relationships between age and infection status, such as the increase in attractiveness to Anophelines as body size increases [83, 84].

Disadvantages of this approach are that it necessarily assumes that reinfection with the same clone is a rare event. Thus, lack of detection of a PCR fragment between two time points that were both positive, was treated as due to either sequestration of the parasite clone at the time of blood sampling or to a density below the detection limit of nPCR. This assumption is justified by the low frequencies of the individual genotypes (Figure 3.7). This approach also treats parasite clearance and detection as independent of age of the infection, which is equivalent to assuming an exponential distribution of durations. The durations of untreated infections fit better to Gompertz or Weibull distributions than to the exponential [49], and to test the robustness of our conclusion, we plan to carry out additional analyses assuming more realistic distributional forms for this parameter.

The generally higher parasite densities in children are expected to lead to higher detectability [85] but we find that detectability decreased gradually with age, whereas density shows a pronounced peak in young children. Correction for this age pattern in detectability makes the age effects in MOI seem less pronounced than they appear in simple descriptions. The MOI observed in our study population is high, though when adjusted for age is lower than the median of 3 clones per parasite positive carrier in 3 months old children found in a study in the Ashanti region of Ghana [86]. Even in high transmission areas MOI is generally low in the youngest children, and shows a peak at about the same age as the prevalence as assessed by microscopy [63, 69, 87]. In lower transmission areas (including another area in Ghana [88]), MOI is both lower and less age dependent [89–91]. Our analyses however suggest that much of the age dependence in MOI in high transmission areas results from lower detectability in adults who simultaneously harbor many sub-patent infections.

Estimates of the age-dependence of duration are even more strongly modified by allowing for age-dependence in detectability. When this is ignored, parasite persistence appears to decrease strongly with age, corresponding to the idea that stronger immunity in older individuals clears infections. In a study in Papua New Guinea a median duration of clonal infections of >60 days was estimated in 4 year olds, but a median of only 15 days was estimated for children of 5-14 years [76]. By allowing for imperfect detection, our analyses found a very different pattern, with the duration of infection shorter in infants than in older children (Figure 3.5). This is in agreement with our earlier analyses that suggested

that there is a decrease in clearance rates with age during the first few years of life in Tanzanian children [58, 60, 92]. Clearance rates do seem to increase again in semi-immune adolescents and adults, leading to somewhat shorter duration of infection, but the oldest individuals have increased persistence that might correspond to loss of immunity in older people, resulting in higher parasite densities.

One factor not considered in our analysis is the potential influence of antimalaria treatment. In Navrongo, treatment was mostly given to <2 years old children, but was relatively infrequent. Thus, the natural clearance is inseparable from curative therapy and duration estimated for the youngest age group is likely an overestimation of natural clearance.

In conclusion, the age patterns of MOI, detectability and duration of infection are the result of the interplay of the infection and clearance processes, but their analysis is complicated by imperfect detection of parasites. The present analysis used data from reliable high-throughput typing and was much larger than any available previously. This gives us much more precise estimates of the age-specific parameters and increases our confidence that effects of parasite clearance can be separated statistically from those of imperfect detection. It supports our previous suggestions that the main effect of acquired immunity is not to increase clearance rates [37], but that there is a non-monotonic relationship between duration and age in endemic populations [69].

Name	Frequency	Mean Peak Height/Size Standard Height*)
Fc27genotype297	87	9.964
Fc27genotype300	4	2.527
Fc27genotype309	17	3.169
Fc27genotype312	15	5.487
Fc27genotype318	21	5.495
Fc27genotype327	22	5.473
Fc27genotype333	5	2.527
Fc27genotype336	459	14.160
Fc27genotype345	6	9.728
Fc27genotype351	13	5.502
Fc27genotype354	5	15.739
Fc27genotype360	83	9.050
Fc27genotype366	6	2.743
Fc27genotype372	284	11.025
Fc27genotype375	5	5.344
Fc27genotype381	11	5.130
Fc27genotype384	4	12.125
Fc27genotype387	3	2.047
Fc27genotype393	11	3.445
Fc27genotype399	10	12.220
Fc27genotype405	18	5.358
Fc27genotype408	59	12.958
Fc27genotype417	325	11.140
Fc27genotype447	3	10.769
Fc27genotype456	13	5.637
Fc27genotype516	14	3.377
3D7genotype206	33	6.020
3D7genotype209	8	7.008
3D7genotype212	28	7.676
3D7genotype215	4	14.065
3D7genotype218	35	3.520
3D7genotype221	9	4.578
3D7genotype224	41	6.255
3D7genotype227	19	7.393
3D7genotype230	38	11.002
3D7genotype233	12	12.591
3D7genotype236	108	9.044
3D7genotype239	14	3.320
3D7genotype242	50	10.425

Figure 3.7: **List of all 100 msp2 genotypes detected in cohort from Northern Ghana (349 individuals)**

CHAPTER 4

The distribution of *Plasmodium falciparum* infection durations

Authors: Michael T Bretscher, Nicolas Maire, Nakul Chitnis, Ingrid Felger, Seth Owusu-Agyei, Tom Smith

4.1 Abstract

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

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

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

4.2 Background

The duration of a *P. falciparum* infections is related to the duration of potential infectivity of the host. It affects the magnitude of transmission from humans to the mosquito population and gains special significance in settings where malaria transmission is seasonal: the fraction of infections surviving a transmission-free dry season constitutes the founder population for the new transmission season. Rational planning of a malaria control or elimination therefore profits from accurate measurements of infection duration. How much variation there is in the duration of natural infections is largely unknown, but important for similar reasons. A case study suggests that single infections may in extreme cases last up to 8 years [93]. Current knowledge about within-host dynamics and the distribution of *P. falciparum* infection durations comes mostly from malariatherapy data [49]: before the arrival of suitable antibiotics, infection with malaria was a common method to treat neurosyphilis.

Analysis of such data is facilitated by the fact that the start- and end-points of every infection are approximately known, and that therefore standard methods of statistical survival analysis can be applied. A comparison of various parametric survival distributions suggested that the Gompertz and Weibull distributions gave the best fit to these data, followed by the gamma, lognormal, and exponential distributions [49]. An average duration of approximately 200 days was found. Infection durations much shorter or longer than the mean were rare.

However, malariatherapy data may not accurately mirror the situation in naturally exposed populations: the patients were immunologically naïve, infected with syphilis, and did not have multiple concurrent infections. Moreover, the *P. falciparum* strains used for therapy were selected by physicians for optimal curative properties as well as for low clinical virulence [94]. Thus it is not clear whether the distribution of infection durations would be the same in human hosts who have experienced high malaria transmission throughout their lives, and possibly have multiple concurrent infections caused by wild-type parasite strains.

A valuable source of information about infection dynamics in natural populations is cohort data on malaria infection, obtained using DNA-based diagnostic methods. These have the advantage that infecting clones can be distinguished on the basis of highly polymorphic genetic markers, such as Merozoite Surface Protein 2 (**msp2**). However, analysis of such data is not straightforward using standard techniques because detection of *P. falciparum* clones is imperfect, even when using polymerase chain reaction (**PCR**). Several dedicated statistical methods, allowing for imperfect detection, have been developed [37,57,58,60,95]. The assumption of a constant clearance rate of infections is common to all these approaches. This has a long tradition and there are practical reasons for doing this: a constant clearance rate means that the rate at which infections are cleared is independent of the age of an infection, which implies an exponential distribution of infection durations. This simplifies the required mathematics enormously since it is not necessary to keep track of the age structure of the infection population. From a biological point of view, however, exponential survival of infections seems not very plausible. The study of how durations of infection are distributed is the quest for a statistical description of one important aspect of within-host dynamics. Such analyses may yield information which can be used to validate

process-based within-host models.
In continuation of [37,57,58,60] we have developed a method to analyse molecular cohort data and measure parameters of infection dynamics. We use a more complete dataset from the study analysed by Sama *et al* [37]. The main findings of [37] are therefore briefly explained: using the statistical model described earlier in [57], Sama *et al* studied seasonality and age dependence of the following parameters of infection dynamics: λ (the force of infection, **FOI**), μ (the clearance rate, which is the inverse duration of infection and implies an exponential distribution of durations) and q (the detectability parameter, denoting the probability of detecting a specific clone in a blood sample, given it is present). In total, Sama *et al* compared twelve different model parameterizations with respect to goodness of fit, using a longitudinal, age-stratified dataset from Navrongo, Northern Ghana. One of the main findings was that the detectability of infections declines with host age in a very pronounced way, suggesting an effect of cumulative exposure, a proxy for acquired immunity, on parasite densities. Contrary to expectation, the duration of infection was hardly affected by host age. The present analyses consider the sensitivity of these results with respect to the assumption of constant clearance rates, and provide estimates of the distribution of infection durations.

4.3 Methods

4.3.1 Study design and sample collection

A one year longitudinal study of malaria infection was conducted in the Kassena-Nankana district, in the upper East region of Ghana [37, 52, 57, 63]. The malariological situation in this area is characterized by very high prevalence and multiplicity of infection [63,64], and year-round transmission with seasonal variation in transmission intensity [37]. A total of 349 individuals of all ages were followed up over one year in 2-monthly intervals. New births were recruited during the follow-up so as to ensure that the age distribution remained the same throughout the study. Blood was collected on ISOCodeStix^{TM} PCR template preparation dipsticks (Schleicher & Schuell, Dassel, Germany).

4.3.2 Genotyping

DNA was eluted from ISOCodeStix^{TM} and screened for presence of *P.falciparum* by PCR. Processing of stix and PCR conditions have been described in detail before [65]. In brief, samples that tested positive for presence of *P. falciparum* were subjected to PCR using primers specific for the msp2 locus. Different alleles were distinguished on the basis of length polymorphisms, by means of automated capillary electrophoresis technology. The obtained data files were further processed using the GeneMapper® software and an in-house generated software, which facilitates identification of known alleles from the raw output of GeneMapper® and transforms the data into different formats suitable for data management and statistical analysis.

4.3.3 Data preparation

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

4.3.4 Models of infection dynamics

A selection of process-based statistical models, similar to the ones presented in [37], were devised and compared to the data. In the models, three main processes are assumed to determine frequencies of the different binary patterns in each human host: acquisition, clearance, and detection of infections. Given mathematical models for each of the three, a likelihood can be calculated as explained in the following section. The simplest possible model represents each process by a single parameter: the force of infection λ (no. of infections acquired per person year), the duration of a clonal infection (in the simplest case modeled as an exponentially distributed random variable, with scale parameter equal to the inverse clearance rate, $1/\mu$), and the detectability q (the probability of detecting any present *falciparum* clone in a blood sample by PCR). Such a simple model is not able to capture several important characteristics of real data, such as seasonality in transmission or changes in detectability with increasing immunity of the host. These have been shown by [37] to be present in the Navrongo dataset and need to be incorporated into a model in order to yield unbiased parameter estimates. As a starting point for our analysis, we use the best fitting exponential model from [37], which was fitted to a partial dataset from the same study: the FOI parameter $\lambda(t)$ was modeled as a function of season alone, meaning that for every two-month season a separate parameter λ_i was estimated. The resulting pattern of seasonal transmission was assumed to have repeated since the birth of every host. We extended the work of [37] to allow the use of four parametric survival distributions for modeling of the clearance of infections: these are the exponential, Weibull, gamma and lognormal distributions (Table 4.1). Except for the exponential distribution, which is characterized by a single (scale) parameter, these distributions require two parameters. In the following we will refer to these as "scale" and "shape" parameters, ignoring possible distribution-specific names. Because the best-fitting model of [37] showed no age-dependence of the duration of infection, and because the present analysis is intended to be a proof of concept, we chose to parameterize the survival models with simple constant values, rather than e.g. modeling them as functions of host age. The age dependence of detectability was modeled as a logit-linear function,

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

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

Table 4.1: **Survival distributions** - Selected properties of the different survival models used in this analysis are given. The exponential is the only single-parameter distribution, and its properties depend entirely on the mean duration of an infection. All other distributions make use of two parameters. We refer to them as scale and shape, respectively, instead of using distribution-specific names. For each survival model, a restricted range of possible "shapes" exists, with restrictions being different among the distributions. Abbreviations: for the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$, for the lower incomplete gamma function $\gamma(s,x) = \int_0^x t^{s-1}\,e^{-t}\,dt$, and for the error function $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

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

Table 4.2: **Parameter estimates from simulated datasets** - The data were produced by stochastic simulation using survival models and parameter values from [49], with constant values of FOI and detectability. Every row corresponds to a survival model tested on a simulated dataset. Columns correspond to the different parameters of infection dynamics, with estimated parameter values shown to the left of the true values. The FOI is given in infections acquired per year, otherwise the time unit is per 2 months (corresponding to the survey interval).

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

$$q(a) = \frac{1}{e^{-l(a)} + 1}. \tag{4.1}$$

4.3.5 Model equations

Let $(n_{k,1}, n_{k,2}, .., n_{k,63})$ denote the realizations of 63 Poisson random variables with means $(\omega_{k,1}, \omega_{k,2}, .., \omega_{k,63})$, where $\omega_{k,i}$ is the expected frequency of observed pattern i in individual k. In order to derive the $\omega_{k,i}$, we firstly derive the expected frequencies of the 21 (hypothetical) true patterns, $\tau_{k,i}$ representing true infection status. True patterns are also indexed using the binary number they encode, but only patterns comprising a single uninterrupted subsequence of "ones" are considered.
Individual k is of age b_k at the time of the baseline survey, t_b, implying that it was born at time $t_b - b_k$. As some parameters of infection dynamics can be functions of time (seasonality), and others functions of the age of the host, we will use t as variable of integration, and refer to the age of an individual at time t as $a_k(t) = t - t_b + b_k$. We denote the length of inter-survey intervals with ε, and assume equally spaced surveys.
Consider for the acquisition of infections a Poisson process with intensity $\lambda(t)$, the force of infection. Since true patterns consist of uninterrupted sequences of "ones", every true pattern can be defined by the times of the first and the last survey where the infection is present, $t_{1,i}$ and $t_{2,i}$, respectively. These imply, given the study design, a time interval where its causative infection may be acquired and a time interval where it must be cleared. Those intervals can be obtained for each true pattern from the age of the individual at baseline, b_k, and the number of surveys, s, in the study.

[1] For the sake of comparing our results with those from [37], $\bar{a}$ was set to 120.72 (in units of 2 months, corresponding to the survey interval). This is the average age in the partial dataset used by [37].

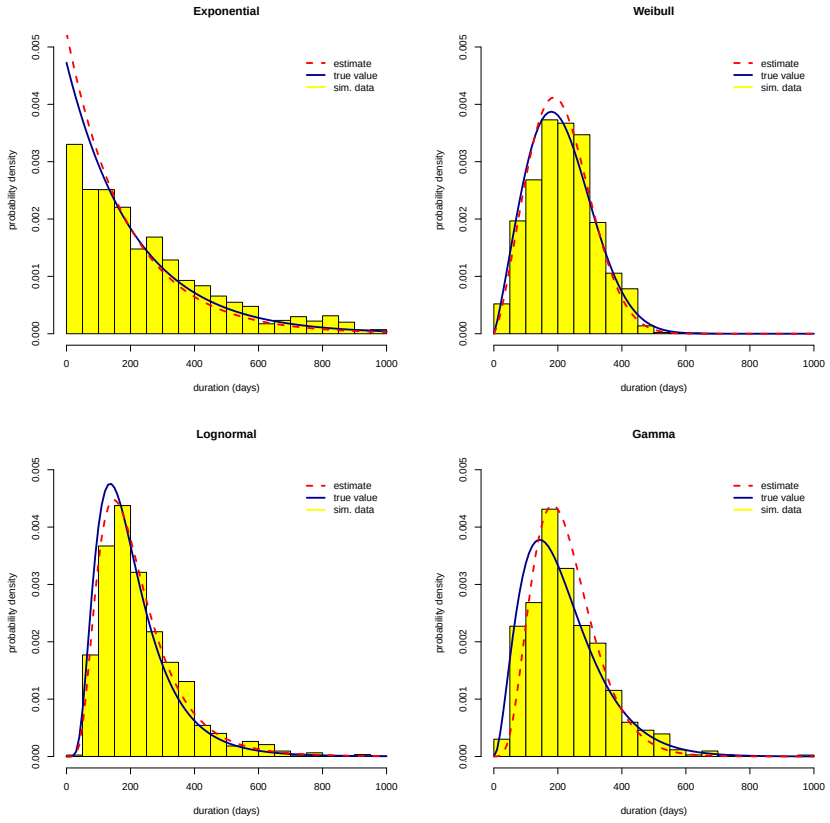


Figure 4.1: **Validation using simulated data** - Exponential, Weibull, lognormal and gamma models were fitted to simulated datasets. The histograms indicate the actual distribution of the duration of those infections which were present at some point during the simulated study. Solid blue lines indicate the PDF of the survival distribution that the actual durations were sampled from - the estimates from malariatherapy data as in [37]. The dashed lines indicate the PDFs of the survival distributions as recovered by our statistical model.

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

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

and

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

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

$$\tau_{k,i} = \int_{\alpha_{k,i}}^{\beta_i} \lambda(t) p_{k,i}(t)\, dt\ . \tag{4.2}$$

This probability depends on the distance to the surveys in time, and the properties of the survival distribution used for modeling clearance of infections. The properties of the survival distribution may in turn depend on the age of the host at time t, when the infection is acquired.
The probability that an infection acquired at time t generates pattern i, $p_{k,i}(t)$, is equal to the probability that the infection is cleared in the interval $[\gamma_i, \delta_i]$. Therefore

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

where f and S are the probability density function and survivor function, respectively, of the survival distribution used to model clearance of infections.
To obtain γ_i , the start of the clearance interval, we consider again patterns 001111 and 110000 as examples, and conclude, trivially, that an infection cannot be cleared before $t_{2,i}$, the time of the last survey it is present conditional on producing pattern i.

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

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

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

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

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

or, in terms of the survivor function S,

$$\tau_{k,i} = \int_{\alpha_{k,i}}^{\beta_i} \lambda(t) \left[S(\gamma_i - t) - S(\delta_i - t) \right] dt . \tag{4.3}$$

A more formal but equivalent approach, explaining the presented heuristics, is outlined in section 4.7.
The expected frequencies $\omega_{k,i}$ of observed patterns in individual k can be obtained using the probability $P_{i,j}$ that true pattern i gives rise to observed pattern j, as follows:

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

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

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

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

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

where q is the host-specific detectability, possibly modeled as a function of the age of

the host. Considering all observed patterns j and all hosts k, and assuming a Poisson distribution of the actual data $n_{k,j}$ with expectations $\omega_{k,j}$ we obtain the overall likelihood

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

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

4.3.6 Model implementation and parameter estimation

All models were implemented using the Java™ programming language. Maximum-likelihood estimates of parameters were obtained by minimization of AIC values using the "UncMin" algorithm by [96]. A Java version of this algorithm was obtained from `http://www1.fpl.fs.fed.us/optimization.html`. Numerical integration was performed using a Romberg integration algorithm with modified stopping criterion[3], from the Apache Commons Math library [97].

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

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

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

4.4 Results

4.4.1 Simulated data

For the purpose of validating our method, simulated datasets were produced using Monte-Carlo simulation. Number and ages of the hosts in the simulated datasets were identical to the Ghanaian dataset, and a constant, homogeneous FOI of 18 infections per person and year was assumed, as approximately measured on average by [37]. The number of infections a person experienced between birth and the last survey round was sampled from a Poisson distribution with mean λa, where λ is the force of infection, and a is the age of the human host at the last survey round. Actual infection time-points were then sampled from a uniform distribution within said interval. Subsequently, a duration was assigned to each infection, using one of four survival distributions with parameter values as measured from malariatherapy data [49]. A Bernoulli random variable with mean $q = 0.5$ was then used to determine for each survey round and clone whether detection was successful or not. All parameters of infection dynamics could be recovered well from the simulated data, as shown in Table 4.2 and Figure 4.1.

4.4.2 Estimates from the Ghanaian dataset

All of the 216 study participants included in the statistical analysis tested positive for *P. falciparum* on at least one survey. Parasite prevalence in the dataset was 48% by microscopy, 75% by PCR, and the mean MOI was 4.5 per person (these measures are not corrected for imperfect detection). A total of 103 different msp2 genotypes were found, with the most frequent genotype representing 10.2% of all fragments detected.
The four different models for infection survival showed the following order of goodness of fit, as measured by AIC: the Weibull model fits the data best (AIC: 8029.1), followed by the gamma (AIC: 8029.4), lognormal (AIC: 8045.1), and exponential model (AIC: 8127.4). Parameter estimates are given in Table 4.3, and the correlation matrix of the Weibull model in Table 4.4. All non-exponential distributions show an increased clearance in the early stages of an infection, i.e. they are positively skewed (to the left). The estimated mean durations, which can be calculated from the scale and shape parameters (Table 4.3) and the distribution-specific expressions for the mean (Table 4.1), are as follows (in days): 139.9 (Weibull), 54.9 (gamma), 205.3 (lognormal), and 219.7 (exponential). There is a substantial difference between the FOI estimates of the gamma and Weibull models, which are very similar in terms of goodness of fit (Figure 4.3). Measurements of detectability are in good quantitative agreement (Figure 4.4). All models measure a detectability of between 40% and 50% in young ages, which decreases to just above 10% in the old ages.

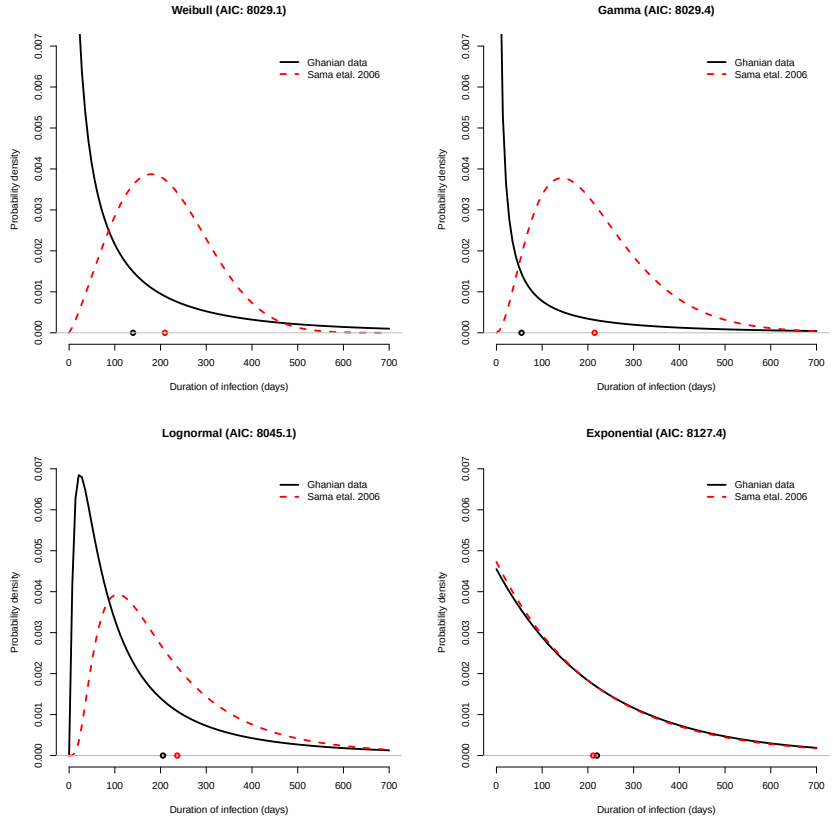


Figure 4.2: **Results from the Ghanaian dataset compared to malariatherapy data** - The PDFs of the distributions of infection duration as measured from the Ghanaian dataset are plotted in solid blue lines together with the estimates from malariatherapy data (dashed red lines) [49], in order of decreasing goodness of fit. Circles on the abscissa indicate the corresponding mean durations. The Weibull survival model fitted the data best, followed by the gamma, lognormal, and exponential models. With the exception of the gamma model, estimated mean durations differ only slightly between malariatherapy data and the data from the exposed population. The estimated non-exponential distributions from the Ghanaian dataset are positively skewed.

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

Table 4.3: **Parameter estimates from the Ghanaian dataset** - The results of fitting four statistical models to the Ghanaian dataset are shown. These models use different survival distributions for modeling of the clearance of infections. Goodness of fit is indicated by AIC, and parameter values are given with the corresponding standard error. The Weibull fits the data best, as indicated by its lowest value of AIC, but is closely followed by the gamma model. Seasonality in the force of infection was accounted for by estimating a separate parameter for every two-month season, assuming the same seasonality pattern had repeated in the past. Scale and shape parameters are specific to each distribution, as explained in Table 4.1. The detectability can be calculated for a host of given age using equation 4.1.

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

Table 4.4: **Correlation matrix** - Correlation matrix for the parameters of the best-fitting Weibull model. An expected overall negative correlation between the average FOI and the mean duration of infection may be obscured, as these are functions of several parameters. However, a clear positive correlation between the shape parameter s_2 and detectabililty of an individual of average age, q_1, indicates that different interpretations of the data are possible concerning clearance of infections and detectability.

4.5 Discussion

By extending the method of [37,57], such that parametric survival distributions can be used for modeling of infection survival, it has become possible to obtain more detailed information on clearance of *P. falciparum* infections in naturally exposed populations. A validation with simulated datasets suggests that parameters are identifiable using our approach and the present study design.

4.5.1 Distribution estimates

All estimated distributions are positively skewed (Figure 4.2), with exception of the exponential distribution, which does not have the freedom to measure more than the mean duration. Positive skewness means that most infections are cleared rather soon after inoculation. This is very different from the estimates of [49] who find that most infections in malariatherapy data last for an amount of time similar to the mean duration, and that very short or very long infections are rare. It also becomes apparent that in the Weibull and gamma models, which fit the data best, this effect is more pronounced than in the lognormal model. We attribute this to fact that the PDF of a lognormal distribution is constrained to be unimodal and therefore does not have the freedom to estimate such extreme early clearance of infections, as would be required in order to attain a better fit to the data. Thus, a consistent picture emerges, where the finding of early clearance of infections is independent of the clearance model. This positive skew of distributions is consistently associated with a lower estimated mean duration, which suggests that assuming an exponential distribution may lead to an overestimation of the mean infection duration. Across the models, shorter estimates of the mean duration are in turn associated with higher estimates of FOI. This

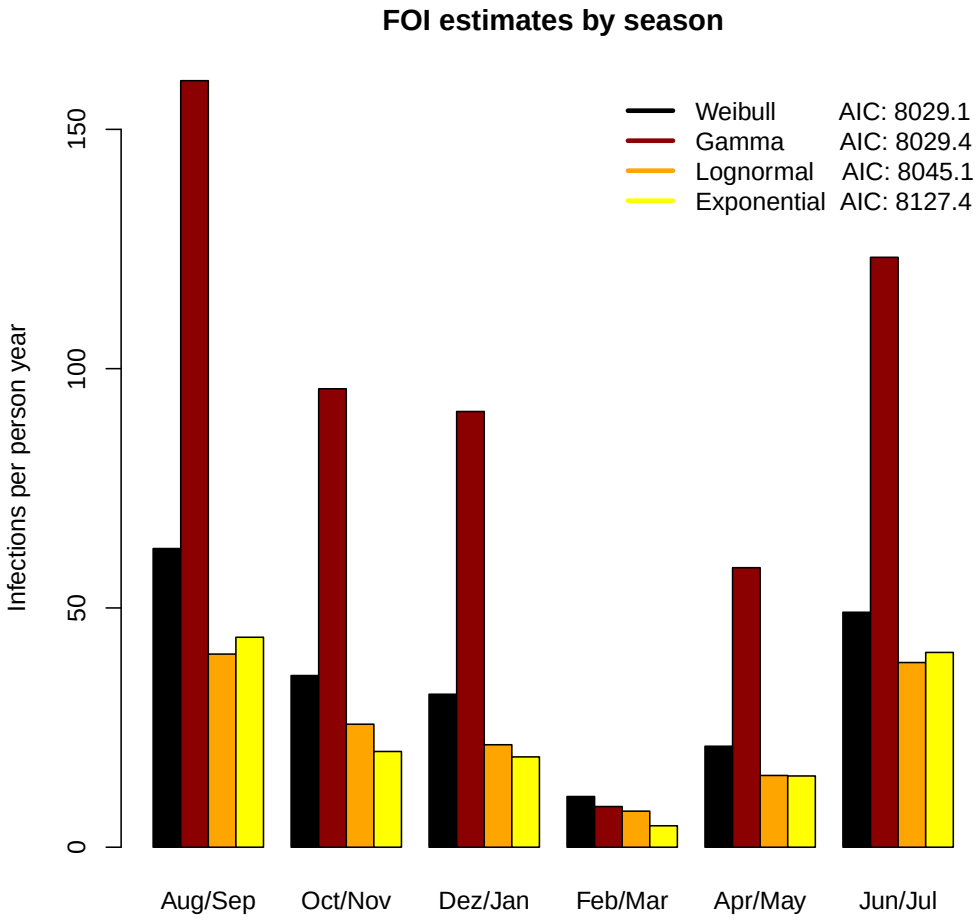


Figure 4.3: **Estimates of the force of infection** - Different seasonal patterns of FOI were measured by the four models. Each group of bars compares the estimates of all the statistical models for a given season, and differences represent the uncertainty in measurements of the FOI with respect to assumptions about clearance of infection. Within one season, estimates are arranged from left to right in order of decreasing goodness of fit of the corresponding survival model. The gamma model estimated the highest FOI, which is consistent with it also estimating the shortest average duration of infection (see Figure 4.2). The overall pattern of seasonality in transmission is consistent across the models.

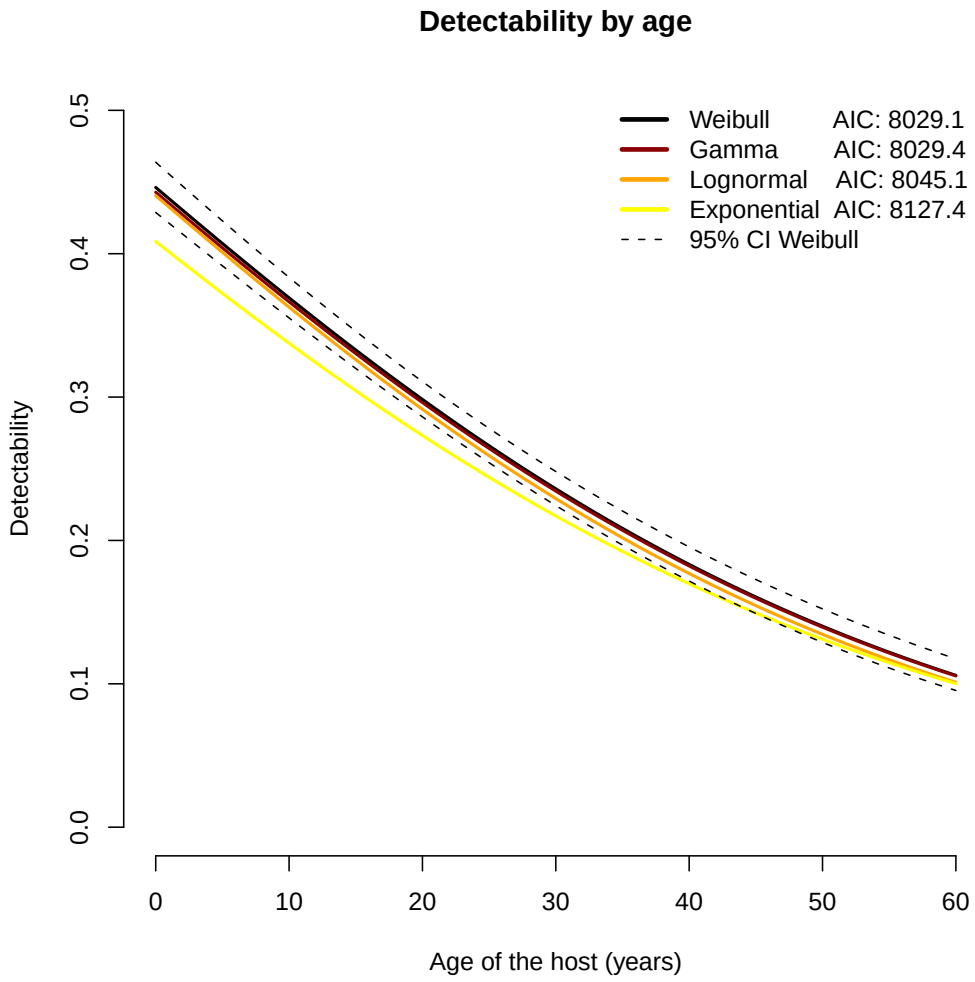


Figure 4.4: **Estimates of detectability** - The obtained estimates of detectability differ only slightly for different assumptions about infection clearance. A detectability below 50% is estimated consistently, and all models agree on a decrease of detectability with host age. The logit-linear relationship of q with age, which is assumed here, does not allow for e.g. a peak in the youngest children. This may lead to inaccurate estimates for these age groups, but since the proportion of infants in the dataset is small, this is unlikely to influence estimates of other parameters.

presumably reflects the simple notion that a given number of detected infections can be explained by either higher FOI and shorter duration, or the other way round. Detectability estimates seem to largely agree across the different survival models, with only the exponential model estimating a slightly lower overall detectability (Figure 4.4).

4.5.2 Validation

The successful recovery of parameters from the simulated datasets suggests that our statistical approach in conjunction with the study design in principle allows for correct identification of the distributions of infection duration. As a caveat, the simulated data is certainly idealized compared to a real dataset: there is no seasonal or spatial heterogeneity in the FOI, and no inter-individual variation in parameters concerning clearance and detection of infections.
Simulated datasets can also be used to establish optimal study designs for measuring a certain quantity of interest. In this case, we could confirm that the study design is suitable for identification of the distribution of infection durations - given it is similar to the published measurements from malariatherapy data [49]. The high discrepancy of the two best fitting models (Weibull and gamma) concerning duration- and FOI estimates from the Ghanaian dataset may be connected to the study design: the differences between the two models fall largely into an interval of infection durations which is shorter than the survey interval of 60 days. The bulk of infections can therefore not be observed very well with the current study design, and shorter survey intervals seem necessary to obtain more accurate measurements. According to the conclusions of chapter 2, there is a lower limit of one week for survey intervals. This has to do with the fact that our method assumes statistical independence of detections at different survey rounds. Due to the complicated temporal behaviour of *P. falciparum* detectability at short time scales, we suggest a survey interval of 2 weeks or more for future, similar studies.

4.5.3 The difference to malariatherapy data

Estimates from the Ghanaian dataset suggested the presence of a large proportion of short infections, which is in dissagreement with previous measurements from malariatherapy data. Hypotheses on the cause of this difference can broadly be divided into two groups: hypotheses attributing the differences to the distinct immune status of the malariatherapy and Ghanaian study populations, respectively, and other hypotheses. Here, we will shortly line out a range of possible explanations.
Acquired immunity could make it harder for the parasite to permanently colonize a host which was exposed before. Since the FOI in the Ghanaian dataset is very high, the population can be considered immune on average, while the malariatherapy patients are immunologically naïve. This hypothesis requires the effects of immunity to last and thus to accumulate over time. Because the age of a human host can be used as a proxy for exposure, an increase of such long-lasting immunity effects with age is expected, which should should lead to more short infections in older individuals. Whether this is indeed the case can be tested by comparing host age groups using the present analysis method (chapter 5).

Short-lived effects of immunity, on the other hand, may include the interaction of concurrent infections within one host: the host populations from the two datasets also differ with respect to their MOI, as there are only single infections in the malariatherapy patients. An effect of interactions between concurrent infections, mediated by short-term effects of immunity, might be confirmed by observing a change of the distribution of infection durations with MOI.
Various factors not related to immunity also have the potential to explain the observed distribution of infection durations. Among these are the following: heterogeneous (unreported) treatment in the population could clear all infections early in some individuals, who mostly treat their infections, and let infections persist in another subpopulation, which rarely treats their infections. Averaged over the study population, this should convey a picture which is consistent with the results of this analysis. This explanation appears, however, unlikely when comparing the number of treatments sold by local health centers to estimates of the expected number of episodes in the area. An alternative explanation attributes the difference of distribution estimates to genetic differences between malariatherapy strains and wild type strains in the Navrongo area. It seems plausible that doctors treating syphilis patients with *P. falciparum* would not favour strains which are cleared after a very short time, requiring a re-infection of the patient. Yet, natural selection may well be doing the same, as a shorter infection duration reduces the R_0 of a strain. Other possibilities include genetic differences between the Ghanaian population and the malariatherapy patients, perhaps with respect to mutations protective against malaria, or an interaction of syphilis with malaria in the patients. In addition, differences in infective dose, or the route of infection may play a role, as some of malariatherapy patients were infected using sporozoites, either through mosquito bites or via subcutaneous injection, and others through infected blood [49, 98].

4.5.4 Limitations of the method

The application of our statistical method to data requires, for now, the assumption that re-infection with the same genetic marker is a rare event. This assumption has been discussed before [56]. However, it gains special significance in cohort studies of long duration: if an immigration-death model is used, what matters is not only the probability to find in a host more than one infecting clone with identical marker genotype at any one time, as considered in [56], but the probability that an individual experiences more than one infection with the same genotype within the study period. The latter must depend on marker diversity, the force of infection as well as on the study duration. For practical purposes, the validity of the assumption can be tested for a given dataset by successively removing the most frequent marker allele from the analysis and observing a possible change in parameter estimates.

4.6 Conclusions

The estimated distribution of *P. falciparum* infection durations in exposed individuals in northern Ghana is different from the distribution in malariatherapy infections [49]. This difference is mainly in the shape of the distributions: in the Ghanaian population, many

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

4.7 Appendix

The relationship between the method described above and previously published methods, most exhaustively explained in [57], is not immediately apparent. We illustrate the mathematical relationship of the presented heuristics with standard approaches of modelling immigration-death processes briefly, and show that the two approaches lead to equivalent expressions.
Rather than calculating the expected frequencies of true pattern types, we consider for the purpose of this illustration the simpler problem of calculating the number of infections present at any time point x, using t as variable of integration. Acquisition of infections at a rate $\lambda(t)$ is assumed to occur within the time interval $[0, x]$. Survival of infections is modeled using parametric survival distributions. These appear in form of the hazard $h(a)$, which for every infection depends on its current age a. The hazard is defined as

$$h(a) = -\frac{S'(a)}{S(a)}, \tag{4.4}$$

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

4.7.1 Exponential survival of infections

We consider first the special case of exponential survival, where the age-independent hazard is often called clearance rate and denoted by μ. In analogy to Equation (4.2) we write the number of infections $n(x)$ at time x as

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

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

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

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

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

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

$$\frac{dn(x)}{dt} = \lambda - \mu n(x) \ , \tag{4.5}$$

with $n(0) = 0$, which constitutes a simple model for superinfection and is explained in [99]. In fact, it was this model of superinfection in connection with the CDF of the exponential distribution which allowed [57] to work out all expected true pattern frequencies. Our approach to calculating these frequencies is therefore equivalent in the case of exponential survival of infections.

4.7.2 Non-exponential survival of infections

In the general case, a model for the age structure of the parasite population within a host is required. Such a model is given by the McKendrick - von Foerster equation [100], a partial differential equation (**PDE**) of the form

$$\frac{\partial u(a,x)}{\partial x} + \frac{\partial u(a,x)}{\partial a} = -h(a)u(a,x) \ , \tag{4.6}$$

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

$$n(x) = \int_0^x u(a,x)da \ , \tag{4.7}$$

the integral of $u(a,x)$ over all existing ages. Equation (4.6) can be solved using the "method of lines", which yields

$$u(a,x) = \lambda(x-a)S(a) \ . \tag{4.8}$$

By inserting the solution for u into Equation (4.7) we obtain the cumulative number of infections of all ages present at a time point x as

$$n(x) = \int_0^x \lambda(x-a)S(a)da \ . \tag{4.9}$$

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

$$n(x) = \int_0^x \lambda(t)S(x-t)\,dt \ . \tag{4.10}$$

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

This expression can also be obtained from Equation (4.3), as the special case when $\alpha = 0$, $\beta = \gamma = x$ and $\delta \to \infty$ [5]:

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

The approach described in this paper therefore represents an extension of the approach by [57], making it possible to use non-exponential survival distributions in models of superinfection.

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

CHAPTER 5

Effects of host age on clearance of malaria infections

Authors: Michael T Bretscher, Nicolas Maire, Nakul Chitnis, Ingrid Felger, Seth Owusu-Agyei, Tom Smith

5.1 Abstract

Background: Acquired immunity against *P. falciparum* is unusual in that it does not prevent re-infection. The parasite thrives even in highly exposed hosts, and superinfection with several clones at once is common. One of the key immune-evasion mechanisms allowing this is thought to be clonal antigenic variation, whereby a given infecting clone changes its dominant antigens in the course of the infection. Clone-specific immunity, developed during the course of an infection, may affect older infections, and clone-transcending immunity, acquired over a lifetime of exposure, may allow older hosts to better control new infections. Analysis of these aspects of immunity requires genotyping data from a cohort study including older, highly exposed individuals. Statistical analyses need to allow for the imperfect detection of *P. falciparum* clones.

Methods: Building on previous work, we use an immigration-death model to re-analyse genotyping data from a one-year cohort study in Navrongo, Northern Ghana. Our statistical analysis method simultaneously estimates force of infection, detectability, and duration of infections. Clearance of infections is modeled using the exponential, Weibull, gamma, and lognormal distributions. By comparing the results from these parametric survival distributions, robust inference about infection clearance is achieved. The effect of host age, and by implication, cumulative exposure, on these estimates offers insights into the effects of acquired immunity.

Results and Conclusions: The average duration of infection was short in children below 5 years, peaked in 5-9 year olds, and subsequently decreased in the older age groups, suggesting a moderate effect of cumulative exposure on infection duration. The Weibull model of infection clearance fitted the data best. An large proportion of *P. falciparum* infections in natural populations appeared to be cleared soon after inoculation, while some infections persisted significantly longer. This was true for all host ages, and therefore cumulative exposure cannot explain the discrepancy between our results and previous distribution estimates from malariatherapy data. The time until the last infections are cleared appeared to be longer in the field than in malariatherapy infections, which has implications for timelines in *P. falciparum* elimination programs.

5.2 Background

With most pathogens, acquired immunity clears the infection and prevents new infections with the same agent. *P. falciparum* parasites are unusual in that they survive and multiply even in highly exposed hosts, who can also be superinfected, both concurrently and sequentially. One of the key immune-evasion mechanisms allowing this is thought to be clonal antigenic variation, whereby a given infecting clone changes its dominant antigens in the course of the infection [101].

A distinction can be made between strain specific immunity, which is accumulated by hosts during the course of an infection, and strain-transcending immunity, accumulated over a lifetime of exposure, which allows older hosts to better control or clear new infections [61, 102]. Strain specific immunity should increase with the age of an infection, and perhaps cause "senescence" of infections, meaning that the clearance rate is higher for older infections. Similarly, strain-transcending immunity should increase with host age, as the cumulative number of infections somebody has experienced increases with age, and should presumably lead to faster clearance of infections in older individuals.

Empirical studies of the effect of acquired immunity on duration can only be carried out in endemic populations since they must include immune individuals, using age as a proxy for cumulative exposure. Such analyses can therefore only be carried out on data from contexts where there are many superinfections. Genotyping technologies are consequently required to track individual infections longitudinally. A change in the duration of clonal infections with age of the human host can then be interpreted as effect of acquired immunity, with the possibility of confounding by other measures which are correlated with age: the multiplicity of infection (**MOI**), or physiological changes of the human body with age. We therefore use the word "cumulative exposure" where it is more accurate.

In empirical studies, *P. falciparum* parasites are often present at low densities, with a small chance of being detected (Chapter 2). Even with high-resolution genotyping of repeated blood samples from the same host, it is challenging to assess the persistence of infections in such contexts. Imperfect detection of the parasite biases estimates of duration by obscuring the start and end-points of clonal infections. Only a few studies have assessed the persistence of individual clones using statistical models that allow for this imperfect detection [37, 57, 58, 60]. Sama *et al* [37, 57] analysed the infection durations of *falciparum* clones in a highly exposed population in northern Ghana. These analyses involved fitting of a deterministic immigration-death model of infection dynamics to "merozoite surface protein 2" (**msp2**) data from a one-year cohort study. Under the assumption that clearance is independent of the age of an infection - which implies an exponential distribution of infection durations - Sama *et al* compared twelve different parameterizations of their model with respect to goodness of fit. One of the main findings was that the detectability of clonal infections declined with host age, while the duration of infection was only slightly affected by cumulative exposure. In addition, it was found that seasonality in transmission is an important aspect of the dataset that was used.

We subsequently perfomed a similar analysis (Chapter 4) of the same study, which allowed for non-exponential distributions of infection duration, thereby dropping the assumption that clearance is independent of the age of an infection. This was achieved by use of

four different paremetric survival distributions to model clearance of clones: the Weibull, lognormal, gamma, and exponential distributions. These survival distributions have the flexibility to model clearance in response to the age of an infection, by including a "shape" parameter. Each distribution has it's own constraints concerning possible shapes.
The estimated distributions of infection duration showed a positive skew, meaning that most infections are cleared soon after inoculation. This is different from estimates obtained using malariatherapy data [49]. Malariatherapy is the treatment of neurosyphilis by intentional infection with *Plasmodium*, which was the method of choice before suitable antibiotics became available. Data obtained from malariatherapy has the advantage that the time points of inoculation and - approximately - clearance of infections are known. An analysis of such data by [49] suggested an average duration of approximately 200 days, and that infection durations much shorter or longer than the mean are rare.
Possible explanations for the difference between our estimates and malariatherapy were brought forward in chapter 4. some of these attribute early clearance of infections to the higher immunity status of the Ghanaian study population. Here we test this hypothesis by allowing the parameters of the survival distributions in to vary with host age.

5.3 Methods

5.3.1 Study design and sample collection

A one year longitudinal study of malaria infection was conducted in the Kassena-Nankana district (KND), in the upper East region of Ghana [37, 52, 57, 63]. The malariological situation in this area is characterized by very high prevalence and multiplicity of infection [63,64], and year-round transmission with seasonal variation in transmission intensity [37]. A total of 349 individuals of all ages were followed up over one year in 2-monthly intervals. Blood was collected on ISOCodeStix^{TM} PCR template preparation dipsticks (Schleicher & Schuell, Dassel, Germany).

5.3.2 Genotyping

DNA was eluted from ISOCodeStix^{TM} Stix and screened for presence of *P. falciparum* by polymerase chain reaction (**PCR**). Processing of stix and PCR conditions have been described in detail before [65]. In brief, samples that tested positive for presence of *P. falciparum* were subjected to PCR using primers specific for the msp2 locus. Different alleles were distinguished on the basis of length polymorphisms by means of automated capillary electrophoresis technology. The obtained data files were further processed using the GeneMapper® software and an in-house generated software, which facilitates identification of known alleles from the raw output of GeneMapper® and transforms the data into different formats suitable for data management and statistical analysis.

5.3.3 Data preparation

Only data of those participants who were present at all survey rounds were included in the analysis. This reduced the number of individuals in the data set to 216. Failure or success to

detect a strain was denoted by 0 or 1, respectively. The 63 possible pattern types containing at least one positive test result were numbered from 1 to 63 using their binary value (e.g., 000010 is pattern 2), and counted. The possiblility of re-infection of a host with the same genotype was ignored for all modeling analyses. This assumption was justified by the high diversity of msp2 alleles in the population. Thus, for every host, a frequency distribution of binary patterns was obtained. Statistical models were formulated to predict the absolute frequency of each pattern taking into account host age and seasonal transmission.

5.3.4 Models of infection dynamics

Three main processes are determining the frequencies of the different binary patterns in each human host: acquisition, clearance, and detection of infections.
Acquisition of infecting clones is represented by the FOI parameter $\lambda(t)$. Seasonality in transmission was found to be an important aspect of the present dataset [37]. For this reason, $\lambda(t)$ takes into account seasonality: for every two-month season i, a separate parameter λ_i is estimated. The pattern of seasonal transmission is assumed to have repeated since the birth of every host, and the FOI is assumed to be independent of host age (in line with previous models for the present dataset [37]).
Clearance of clonal infections is represented by four different parametric survival distributions: the exponential, Weibull, gamma and lognormal distributions (Table 5.1). These represent competing hypotheses about clearance of clonal infections, and are compared with respect to goodness of fit. The exponential distribution is characterized by a single scale parameter, equal to the mean duration of a clonal infection. It does therefore not have any flexibility to yield detailed information on the properties of the clearance process. Nevertheless, it has been widely used in infectious disease models, mainly due to it's mathematical simplicity [1]. The Weibull, lognormal and gamma distributions are characterized by two parameters. In the following we will refer to those as "scale" and "shape" parameters, ignoring possible distribution-specific names. For each of the following age groups, a separate set of scale and shape parameters was estimated (in years): <5, 5-9, 10-19, 20-39, 40-59, ≥60.
The probability of detecting an infecting clone was found to be dependent on host age [37]. The age dependency of detectability was modeled as a logit linear function,

$$l(a) = q_0 + q_1(a - \bar{a}),$$

where a is the age of a host (in 2-month units), and $\bar{a}$ is the average age in the dataset [2]. The actual detectability of any given host of age a is then given by the inverse logit function:

$$q(a) = \frac{e^{l(a)}}{1 + e^{l(a)}} \,. \tag{5.1}$$

[1] Assuming a constant clearance rate per time implies an exponential distribution of infection durations and makes it possible to ignore the age structure of the infection population within one host.

[2] For the sake of comparing our results with those from [37], $\bar{a}$ was set to 120.72 (in units of 2 months, corresponding to the survey interval). This is the average age in the partial dataset used by [37].

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

Table 5.1: **Parametric survival distributions** - A summary of the parametric survival distributions used in this analysis to model clearance of clonal infections. The exponential distribution is specified by a single scale parameter, denoting the mean duration of infection. All other distributions have increased flexibility due to an additional shape parameter (distribution-specific parameter names are ignored). The survivor function of each distribution is defined as 1-CDF (Cumulative distribution function). The following abbreviations are used in the table: for the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$, for the lower incomplete gamma function $\gamma(s,x) = \int_0^x t^{s-1}\,e^{-t}\,dt$, and for the error function $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

5.3.4.1 Calculating expected pattern frequencies

Firstly, the detection process is ignored in order to determine the frequency distribution of the expected "true pattern" frequencies, representing true infection status. Every true pattern i consists of an uninterrupted sequence of "ones", since detection is assumed perfect. It can therefore be uniquely identified by two time intervals: $[\alpha_{k,i},\beta_i]$ (the interval where infections leading to true pattern i can be acquired), and $[\gamma_i, \delta_i]$ (the interval where infections must be cleared in order to produce true pattern i).

The expected frequency τ_i of true pattern i in host k is then given by the expression

$$\tau_{k,i} = \int_{\alpha_{k,i}}^{\beta_i} \lambda(t) \left[S(\gamma_i - t) - S(\delta_i - t) \right] \, dt, \tag{5.2}$$

where S is the survivor function of the parametric survival distribution in use (Table 5.1). The scale and shape parameters of S depend on the age of the host at time t, as outlined above.

In a second step the detection process is incorporated by means of the age-dependent detectability q. In short, the expected frequencies of all the 63 "observed" patterns j are calculated based on the frequencies of the true patterns i and the probabilities of these giving rise to observed pattern j through detection failure. A binomial model of detection is assumed, implying that detection outcomes at different time points are independent of each other.

A detailed description of the statistical methods was given previously, and all models were implemented in the Java ^{TM} language and fitted to the data by maximum likelihood, as described in chapter 4. The software can be obtained from the authors.

5.4 Results

The prevalence of *P. falciparum* in the dataset is 48% by microscopy and 75% by PCR. The age distribution of PCR positivity showed a peak in the 5-9 year old children with 93% of these children being parasite positive (data not shown). Mean multiplicity measured in the PCR positive samples also peaked in the age group of 5-9 years. Both prevalence and the MOI were lowest in the 60+ age group. A more comprehensive epidemiological description of the dataset is given in chapter 3. A total of 103 different msp2 genotypes were found. The most frequent genotype represented 10.2% of all fragments detected.

The four parametric survival distributions as models for clearance of clonal infenctions were compared using Akaike's Information Criterion (AIC) as a measure of goodness of fit (Tables 5.3 and 5.4): the Weibull model fitted the data best (AIC: 7908.5), followed by the gamma (AIC: 7944.8), lognormal (AIC: 7978.3), and exponential (AIC: 8022.6) models. These results can be compared to the results in chapter 4 because the same dataset and methodology were used.

Survival model	FOI by season (person^{-1} year^{-1})						Detectability		AIC
	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	q_0	q_1 (2-months^{-1})	
Weibull	74.5	41.6	36.3	13.2	22.7	56.9	-0.72	-0.003727	7908.5
Gamma	169.1	93.5	89.2	9.1	57.1	126.3	-0.84	-0.004692	7944.8
Lognormal	41.7	26.0	21.6	7.5	15.1	38.8	-0.89	-0.004985	7978.3
Exponential	45.1	19.3	18.1	7.4	12.3	39.8	-0.84	-0.002865	8022.6

Table 5.2: **Force of infection and detectability** - A separate FOI parameter ($\lambda_1, \lambda_2 \ldots \lambda_6$) was estimated by the statistical models for each of the six 2-month-seasons. The numerical value of detectability can be calculated for any host of given age using equation (5.1) and parameters q_0 and q_1.

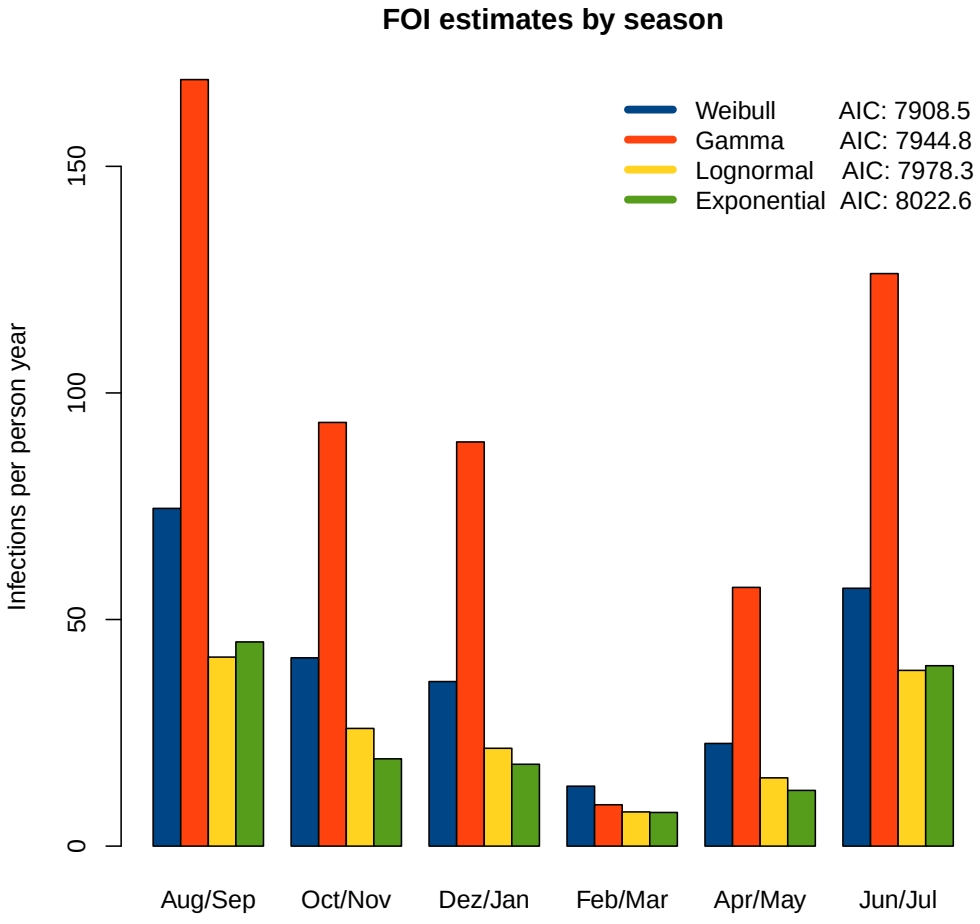


Figure 5.1: **The force of infection** - Each group of bars shows the FOI estimates across the four models for the corresponding season.

Survival model	Scale						Shape						AIC
	r_1	r_2	r_3	r_4	r_5	r_6	s_1	s_2	s_3	s_4	s_5	s_6	
Weibull	1.00	1.31	1.01	0.75	0.94	1.23	0.49	0.47	0.67	0.58	0.66	0.75	7908.5
Gamma	9.02	12.37	6.24	6.64	7.06	8.44	0.10	0.10	0.12	0.09	0.10	0.12	7944.8
Lognormal	1.68	1.86	1.54	1.51	1.62	1.71	1.32	1.37	1.06	1.06	1.18	1.24	7978.3
Exponential	3.73	5.30	2.93	2.15	2.10	2.18	-	-	-	-	-	-	8022.6

Table 5.3: **Parameters related to clearance of infections** - Scale and shape parameters are shown for all models and host age groups. The exponential distribution is defined by a single "scale" parameter per age group, while all other distributions require an additional "shape" parameter. Derived properties of the estimated survival distributions can be obtained from these parameters in conjunction with Table 5.1. The survey interval of two months was used as time unit where applicable. Lower AIC values indicate a better fit to the data.

5.4.1 Force of infection

The overall pattern of seasonality is consistent among all models of infection survival, but differences exist in the numerical estimates of the FOI (Figure 5.1 and Table 5.2). Estimates of the Weibull model range from 74.5 infections per person year in August and September to 13.2 infections in February and March, while the corresponding results from the exponential model are 45.1 and 7.4, respectively. The gamma model estimates a considerably higher FOI of 169.1 in the high-transmission season, but the FOI estimate for the low-transmission season (9.1) agrees well with the other models.

5.4.2 Clearance of infections

The probability density functions (**PDF**) estimated by the best-fitting Weibull model are compared to the estimates from the exponential model, which gave the worst fit to the data (Figure 5.2). The distribution of infection durations under the Weibull model is positively skewed in all age groups, compared to the exponential estimates. This is the case for all survival models (not shown), and most extremely for the gamma estimates. A positive skew of distributions indicates that a significant proportion of infections is cleared rather soon after inoculation. The PDF of an exponential distribution depends by definition entirely on the mean duration and does not have the flexibility to measure the higher moments of the distribution in the data. The corresponding CDF's of distribution estimates are shown in Figure 5.6

The raw estimates of scale and shape parameters provide limited information on properties of interest. Mean and standard deviations and their dependence on host age are therefore reported. These can be calculated from the scale and shape parameters (Table 5.3) and the distribution-specific expressions for the mean and variance (Table 5.1).

All models suggest an intermediate mean duration of infection in the youngest age group, a peak a in the 5-9 year olds and a subsequent decrease of duration with increasing host age (cumulative exposure). The average durations across all age groups are (in days): 102.7 (Weibull), 52.0 (gamma), 210.4 (lognormal), and 184.0 (exponential). Results for all age

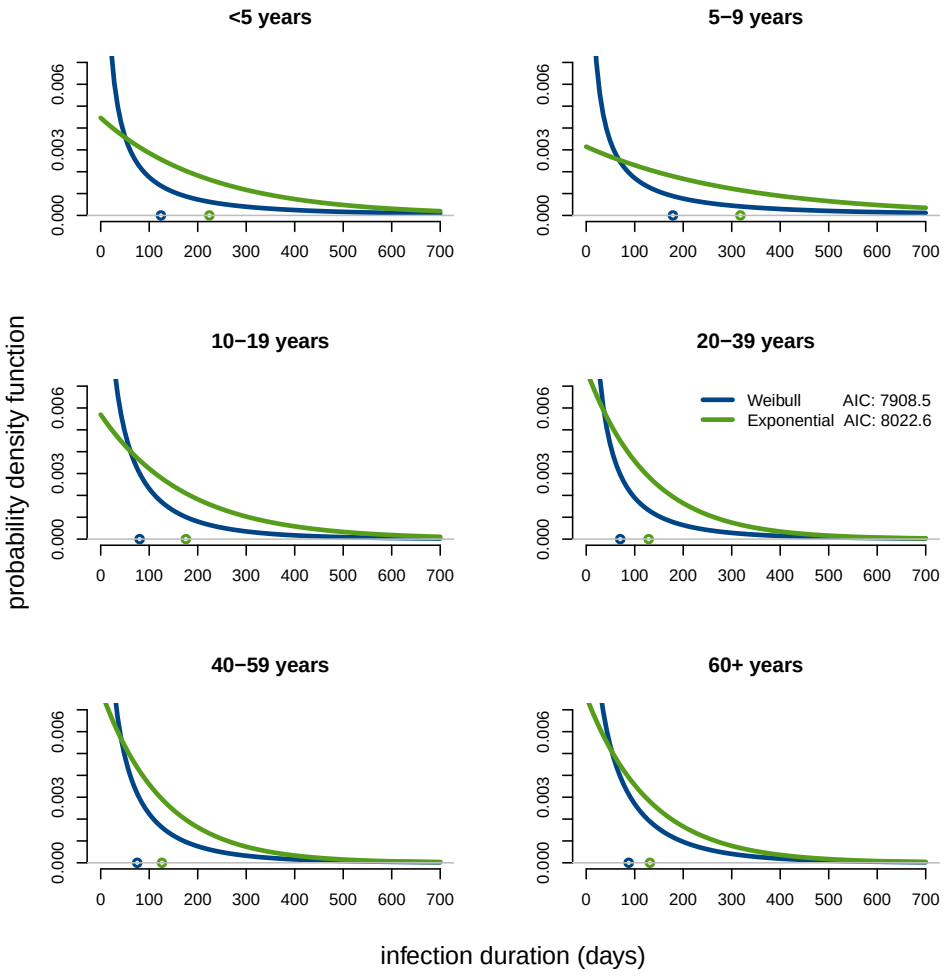


Figure 5.2: **Clearance of infections** - The PDF's of the best-fitting Weibull model and the exponential model by age groups. Mean durations are indicated by circles on the abscissa.

Survival model	< 5yo	5-9yo	10-19yo	20-39yo	40-59yo	> 60yo	AIC
Weibull	124.2	178.8	80.2	70.2	75.3	87.7	7908.5
Gamma	54.6	77.4	43.1	36.8	40.6	59.7	7944.8
Lognormal	241.9	285.5	161.7	158.4	194.6	220.4	7978.3
Exponential	224.1	317.9	175.6	128.8	126.3	131.4	8022.6

Table 5.4: **Mean durations** - The mean duration of infection (in days) is indicated for all age groups under different assumptions concerning infection clearance. It can be calculated from the parameter estimates (Table 5.3) and the distribution-specific expressions for the mean (Table 5.1). Lower AIC values indicate a better fit to the data.

groups separately are reported in Table 5.4 and Figure 5.3.

Estimates of the standard deviation of infection duration consistently show less variation in the non-exponential survival models than estimated under the exponential model (Figure 5.4). In an exponential distribution, the standard deviation is always equal to the mean. All models show a peak in the standard deviation in the 5-9 year olds, very similar to the estimates of the mean duration.

The mean residual lifetime (**MRL**) is the additional time an infection is expected to last [103]. This quantity may change with the age of an infection, depending on the properties of the distribution of infection durations. It can be calculated from the shape and scale parameters as $\frac{1}{S(\xi)} \int_0^\infty t\, f(t+\xi)\, dt$ where ξ is the current age of an infection and S and f are the survivor function and PDF, respectively, of the parametric survival function in use. MRL is plotted against age of infection in Figure 5.5. Only estimates from the best-fitting Weibull model and the worst-fitting exponential model are shown. The MRLs of the exponential estimates appear as horizontal straight lines because under an exponential model, clearance is independent of infection age. Weibull estimates, conversely, show an increase of MRL with the age of an infection. This effect is strongest in the first days after inoculation, and in the younger host ages.

The question of how much time is needed until all or almost all infections in a human population are cleared, in absence of transmission, is of relevance for the planning of malaria elimination, although other factors need to be considered as well. In line with [104], who develop a theory of elimination, the times until 99% of infections are cleared is shown across host ages (Figure 5.7). Times to near-elimination show a peak in the 5-9 year olds, and there appears to be considerable variation of estimates depending on which survival distribution is used.

5.4.3 Detectability

Measurements of detectability are in good quantitative agreement (Figure 5.8). All models measure a detectability of ca. 40% in the young ages, which decreases to ca 20% in the older age groups. However, measurements in the very young may be biased, as the logit-linear curve in use to model the relationship of detectability with age did not allow for an expected "peak" of detectability in those ages.

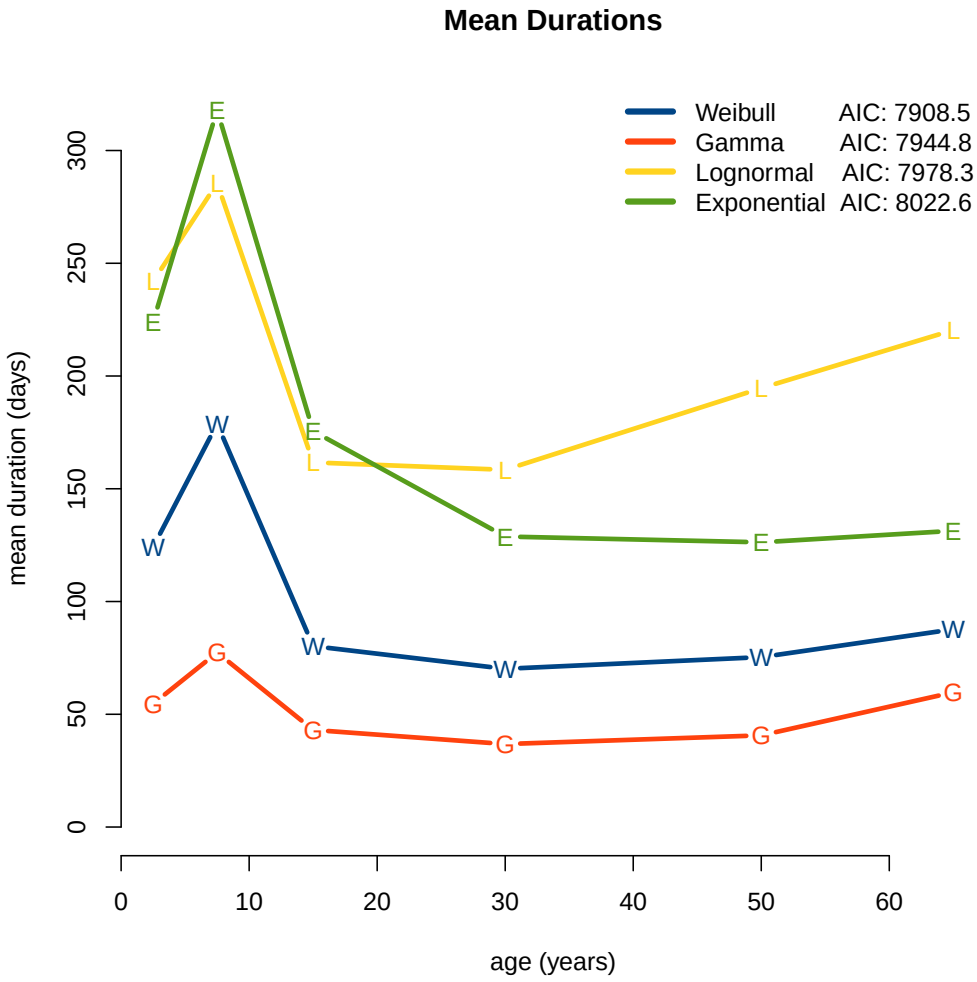


Figure 5.3: **Average duration of infection** - The mean duration of a clonal infection is plotted against the midpoint of each age group, for all four models of infection survival.

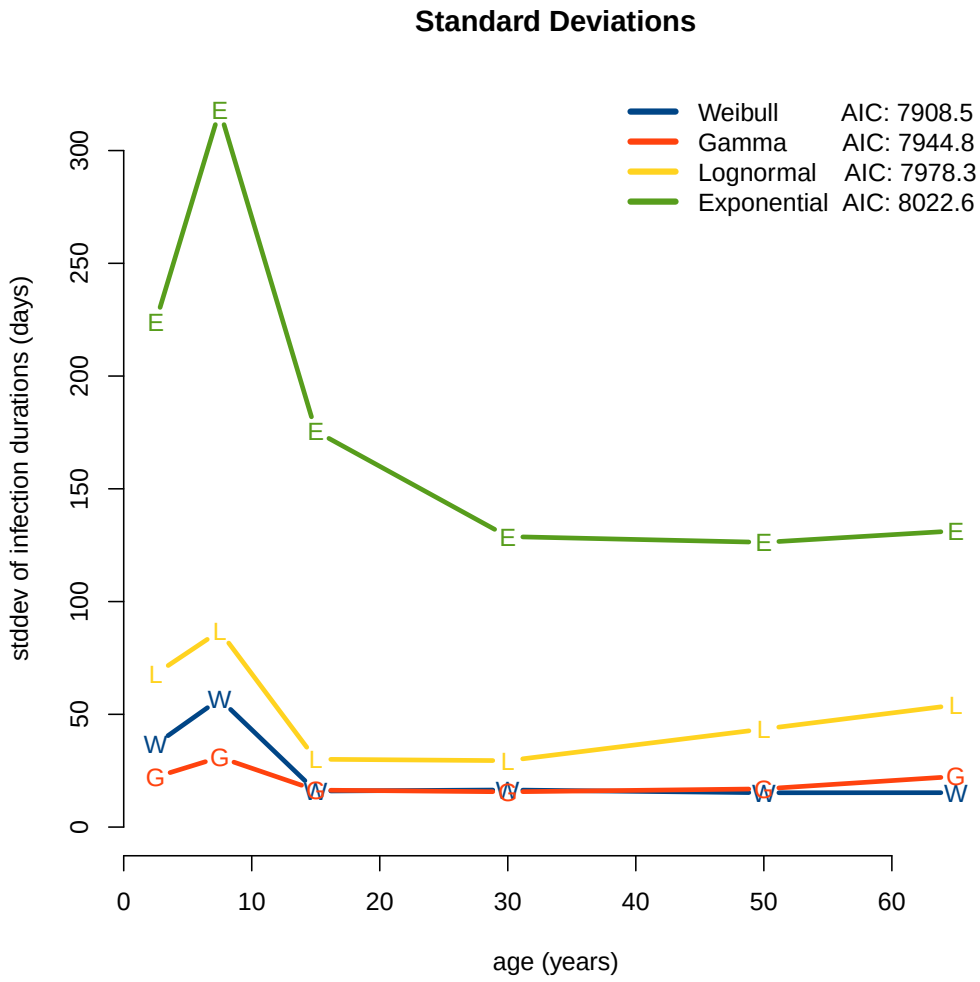


Figure 5.4: **Variation in the duration of infection** - The standard deviation of infection durations is plotted against the midpoint of each age group. While in an exponential distribution, the standard deviation is always equal to the mean, the other distributions have greater flexibility.

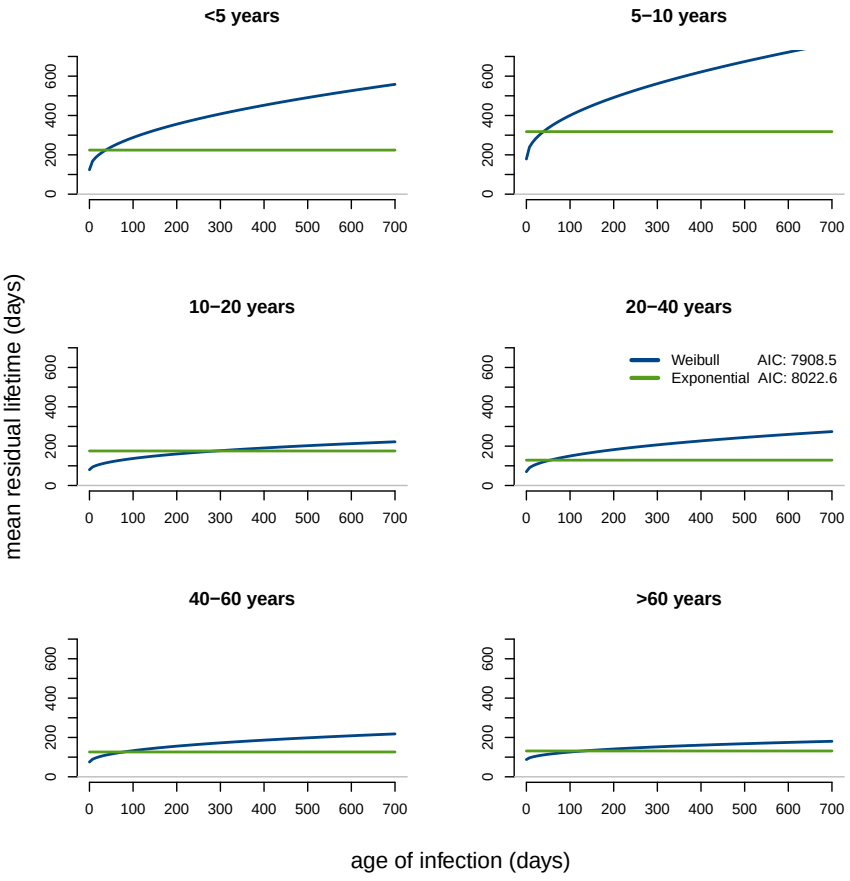


Figure 5.5: **Mean residual lifetime** - The mean residual lifetime (MRL) is the additional time an infection is expected to last. It depends on the current age of an infection and the distribution of infection durations. MRL is plotted against age of infection separately for each host age group.

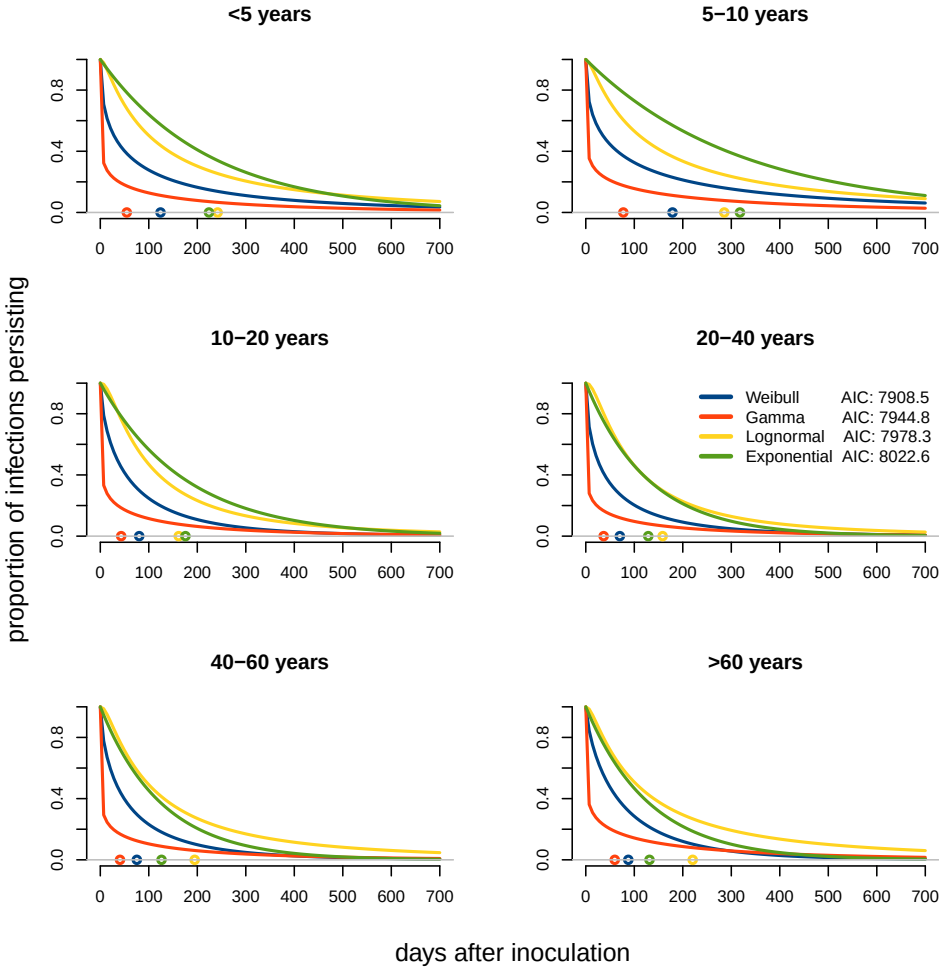


Figure 5.6: **Persistence of infections** - The proportion of clonal infections surviving is plotted against time after inoculation, for different models of infection clearance and across host age groups. Mean infection durations are indicated as circles on the abscissa.

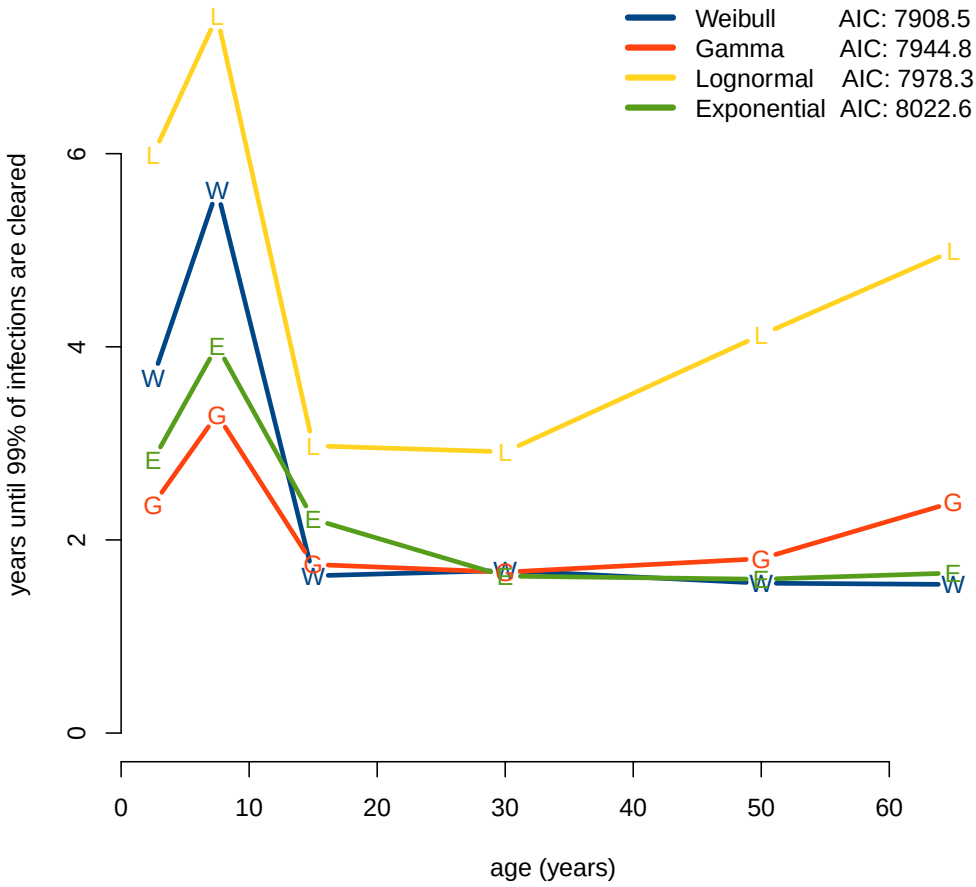


Figure 5.7: **Time until most infections are cleared** - The time is shown until 99% of infections are cleared after complete interruption of transmission. All models agree that the time required is longest in the age group of 5-9 years.

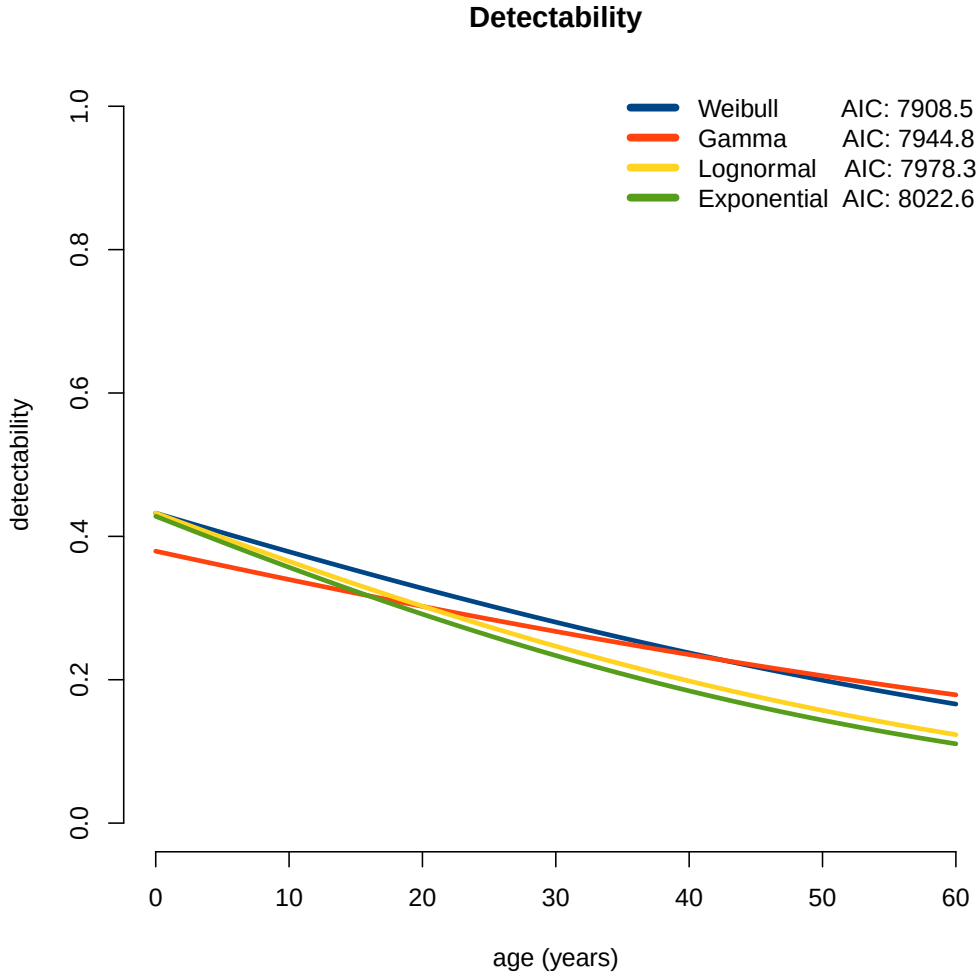


Figure 5.8: **Detectability** - Detectability estimates across the different clearance models show almost no difference. The numerical values of detectability range from ca. 40 % in the younger ages to ca. 20 % in older hosts. Modeling the age depencence of detectabilty as a logit-linear relationship (which cannot represent peaks) may bias detectability estimates in the very young host ages.

5.5 Discussion

The introduction of separate scale and shape parameters for every age group improved model fit considerably in comparison to the age-independent models of chapter 4. This indicates that changes in infection clearance with host age are important. The ranking of the different survival models with respect to goodness of fit was not affected: in both cases the Weibull model gave the best fit to the data, followed by the gamma, lognormal and exponential models[3]. A major difference is that in the present analysis the model ranking is very clear, while in chapter 4 the Weibull and gamma model showed a very similar goodness of fit.
The significance of host age lies in its correlation with cumulative exposure to infection. If a change of certain parameters with host age is observed, it may be interpreted as an effect of acquired immunity. This is only partly correct, since confounding by other age-dependent factors may occur, such as by MOI. We therefore prefer the term "cumulative exposure" instead of "immunity". The absence of change in parameter values or derived measures with host age, conversely, may well be interpreted as "no effect of immunity" - ignoring the unlikely possibility that a confounder might exactly balance the effects of acquired immunity. By fitting models with infection clearance depending on host age we are now able to better understand the findings presented in chapter 4, and it possible to rule out some of the hypotheses given in chapter 4, which aimed at an explanation of the observed differences in infection clearance between a dataset from malariatherapy patients [37] and the present dataset.

5.5.1 Force of infection

FOI estimates are practically identical to previously reported results (chapter 4), using models which did not estimate separate clearance parameters for each age group. All models broadly agree with respect to the pattern of seasonality (Figure 5.1). The Weibull model estimates on average a slightly higher FOI than the exponential model, indicating that previous analyses may have underestimated the FOI. The most extreme estimates are by the gamma model, which also estimated a very short infection duration (Figure 5.3). A major difference to chapter 4 is, as mentioned above, that the ranking of models is much clearer in the present analysis. This considerably increases our confidence in the numerical estimates of the FOI obtained from the Weibull model. Several potential properties of the FOI, however, remain unexplored in the present analysis. Firstly, the dataset is expected to comprise a wide range of transmission intensities. Secondly, it is assumed that the FOI does not depend on the age of the host. This neglects both effects of body size and acquired immunity on the FOI. It is unclear what the effects of not accounting for age-dependence of FOI are. Because the number of clones present at any one time should increase with both FOI and mean duration of infections, it is expected that not accounting for age-dependence of the FOI may bias estimates of the mean duration. Future analyses need to explicitly

[3]The AIC values in the age-independent models of chapter 4 were: 8029.1 (Weibull), 8029.4 (gamma), 8045.1 (lognormal), 8127.4 (exponential). These are comparable to the present analysis because the same dataset was used.

model age-dependence of the FOI.

5.5.2 Clearance of infections

A positive skewness of all distribution estimates across all age groups emerged. The Weibull PDF, shown in Figure 5.2, illustrates this. It means that most infections are cleared soon after inoculation, with the PDF flattening out at older ages of infection. In fact, only a fraction of those infections shorter than the survey interval of two months can be picked up with the current study design, even if detectability were high. It is therefore desirable to have shorter survey intervals in future, similar studies, in order to collect more information on these short infections. The picture of infection clearance is radically different from estimates obtained from malariatherapy patients [49], where most infections were cleared after an intermediate amount of time. The distribution of infection durations in the malariatherapy patients was approximately symmetrical about the mean, with relatively few infections cleared soon after inoculation.

The absence of a change of skewness with host age in the present dataset indicates that cumulative exposure is not the likely cause for the early clearance of many infections. This was proposed in chapter 4 as a possible explanation for the observed differences between malariatherapy data and the the present Ghanaian dataset, because malariatherapy patients were immunologically naïve. But no effect of cumulative exposure does not necessarily mean no effect of immunity. It is imaginable that a short-lived effect of the frequent exposure to *P. falciparum* antigens could raise overall immune activity, which would in turn make it hard for newly arriving infection clones to survive the critical phases of host entry and establishment of a stable blood stage infection. A short term effect of exposure would mean that the distribution of infection durations should change with the FOI. This could perhaps be tested on the present dataset, as it presumably contains considerable heterogeneity in the FOI. A very similar but distinct hypothesis is that the presence of infections in a host would make it harder for newly arriving infections to last.

Genetic differences between the wildtype strains and those used in malariatherapy represent another possible explanation. It is known that the malariatherapy strains were selected for desired properties with respect to the treatment of neurosyphilis [94], and in fact, Collins and Jeffery [35] report shorter infection durations of the "McLendon" strain compared to the "El Limon" or "Santee-Cooper" strains. Even without doctors intentionally selecting parasite strains, the serial passage of strains from patient to patient, and the absence of within-host competition through absence of superinfection could arguably have changed the genetic properties of the malaria-therapy strains.

Unknown genetic differences between the two human populations, such as a possible increased prevalence of sickle cell anaemia in the Ghanaians, or a predisposition for contracting syphilis in the patients, may further play a role. Or, there may be an interaction between malaria and syphilis in malariatherapy patients, perhaps through modulation of the immune response.

Mean residual lifetime of clonal infections

The MRL of an infection is the expected additional time the clone will persist. Plotting MRL versus the age of infection is a way of visualizing the properties of the distribution of infection durations. The MRL's under the Weibull and exponential models are compared in (Figure 5.5). Exponential MRL appears as a horizontal line because the clearance rate is independent of infection age. The Weibull MRL, conversely, increases with age of infection, particularly in the early days after inoculation. This effect is strongest in the younger host ages, and decreases with age (cumulative exposure). This can be interpreted as indication that the early stages of a *P. falciparum* blood-stage infection are a vulnerable point in the life cycle of the parasite. The continous increase of MRL with infection age gradually disappears with cumulative exposure. This may be connected to increased acquired immunity or higher MOI. Both should have the effect that more and more antigenic variants are either known to the host from previous exposure, or already "in use" by another infecting clone, which would close the corresponding routes of immune escape. The lognormal and gamma models do not fit consistently into this view when ordered by goodness of fit (not shown).

Mean and standard deviation of infection durations

The Weibull model estimates shorter mean infection durations than the exponential model. This suggests that analyses which assume a constant clearance rate tend to overestimate infection duration. As a general trend, the models with shorter mean estimates also yield higher estimates of the FOI. The most extreme example is the gamma model, which estimates very high FOI and very short infection durations. It's goodness of fit is clearly inferior to the Weibull model, which suggests that the estimates of the latter are more reliable. The gamma model also suggests that a massive proportion of infections are cleared after only a few days. But most of these infections never appear in the data, given the 2-month survey interval of the study. This supports the view that the gamma estimates are unrealistic, and that the extremely short (unobserved) infection durations are balanced by a high FOI, such that the mean number of infections present at any time is consistent with the data.

Infection duration is short in children below 5 years, and peaks in the 5-9 year olds (Figure 5.3). The reason for this is unclear. Physiological differences between age groups might explain this observation, but it is possible that the absence of age-dependence of the FOI, in the statistical models, might lead to biased estimates of the duration. In the present analysis, the FOI is modeled as a function of seasonality alone, but [84] and others found that the FOI usually increases with body size. Explicit modeling of the age dependence of FOI is therefore required to gain futher insight.

The decrease of average infection duration in individuals of 10 years and older is not in line with a main conclusion of [37], who analysed a partial dataset from the present study: Sama *et al* found that assuming clearance rate independent of host age gave the best fit to the data, and concluded that, contrary to common belief, cumulative exposure does not substantially influence infection clearance, but mainly causes a decrease in detectability. This decrease in detectability is able to explain the often observed decrease in prevalence and MOI in older ages. The relationship of infection duration with age, as suggested by

our analysis, could not be captured by the model of Sama *et al*, as it only allowed for monotonous changes of the mean duration with age. However, our estimates indicate that the mean infection duration remains relatively constant after 15 years of age.
The standard deviation of infection durations, as estimated by the non-exponential models, is considerably smaller than suggested by the exponential model (Figure ...). Since the standard deviation of an exponential distribution is always equal to the mean it is likely that exponential models overestimate the variation of infection durations. Interestingly, while the Weibull estimates from the Ghanaian dataset show less variation than the corresponding exponential estimates, they show more variation than the durations in malariatherapy data. This might be a consequence of selection bias in the malariatherapy strains, as doctors may have avoided using strains that only lead to short infections.

5.5.3 Detectability

Detectability estimates largely agree across the different survival models and show a clear decrease of detectability with cumulative exposure (Figure 5.8), in agreement with [37] and chapter 4. Part of this decrease might also be attributed to increased MOI in older ages as well as to increased levels of acquired immunity, which acts to reduce parasite densities.

5.5.4 Time to near-elimination

The discussion about local elimination or global eradication of malaria is again on the table, and the waiting time until the last infection is cleared is of considerable interest in this debate. However, estimates of the expected waiting time until 99% or 99.9% of infections are cleared, obtained from a distribution, do not form a complete theory of elimination or eradication: the size of the human population, the MOI, the intensity of residual transmission, the number of people at risk of infection, the intensity of surveillance through the health system and the rate of immigration of cases from elsewhere need to be considered as well. They distribution of the durations of clonal infections does, however, form part of such a theory, as presented by [104]. Smith *et al* consider "senescence" of infections, meaning that the rate at which infections are cleared (the hazard) increases with their age. Senescence of infections appeared to decrease the time to elimination in those simulations. The opposite, namely that the clearance rate would decrease with the age of infection, should therefore prolong the time needed until near-elimination. The increasing MRL estimated by the Weibull model (Figure 5.5) shows exactly this: the older an infection, the longer it will stay. Anecdotal evidence suggests that *P. falciparum* infections may last up to 8 years [93], which may well be an upper limit. This renders some our estimates of time to near-elimination unrealistic. This is likely because infection durations longer than one year could not be directly observed within the present study but, represent projections outside the data. However, our results show that the choice of the statistical model influences estimates of time to near-elimination, identify the age group of 5-9 year olds as important target group for surveillance of activities during a potential elimination campaign.

5.6 Conclusions

We investigated the properties of infection clearance by use of an immigration-death model with parametric survival distributions for modeling infection clearance. The force of infection, parameters of the clearance process, and detectability were simultaneously estimated from a genotyping dataset from northern Ghana. The present analysis extended previous work by allowing for host-age dependent estimates of parameters describing infection clearance. We found that the Weibull distribution constitutes the best model for infection clearance in the present dataset, as it was clearly superior in terms of goodness of fit compared to the second-best fitting gamma distribution.

A prominent property of the estimated distributions of infection duration is positive skewness, which means that a large proportion of infections does not persist for a very long time, but is rather cleared shortly after inoculation. This differs from analyses of malariatherapy data [49], which suggested that infection durations are rather symmetrically distributed about the mean. Possible causes for this discrepancy include selection of strains with good curative properties by the physicians, interaction of infections in the Ghanaian population, or different genetics of the two study populations.

Because no large differences concerning early clearance were found among host age groups in the Ghanaian population, a different immune status of the individuals in either dataset cannot account for differences in early clearance of infections.

There was an effect of cumulative exposure on other aspects of infection survival: the mean residual lifetime (MRL) of an infection increased with the age of an infection under the best-fitting Weibull model. This effect nearly vanished in the older host ages, suggesting that acquired immunity or higher MOI does make it harder for a clone to establish a lasting infection.

The average infection duration peaked in the age group of 5-9 years, and decreased in older host ages.

An overestimation of the average duration of infection by exponential models was noted. This overestimation of the duration appears to go along with a slight underestimation of the force of infection. There might be datasets where assuming an exponential decay of infections can be a good assumption in order to reduce the number of parameters in the statistical model. Such a situation would occur if the total duration of a study is too short to contain sufficient information on the higher moments of the distribution of infection durations.

It is likely that data with 2-month survey interval does not contain accurate information about the presence or absence of very short infections. This argues for experimentation with different study designs, ideally using a combination of short (ca. 2 weeks) and long time intervals [4].

[4] The present statistical method can be extended to studies with non-equally spaced survey intervals

CHAPTER 6

Are all malaria parasites equal? Human *Plasmodia* compared.

Authors: Cristian Koepfli, Sonja Schöpflin, Michael Bretscher, Enmoore Lin, Benson Kiniboro, Peter A Zimmerman, Thomas A. Smith, Ivo Mueller, Ingrid Felger

Abstract

Background: In endemic areas, most people are simultaneously infected with different parasite clones. Detection of individual clones is hampered when they fluctuate around the detection limit and, in the case of *P. falciparum*, by sequestration during part of their life cycle. This might have important implications for basic measures of epidemiology or for the outcome of clinical trials. This study aimed at measuring the detectability of individual *P. falciparum* and *P. vivax* parasite clones in consecutive samples of the same patient and at investigating the impact of sampling strategies on basic epidemiological measures, such as multiplicity of infection (**MOI**).

Methods: Samples were obtained in a longitudinal field survey in 1-4.5 years old children from Papua New Guinea which were followed up in 2-monthly intervals over 16 months. At each follow-up visit, two consecutive blood samples were collected from each child at intervals of 24 hours. Samples were genotyped for the polymorphic markers msp2 for *P. falciparum* and msp1F3 and MS16 for *P. vivax*. Mean MOI estimated from single samples per host was compared to combined data from sampling twice within 24h.

Findings and Conclusion: Detectability was high in our data set (0.79 for *P. falciparum* and depending on the marker 0.61 to 0.73 for *P. vivax*). The effect of combining results from samples collected over 24 hours on prevalence and mean multiplicity of infection was comparable in *P. falciparum* and *P. vivax*. Differences in parasite densities have likely a greater impact on clone detectability than effects of periodic parasite sequestration. The additional efforts and costs of such a study design do not generally justify short term sampling in epidemiological studies with a main focus on prevalence, however, this sampling strategy is advised when individual clones in low parasitaemia samples shall be detected.

6.1 Background

Many epidemiological health surveys rely on data of the presence of parasites or parasite clones in populations. In recent years the Malaria Atlas Project (**MAP**) has created detailed maps of global malaria risk using parasite prevalence as the main indicator for transmission and estimated global *P. falciparum* disease burden based on a empirical relationship between prevalence of infection and incidence of clinical disease [105–108]. Within populations, the identification of individual clones by PCR based genotyping techniques has substantially increased knowledge of the infection dynamics of malaria by providing precise estimates of multiplicity of infection, incidence and clearance rates [76, 78, 109, 110]. In addition, it allows classifying drug failure in recrudescence and new infection.
However all methods for the detection of malaria infections or individual parasite clones in a blood sample are imperfect. Malaria parasites can remain undetected because parasite densities are below the detection limit of a given diagnostics or in the case of *P. falciparum* because of mature forms sequester in peripheral blood vessels for 24-28 hrs of its 48-hour life cycle. Although sequestration of *P. falciparum* parasites has been reported several decades ago [111, 112], and the sensitivity threshold of both light microscopy and PCR-based diagnostic methods have been well studied [52, 113, 114], their consequences on estimates of prevalence and other epidemiological parameters have so far little been discussed. Also outcome of drug efficacy studies might be compromised. According to the clinical trial guidelines from WHO treatment failure is defined by at least one identical parasite clone observed before and after treatment [115]. Imperfect detection potentially leads to underestimation of treatment failure when a persisting parasite clone is detected only once.
When children were followed-up on a daily basis a highly complex dynamics of *P. falciparum* clones was observed with the composition of infections unstable over time or even changing from one day to another [62]. Consequently, a single blood sample is likely to only partly represent the true parasite population present in a host. Several studies have addressed the issue of imperfect detectability and described mathematical models to estimate infection dynamics of *P. falciparum* under conditions of imperfect detection of parasites [37,57,60]. The model by Sama *et al* [57] was applied for a longitudinal field study in Ghana, where consecutive blood samples were collected from each participant in 2 monthly intervals. From this dataset, it was estimated that 47 per cent of all parasites present in a host are detected by PCR-RFLP in a single blood sample [57].
Sequestration is thought not to occur in *P. vivax*. This view was questioned only very recently when cytoadherence of *P. vivax* was shown in vitro [116]. In addition parasite loads of *P. vivax* are usually lower than those of *P. falciparum* [70, 117]. As a consequence imperfect detection of *P. vivax* is to be expected when *P. vivax* densities are near of below the detection limit of light microscopy or molecular diagnostics. Only few studies have addressed the infection dynamics of individual *P. vivax* clones [76]. Daily fluctuations in detectability of *P. vivax* clones have not yet been investigated. Indication of limits in *P. vivax* detection comes from a report in Brazil [118]. A high proportion of confirmed multiple clone infections were not detected, despite the application of high-resolution markers. This is most likely due to imperfect detection of clones.

Repeated blood sampling after short time intervals might provide an improved picture of the parasite population present in a host. On the other hand it leads to a considerable increase in efforts in the field and laboratory, added costs, and additional discomfort for study participants. The effect of such a sampling scheme on epidemiological measures such as prevalence or multiplicity of infection is little known.
We present here data from a large set of paired samples that were collected 24 hours apart. Within this time interval re-infection with a new parasite clone can be excluded. Samples derived from a cohort study performed in Papua New Guinean children 1-4.5yrs of age living in an area highly endemic for both *P. falciparum* and *P. vivax*. Parasites of both species were genotyped to calculate the detectability of infection and to investigate the benefit of collecting 24h bleeds on basic and molecular measures of epidemiology such as prevalence and multiplicity of infection.

6.2 Methods

6.2.1 Field survey and patients

This study was conducted in a rural area near Maprik, East Sepik Province, Papua New Guinea. A detailed description of the study was given previously [117]. Briefly, 269 study participants were enrolled at an age of one to three years starting in March 2006 and regular follow-up visits were conducted over a period of 16 months until July 2007. At seven time points separated by 8-weekly intervals from each study participant two consecutive blood samples were collected at intervals of 24 hours (in the following termed: 24h bleed). Antimalarial treatment with Coartem® (Novartis, Switzerland) was administered upon a positive rapid diagnostic test or if haemoglobin levels were <7.5 g/dl. Informed consent was sought from all parents or guardians prior to recruitment of each child. Scientific approval and ethical clearance for the study was obtained from the Medical Research and Advisory Committee (MRAC) of the Ministry of Health in PNG and from the Ethikkommission beider Basel in Switzerland.

6.2.2 Laboratory procedures

All finger prick blood samples were separated into plasma and cells. DNA was extracted from cell pellets using QIAamp® 96 DNA Blood Kit (Qiagen, Australia) according to the manufacturer's instructions. All samples were genotyped for the polymorphic marker gene merozoite surface protein 2 (msp2) by use of capillary electrophoresis for fragment sizing as previously described by Falk *et al* [52] with some minor changes and adaptations of PCR conditions for highly purified DNA as described by Schoepflin *et al* [119].
P. vivax genotyping was performed as described previously [120] with the following modifications: A multiplex primary PCR was done with the primers for the 2 markers msp1F3 and MS16 followed by individual nested PCRs for msp1F3 and MS16. Primary PCR was performed in a volume of 20 μl containing 1 μl template DNA, 0.25 μM of each primer (Eurofins MWG Operon), 0.3 mM dNTPs (Solis BioDyne), 2 mM MgCl2, 2 μl Buffer B (Solis BioDyne) and 5 U Taq FIREPol (Solis BioDyne). 0.5 μl primary PCR product was

used as template for nested PCR performed in a volume of 20 μl containing 0.25 μM of each primer (Applied Biosystems), 0.2 mM dNTPs (Solis BioDyne), 2 mM MgCl2, 2 μl Buffer B (Solis BioDyne) and 1.5 U Taq FIREPol (Solis BioDyne). The forward primers for the nested PCR were labelled with fluorescent dyes: 6-FAM for msp1F3, NED for MS16. Cycling conditions were as follows: Initial denaturation 95°C for 1 minute, then 30 cycles (primary PCR) or 25 cycles (nested PCR) with 15 seconds denaturation at 95°C, 30 seconds annealing at 59°C and 30 seconds elongation at 72°C plus a final elongation of 5 minutes at 72°C.

All samples negative after the first round of PCR amplification, were repeated once. Repeats and all microscopy negative samples (due to an expected lower parasitaemia) were done under similar conditions with the exception that 2 μl DNA solution were used as template for the primary PCR, and 1 μl primary PCR product as template for the nested reaction. All samples with high amount of background or stutter peaks were repeated in a 50 μl reaction volume for both the primary and nested reaction, 0.2 μM of each primer, 0.2 mM dNTPs, 5 μl Buffer B, 2 mM MgCl2 and 2.5 U Taq DNA polymerase and 1 μl DNA for the primary PCR and 0.5 μl of primary PCR product for the nested PCR. Capillary electrophoresis was done as described earlier [120]. As the msp1F3 nested PCR in general led to more amplification product compared to the MS16 PCR, twice as much MS16 PCR product was analysed by capillary electrophoresis.

6.2.3 Data analysis

Analysis of 24h interval bleeds. Sample pairs collected 24 hours apart from the same patient were compared. Sample pairs were excluded from the analysis if antimalarial treatment was given on the first day of a paired sampling. Individual genotypes were classified by positivity on each of two consecutive days, leading to two categories for each genotype: one day positive (genotype observed on either day of paired sampling, n1) and both days positive (n2). An estimate of the detectability q was calculated as follows:

$$q = \frac{2n_2}{n_1 + 2n_2} ,$$

according to [82]. An approximate confidence interval was calculated as follows: CI $[q \pm 1.96\ \mathrm{SE}(q)]$, where the standard error is:

$$SE(q) = \frac{2\sqrt{n_1 n_2 (n_1 + n_2)}}{(n_1 + 2n_2)^2} .$$

The detectability was calculated for different age groups of patients. Comparison of detectability between day 1 and day 2 was done by McNemar's exact test for paired data.

All statistical analysis was performed using STATA® 9.1 statistical analysis software (Stata Corporation, College Station, TX).

	P. falciparum	P. vivax
No. of positive samples		
only 1st day positive	39 (16.1%)	57 (10.1%)
only 2nd day positive	47 (19.4%)	68 (12.1%)
Both days positive	156 (64.5%)	438 (77.8%)
Total No.of positive pairs	242	563
Prevalence		
Prevalence on 1st day	0.19	0.49
Prevalence 2 days combined	0.24	0.55
Detectability q incl. CI	0.78 [0.76 – 0.81]	0.88 [0.86 – 0.89]

Table 6.1: **Effect of repeated sampling on prevalence as determined by microscopy**

6.3 Results

6.3.1 Effect of repeated sampling on prevalence

For the analysis of paired samples collected in a 24h interval a total of 1019 pairs were eligible. By light microscopy 242 and 563 pairs were positive *P. falciparum* and *P. vivax*, respectively, at least on either day (Table 6.1). Prevalence did not differ between individual days but it increased when both days were combined. *P. falciparum* prevalence was 19.1% at day 1 and 20.0% at day 2 (McNemar's test: χ^2=0.65, p=0.42). When both days were combined prevalence increased to 23.7%. Prevalence of *P. vivax* was 48.6% on day 1 and 49.7% at day 2 (McNemar's test: χ^2=0.88, p=0.35), and 55.3% when both days were combined.
The 1019 sample pairs were genotyped using msp2 as *P. falciparum* marker and msp1F3 and MS16 as *P. vivax* markers. After PCR the number of pairs positive at least on one of both consecutive days increased to 311 for *P. falciparum* and 616 for *P. vivax*. Table 6.2 summarizes the PCR results on day 1 and 2 of paired samples. Again the prevalence of *P. falciparum* as well as *P. vivax* infection did not differ significantly between both days (*P. falciparum*: 27.8% on day 1 vs. 28.5% on day 2; McNemar's test: χ^2=1.0, p=0.3, *P. vivax*: 58.5% on day 1 vs. 60.5% on day 2; McNemar's test: χ^2=2.21, p=0.14). When typing results from both days were combined, the prevalence increased only marginally: from 28% to 30.6% for *P. falciparum* and from 59.3% to 64.0% for *P. vivax*.

6.3.2 Effect of repeated sampling on detection of individual clones

When assessing the persistence of individual alleles on the consecutive days of sampling, considerable turn over in allele composition was observed. For the msp2 marker of *P.*

	P. falciparum msp2	P. vivax msp1F3	P. vivax MS16	P. vivax 2 markers combined
Positivity in paired samples				
only 1st day positive	21 (6.8%)	39 (6.3%)	49 (7.8%)	40 (6.1%)
only 2nd day positive	28 (9.0%)	56 (9.1%)	58 (9.1%)	55 (8.4%)
Both days positive	262 (84.2%)	521 (84.6%)	527 (83.1%)	557 (85.5%)
Total No. of positive pairs	311	616	634	652
Prevalence				
Prevalence on 1st day	0.29	0.55	0.57	0.59
Prevalence 2 days combined	0.31	0.60	0.62	0.64
Detection of parasite clones				
No. clones detected only on day 1	93 (18.0%)	382 (23.8%)	495 (25.2%)	
No. clones detected only on day 2	90 (17.4%)	307 (19.2%)	617 (31.3%)	
No. clones detected on both days	335 (64.7%)	912 (57.0%)	855 (43.5%)	
Total No. of clones	518	1601	1967	
Detectability *q* incl. CI	0.79 [0.76 - 0.82]	0.73 [0.71 - 0.75]	0.61 [0.58 - 0.63]	
Multiplicity of infection				
MOI on day 1	1.52	2.31	2.34	2.78
MOI on day 2	1.47	2.11	2.52	2.77
MOI on both days	1.68	2.60	3.10	3.37

Table 6.2: **Effect of repeated sampling on detection of parasites and alleles by PCR and on multiplicity if infection** - Combining of markers for *P. vivax*: A sample was defined positive if any of the two markers msp1F3 or MS16 was amplified. The highest value for MOI of either marker was used.

falciparum 64.7% of all genotypes were observed on both days. For *P. vivax* two molecular markers, msp1F3 and MS16, showing slightly different diversity were genotyped. 57.0% of the msp1F3 alleles and 43.5% of the MS16 alleles were observed on both consecutive days (Table 6.2).

Combining the genotyping results from 24h bleeds makes it possible to assess the effect of repeated sampling on other molecular epidemiological parameters, e.g. multiplicity of infection. In *P. falciparum* combination of results from both days lead to a small increase in mean MOI to 1.68 compared to mean MOI of 1.5 based on a single day (t test for paired data: t=8.5, p<0.001). In *P. vivax* the mean MOI based on msp1F3 increased from 2.21 detected on a single day to 2.60 (t = 7.6, p<0.001) detected on both days. For the highly diverse marker MS16, mean MOI increased from 2.43 to 3.10, respectively (t = 8.7, p<0.001). When both markers were considered to establish mean MOI, i.e. for each pair the highest number of clones observed was counted, MOI increased from 2.78 based on a single day bleed to 3.37 based on results of both consecutive days (t = 18.4, p<0.001).*q* was in the same range for both *Plasmodium* species: for *P. falciparum* clones detectability was 0.79 and for *P. vivax* detectability was 0.73 based on msp1F3 marker and 0.61 based on microsatellite MS16. Table 6.3 lists the detectability calculated for different age groups (0-2 years, 2-3 years, >3 years). For *P. falciparum* msp2 and *P. vivax* msp1F3 no substantial difference was observed between the 3 groups (95% CI overlap). Detectability of *P. vivax* MS16 was lower in children older than 3 years (no overlap of 95% CI).

	P. falciparum msp2		P. vivax msp1F3		P. vivax MS16	
	Detectability	CI	Detectability	CI	Detectability	CI
0-2 years	0.80	[0.71 – 0.88]	0.72	[0.68 - 0.76]	0.63	[0.59 - 0.67]
2-3 years	0.81	[0.77 – 0.85]	0.73	[0.70 - 0.76]	0.63	[0.60 - 0.66]
>3 years	0.75	[0.70 – 0.81]	0.72	[0.68 - 0.76]	0.55	[0.51 - 0.59]

Table 6.3: **Detectability of parasite clones by PCR in different age groups**

6.4 Discussion

Many epidemiological surveys are crucially dependent on accurate diagnosis of presences and/or complexity of infecting parasites strains. In the case of *Plasmodium* species imperfect detectability in blood samples can occur when parasites sequester in the deep blood vessels for part of their life cycle or when parasite densities fluctuate around the detection limit of light microscopy or PCR. As a consequence, a single blood sample might not be representative for the entire parasite population in a host at a given time and parameters such as prevalence, MOI, duration of infection or outcome of treatment failures can be compromised. This study aimed at investigating changes in the genotypic profile of *P. falciparum* and *P. vivax* populations within a host over a period of 24 hours in children 1-4.5 years of age.

Our analysis of samples collected 24 hours apart revealed imited day to day fluctuations in the detection of *P. falciparum* and *P. vivax* infection. This indicates that short-term sampling has only a small impact on prevalence - regardless whether infections are detected by light microscopy or PCR. Prevalence by PCR increased only between 7 and 12% (depending on the marker used) when 2 days were combined. A more pronounced difference was only observed for *P. falciparum* prevalence determined by light microscopy that increased by 24%. These finding suggests that for *Plasmodium* species conducting repeated sampling within 24 h does not substantially increase the total number of infections detected.

Detection of individual clones was very high in *P. falciparum* (q =79%). Accordingly, combining genotypes from both days resulted in a small increase in mean MOI rising from 1.52 based on one day only to 1.68 for both days. Children <5 years have not yet developed a strong immunity to *P. falciparum* [117] and therefore carry high parasite densities (mean parasite density: 2558 parasites/μl), which leads to a better chance to detect most of the parasites present by PCR.

In contrast to *P. falciparum*, sequestration of late stage parasites is not reported from *P. vivax*. Despite this biological difference, the detectability of clones was lower in *P. vivax* than in *P. falciparum*. A larger number of *P. vivax* clones was only deteced on either day for both *P. vivax* markers analyzed. In *P. vivax* parasite densities are generally much lower, in our study mean *P. vivax* density was 498 parasites/μl compared to 2558 parasites/μl of *P. falciparum* positive samples. Our results suggest that the low parasitaemia of *P. vivax* has a larger impact on detectability than sequestration plus occasional low parasitaema levels in *P. falciparum*. This is in line with a recent study by Bretscher *et al* [82] that failed to find a 48 hours periodicity in *P. falciparum* positive samples. A further reason for the difference in detectability of clones might lie in the more complex *P. vivax* infections. *P. vivax* mean MOI

was 3.37 (2 markers combined) versus 1.7 of *P. falciparum*. In samples harbouring a high MOI, the chance of amplifying only the dominant clone in a PCR reaction and of missing minority clones due to template competition is higher than in samples where infections are less complex. Indeed *P. vivax* detectability decreases in our data with increasing MOI (data not shown). The different results for the two *P. vivax* markers likely represent the overall better amplification of the msp1F3 PCR compared to the MS16 PCR. In addition different resolution of markers adds to the effect as two clones share more often the same msp1F3 allele (virtual heterozygosity HE=0.88) than the same MS16 allele (HE=0.98).

The lower detectability of the *P. vivax* markers msp1F3 (q=0.73) and MS16 (q=0.61) compared to 0.79 in *P. falciparum* suggests that a single bleed does usually not fully reflect the complexity of concurrently infecting *P. vivax* clones. The effect of repeated sampling on prevalence, however, was in the same range in *P. falciparum* and *P. vivax*. In both cases around 15% of all infections were missed on one day. This is astonishing, as *P. falciparum* sequestration and as a consequence absence of parasites from the blood stream had been thought to be a main difference between the two parasites. In fact the ability of *P. falciparum* to sequester is considered as a major reason for the generally more severe outcome of *P. falciparum* infection [121]. Future research will show whether the recent observation of adhesion of *P. vivax* infected red blood cells to different human cells indicates a certain level of sequestration in *P. vivax*. Our data suggest that regarding presence of parasites in the blood stream the two species are less different than previously assumed.

It had been shown that in the age group studied densities of *P. vivax* drop with increasing age, while *P. falciparum* parasitaemia remains unchanged. If densities directly impact detectability, a decrease in *P. vivax* detectability is expected with increasing age. In the present study, detectability of *P. vivax* for marker MS16 (but not for msp1F3) was indeed lower in children above 3 years of age. Frequent multiple clone infections combined with low parasite loads likely lead to lower detectability. The often-observed decrease of prevalence in age at moderate to high levels of transmission might at least partially be due to a lower detectability associated with decreasing parasite levels.

Our *P. falciparum* results can be compared to a previous study conducted in a highly endemic area in Ghana, where overall detectability was only 35% [52]. However, a strong age-dependency of detectability [37] was noted and in individuals of the same age group detectability ranges from 0.51 and 0.55 [82]. The low overall value in Ghana was caused by a detectability of only around 10% in highly immune adults. Due to the narrow age range in our study (1 to 4.5 years) similar age trends were not observed. The lower transmission and therefore slower acquisition of immunity in PNG might lead to generally lower age dependence in detectability. Our current results reflect the situation in children harbouring high parasite densities and might differ in adults. In addition it is possible that the higher endemicity in Ghana compared to PNG might have lowered the estimates of detectability of P. falciparum. The discussed effect of MOI on detectability likely influences the different estimates from the 2 countries.

We conclude that measures of prevalence obtained from cross sectional surveys give a reasonable good picture of the overall malaria situation in a region. In this context it is also important to point out that prevalence is a very stable measure and is surprisingly independent of other factors such as use of insecticide treated bed nets [11] or entomological

inoculation rate [122]. In a study in the same area in PNG almost 20 years ago only a very weak correlation between entomological inoculation rate and prevalence was observed: a ten-fold increase in entomological inoculation rate led only to 28% increase in *P. falciparum* prevalence and a 7% increase in *P. vivax* prevalence [122].
We observed a limited increase in precision of estimates of epidemiological parameters in our study for both *P. falciparum* and *P. vivax*, with surprisingly little difference between the two parasites. We conclude that routine 24 hrs bleeds may not be justified in children below 5 years. In older children, however, and in highly endemic areas where individuals often carry very low parasite density as a result of acquired immunity, the considerable logistical efforts required for this sampling scheme might well be justifiable or even required.

CHAPTER 7

General discussion

The aim of this thesis was to measure the distribution of *Plasmodium falciparum* infection durations, in continuation of the thesis of Wilson Sama [123]. Sama's work represented significant advances in measuring the duration of clonal infections from longitudinal genotyping data [37, 57], but a descriptive analysis of malariatherapy data [49] had revealed that exponential models might significantly misrepresent the survival of *P. falciparum* infections. It was therefore envisaged to drop the assumption of a constant clearance rate in analyses of longitudinal genotyping data. A mathematical approach to modeling the age structure of an infection population was undertaken by Wilson Sama in collaboration with Klaus Dietz ([123], Appendix), yielding expressions for the relative frequencies of all true pattern types. While the obtained solution was correct, the algebraic expressions were unpractically complicated, and it was decided for this thesis to use stochastic simulation techniques instead. Stochastic simulation was pursued for 2 years, and later abandoned for an approach using explicit calculation of the likelihood, due to problems with model fitting with stochastic lossfunctions. Although model fitting using explicit likelihoods proved demanding for other reasons, those problems were eventually solved and the method was applied to a dataset from Navrongo, Northern Ghana, in chapters 4 and 5.

7.1 Review of chapters

Chapter 2 explored the short-term dynamics of detectability of clonal infections. The reason for this in the context of the whole thesis is that the immigration-death models presented by [37, 57] as well as the approaches used in chapters 4 and 5 rely on binomial models of detection. This implies that detection of an infecting clone at two different time points are assumed independent. Yet, numerous publications discuss whether detection of parasites may be connected to periodic behaviour of parasite densities - influenced by the 48 hour erythrocytic cycle of *P. falciparum*, and sequestration, i.e. binding of the late intra-erythrocytic stages of the parasite to the vascular endothelia. Such periodicity of detectability could compromise estimates from immigration-death models, because the assumption of independence of detections would be violated. Chapter 2 analysed genotyping data with short sampling intervals and discussed the implications for using immigration-death models. The main conclusion was that when sampling at intervals of 7 days (or more) it is safe to assume indepence of detections. No 24 h periodicity was found. A simple formula was provided for estimating detectability from pairs of surveys, with the aim to convey the concept of detectability to research fields which currently pay little attention to it (e.g. drug efficacy trials), and provide field-epidemiologists with a practical tool for measuring and

reporting detectability alongside standard epidemiological measures.

Chapter 3 reported for the first time the complete genotyping dataset from a longitudinal molecular study in Navrongo, Northern Ghana, providing a comprehensive description of the molecular results. The exponential model from chapter 5 was used with eight age groups. Rates of infection clearance and the detectability were estimated and compared to analogous estimates from the simpler "triplet" model [58]. The results broadly agreed, with the "triplet" model estimating higher detectability in children below one year of age, because its detectability estimates were not constrained to be a monotonic function of host age.

Chapter 4 developed a statistical method for analysing longitudinal genotyping data. The method is based on the approach of [37, 57], but not restricted to exponential models of infection survival. Rather it is possible to use Weibull, lognormal, exponential or gamma distributions to model clearance of infections. This allowed investigation of the relationship between age of an infection and infection clearance. The dependence of infection clearance on host age was ignored. This was justified by the results of [37], which reported that their best-fitting model assumed a single clearance rate across all host ages. An unexpected difference between the distribution of infection duration estimates from the Ghanaian dataset and analogous measurements from malariatherapy data [49] emerged: a vast majority of infections in the naturally exposed population appeared to be cleared soon after inoculation. Several competing hypotheses were brought forward, aiming to explain the observed difference between the two datasets. These include hypotheses considering acquired immunity as a possible cause as well as factors which are not related to acquired immunity. It was noted that by using host age as proxy for cumulative exposure, and therefore acquired long-term immunity, some of the immunity-related hypotheses could be falsified.

Chapter 5 expanded on the parameterizations of the immigration-death model in chapter 4 by estimating a separate set of distribution parameters for each host age-group. This meant that lasting effects of cumulative exposure, which are assumed to increase in intensity with host age, became observable in form of the change of distribution parameters with age. The results indicated, that the previously observed early clearance of most infections cannot be explained by long-lasting acquired immunity. Various quantities derived from the parameter estimates for the Weibull, lognormal, exponential and gamma models were plotted against host age to describe the effects of immunity. These include foremost a moderate reduction of the average duration of infection and a pronounced reduction of detectability. The expected time (after inoculation) until nearly all infections are cleared was found to be larger than previsouly assumed using exponential models. This has implications for planning of strategies for local elimination. However, the predicted times lie outside the range of observations, since the models were fitted to data from a one-year study. They They may therefore be influenced by assumptions about the highly seasonal force of infection (**FOI**) in the past or by the range of possible shapes permitted by the various survival distributions.

Chapter 6 analysed blood samples from a study in Papua New Guinea, where children be-

tween 1 and 4.5 years of age were followed up in two-monthly intervals for 16 months. Additional blood samples were taken 24 hours after each follow-up visit. All samples were analysed for presence of *P. falciparum* and *P. vivax* using PCR. A formula from chapter 2 was used to estimate the detectability of clonal infections, and estimates were compared between species. It was found that the detectabilities of *P. vivax* and *P. falciparum* were similar and high (between 60% and 80%), whereas parasite densities were higher in *P. falciparum* than in *P. vivax*.

7.2 Results of the thesis

The main achievement of this thesis is to have developed a method which makes it possible to use non-exponential survival distributions for the statistical analysis of longitudinal genotyping data. The main finding is that a large proportion of clonal infections in the Ghanaian study population is cleared relatively soon after inoculation. This is different from the results of an analysis of malariatherapy data [49]. The presence of early clearance of infections in all host ages revealed that this difference cannot be explained by a lasting effect of previous exposure.
Some of the assumptions underlying the statistical method used to fit survival distributions to the Ghanaian dataset, one could argue, may be incorrect. As an example, the detectability of infections is assumed to be constant throughout the course of infections, even though parasite densities are known to decrease with time, and detectability appears to be related to parasite densities. The possibility that the main result of fast clearance of infections in the Ghanaian dataset is merely an artefact can therefore not be excluded with certainty if only a single statistical method is used. In addition, the short infection durations postulated by the analyses in chapters 4 and 5 cannot be well observed using survey intervals of 2 months. To verify if early clearance of infections is common in the analysed dataset, an additional analysis was performed (Figure 7.1), and the main finding of the thesis could be confirmed with a different statistical method.

7.2.1 Interpretation

Several competing hypotheses aiming to explain the high abundance of short infection durations in Ghana compared to malariatherapy data have been discussed in chapters 4 and 5. These identify genetic differences between the *P. falciparum* strains or between the host populations, and interaction of infections as possible causes. While there is a clear possibility that malariatherapy doctors selected against strains with short infection durations, it is almost certain that natural selection would do the same: the duration of infection is related to the duration of potential infectivity to mosquitoes and directly connected to parasite fitness. That host genetics may be playing a role cannot be excluded with certainty, but given the large proportion of short infection durations, genetic mutations conferring resistance against malaria to the carrier would need to be present at very high frequencies.
It is, however, very hard to argue that there is no interaction between clones which are simultaneously infecting the same host: the immune response against an ongoing infection

is likely to also affect newly arriving clones. In fact, any type of cross-reactive immunity should be strongest when another strain is present in the host, stimulating both innate and adaptive immune responses by constant antigen presentation. The parasite has adapted to escape the immune response of the host, and it seems plausible that the initial phase of a blood-stage infection may be a very vulnerable stage in the life cycle of the parasite: the population size of a newly arriving clone may be small initially, making it vulnerable to extinction, and a suitable var-gene variant for optimally escaping host immune defenses may not yet have been found.
Most mathematical within-host models consider only single infections, and often use malariatherapy data to calibrate parameters. A model of Gatton & Cheng [124] allows multiple concurrent infections per host. In the particular example shown in [124], 10% of blood stage infections did not last for longer than 10 days after acquisition of few infections. This indicates that rapid clearance of clones infecting an already infected host is implied by current ideas about within-host dynamics and immunity. Schöpflin *et al* (personal communication) found that multiplicity of infection (**MOI**) significantly reduced the number of new *P. falciparum* infections in the dataset used in chapter 6. It is generally problematic to use MOI as explanatory for the FOI parameter, because of the known causal influences of the FOI and duration of infection on MOI, resulting in circularity of causal relationships. In concrete terms, one would expect MOI to be positively associated with the FOI, with individuals experiencing higher exposure also showing a higher MOI. Since the measured effect of MOI on the FOI was negative, however, the result can be interpreted as showing a protective effect of pre-existing infections against the establishment of new blood-stage infections.
Experimental infection of *Saimiri* or *Aotus* monkeys with multiple *P. falciparum* strains [125–127] could provide further insight into within-host interactions in *falciparum* malaria. Alternatively, a confirmation that the abundance of short infection durations is caused by interaction of infections could be obtained through analysis of field data from low-transmission areas, where superinfection is rare, using the method developed in this thesis. However, this type of data may not be easily available, as people in low transmission areas rarely have asymptomatic infections and would need to be treated.

7.2.2 Virulence of *P. falciparum*

There are several approaches in evolutionary biology aiming to explain parasite virulence. Some emphasize the importance of trade-offs between virulence and life-history traits which directly influence reproductive success of the parasite, others stress the role of inter-species or intra-species competition, or whether a parasite is a specialist or a generalist with respect to its host range [128]. Interaction of infections within a host, such that a large proportion of newly arriving infections are cleared quickly, is synonymous with strong within-host competition among co-infecting strains: a strain which cannot establish a lasting infection is unlikely to transmit to mosquitoes, and therefore has a selective disadvantage. Due to the absence of a lasting liver stage, the cost of not being able to establish a lasting blood-stage infection must be particularly high for *P. falciparum*, compared to other human malaria parasites.

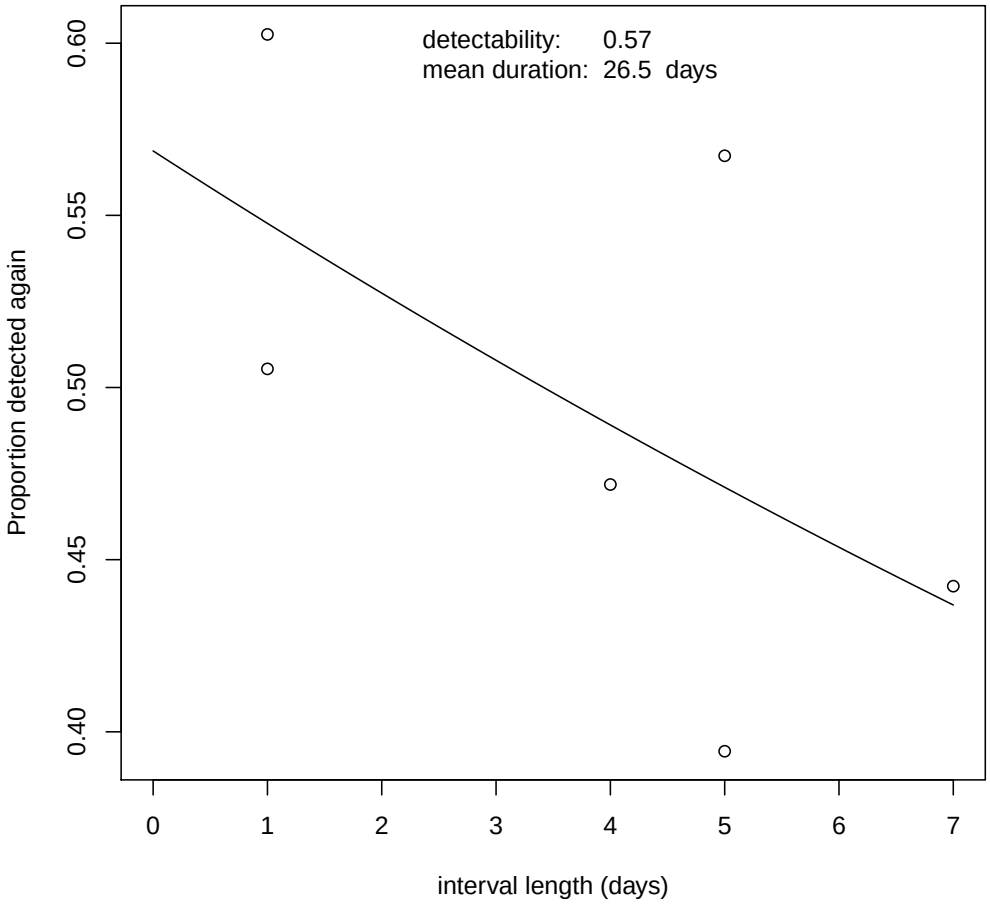


Figure 7.1: **Early clearance of infections confirmed** - An estimate of the duration of infection was obtained using a different methodology than the one developed in chapter 4. For all possible survey pairs from the short-term study described in chapter 2, the proportion of infections detected at the second survey, conditional on confirmed presence at the first survey, was calculated (data points). A model for this proportion p, depending on detectability q, clearance rate μ, and time t between survey pairs, was fitted using log-transformation of the data and linear regression. The model was of the form $p = qe^{-\mu t}$. Conclusion: a large proportion of infections is cleared within short time. Most of these can only be observed directly when sampling intervals are short, but the method developed in chapter 4 correctly interprets data from studies with long sampling intervals.

Kin-selection models of virulence argue that such intra-host competition should result in increased virulence of parasites [129]: in absence of competition, trade-offs between host-death (one aspect of virulence) and reproductive output of the parasite would lead to intermediate virulence. Conversely, when other strains are present, a higher virulence is expected because parasites that treat the host kindly at their expense of reproductive output are outcompeted by those that exploit the host more vigorously. If the host dies, the prudent and the ruthless clone share the reproductive cost of killing their host in equal parts. If the host survives, the prudent clone has a reproductive disadvantage and is therefore removed from the population by natural selection. This "tragedy of the commons" appears in many areas of evolutionary biology, and well beyond (e.g. in economics).
What would the concrete nature of this evolutionary game be in the present case, where an newly infecting clone is faced with a competitor which has already established an infection? What would increase the chances for the new clone of not being cleared by cross-reactive immunity, originally mounted against the pre-existing strain, or of securing enough of any limiting resource in order to persist? It appears that fast replication should be advantageous for the intruder under several possible scenarios: again we encounter a "tragedy of the commons", this time not with the host as reproductive resource for between-host transmission, but with competition for limited resources for within-host replication, such as erythrocytes. When the resource is used up, both competitors face a shortage, but as long as it is still available, whoever uses it faster gets more out of it. Even for immune-escape, which may not exactly fit the concept of a resource, it can be argued that faster replication should allow the new parasite to try a larger number of antigenic variants before it suffers the delayed effects of a targeted immune response against itself, which is stimulated as soon as it reaches high population densities.
The incidence of clinical disease in *falciparum* malaria is strongly correlated with peak parasite densities, even across species [130]. The virulence of strains from high transmission areas should therefore be higher, as superinfection is more common. As Read *et al* [129] point out, it is difficult to find evidence for this from epidemiological data, because immunity may be masking a relationship between high R_0 (leading to high incidence of superinfection) and virulence: highly superinfected hosts rather tend to have less episodes, presumably because they have a longer history of exposure and therefore increased immunity against clinical disease.
Provided the assumption that early clearance of infections in the Ghanaian dataset is caused by pre-existing infections can be confirmed, the methods presented in this thesis provide evidence for the existence of fierce intra-host competition among co-infecting strains, and thereby indirectly support the hypothesis that high transmission in conjunction with the absence of a permanent liver stage in *P. falciparum* are likely reasons for the high pathogenicity of *falciparum* malaria.

7.2.3 Consequences for control and elimination

While there is a range of models of malaria transmission which consider multiple infections, it appears that these mostly treat concurrent infections as independent of each other. This is likely due to the absence of reliable information on the nature of such interactions.

In fact, it is analyses like the ones presented in this thesis, or developments thereof, which may have the potential to supply such quantitative information from longitudinal genotyping data. Alternatively, interaction of infections could be accounted for via process-based within-host models, such as the one presented in [124].
Transmission models ignoring interaction of infections, such as the one in [44], should yield biased predictions of malaria epidemiology in high transmission areas. Assumed values of R_0, based on measurements of prevalence or force of infection might be too high, as a large proportion of infections is never able to transmit to mosquitoes. It is plausible that the predicted impact of vector control measures in such high transmission areas may be smaller than expected, because infections become longer as transmission intensity decreases. However, as both fitting of models to data and predictions are performed assuming independence of infections, the consequences might be counterintuitive. A way of testing predictions of the impact of control measures by transmission models would be to assume, as in the Ross-MacDonald model, that no superinfection occurs at all. A comparison of the two extreme assumptions of no superinfection versus independence of if concurrent infections should allow a quantitative assessment of neglecting interaction of infections on predictions.

7.3 Methodology

Data from longitudinal studies contain more information than data from cross-sectional studies. Every study participant may serve as it's own control, permitting identification of heterogeneities, and the stochastic processes of interest can be more directly observed.
In the case of genotyping data, it is impossible to distinguish between absence of clones and detection failure in cross-sectional studies. Correct interpretation of such data therefore requires cohort data and a process-based statistical approach. A series of such methods for the analysis of malaria genotyping data was developed over the years [37, 57, 58, 60]. A short review of these is given below.

7.3.1 Short history of methods

Common to all the currently available process-based approaches to analysing genotyping data [37, 57, 58, 60] is the distinction of different types of binary sequences. Depending on the model, either a multinomial likelihood function is used, making use of the relative frequencies of binary patterns in the data, or a Poisson likelihood is used, making use of absolute pattern frequencies. That either of the two can be used reflects an intimate relationship between these distributions: a set of Poisson-distributed random variables is multinomially distributed, conditional on their sum. In practice, this sum represents the total number of infections present in a host, which is not exactly known due to the detection problem in *falciparum* malaria.
The first approach to analysing longitudinal genotyping data by Smith *et al* [58] uses all binary patterns of length three where a clone was detected on the first occasion. The expected relative frequencies of these patterns are calculated depending on whether the clone is detected at the two survey rounds which follow the first observation, e.g. the pattern 101

arises with a probability $(1-r)(1-q)(1-r)q$, using the clearance probability between two observations, r, and the detectability q. This approach does not estimate a force of infection, but limits the analyses to a subset of the data where an infection is known to be present. Estimates of infection duration and detectability agree with estimates using later developments, as shown in chapter 3. A potential problem of this approach is that the inclusion of binary patterns into the analysis depends on the detectability itself: data from hosts with lower detectability yields fewer "triplets" where a clone is detected on the first occasion. The effects of this variable selection bias on estimates is unclear. However, by only looking at subsets of the data where an infection is known to be present the need to make assumptions about the force of infection in the past does not arise.
The method of Smith & Vounatsou [60] for the first time simultaneously estimated the force of infection along with infection duration and detectability, making use of the entire dataset. A Hidden Markov Model is considered for every clone which appears in the data at least once, and transition probabilities between presence and absence of a particular genotype are derived from a continuous-time reversible catalytic model. A multinomial likelihood function is used to fit the models to data. The high number of parameters made fitting of the models challenging, even with MCMC, and it is not obvious how this method could be extended to non-exponential distributions.
A different approach was chosen by Sama *et al* [37, 57]: instead of predicting the expected relative frequencies of binary patterns for each genotype separately, absolute frequencies of binary patterns are considered. Which genetic marker is producing a particular binary pattern can be ignored because only the sum of patterns is needed to evaluate the likelihood function. The expected number of patterns in the data is determined using a model of superinfection, represented by the differential equation $\frac{dn}{da} = \lambda - \mu n$, with the number n of infections in a host of age a, the FOI λ and clearance rate μ.
In [123] Sama *et al* presented a statistical approach which allows modelling of infection survival using parametric survival distributions, in contrast to the earlier methods, which were limited to exponential distributions of infection duration. By use of a partial differential equation describing the age structure of the infection population within one host, the expected relative frequencies of binary patterns are determined. However, this lead to very complicated mathematical expressions, which prevented the method from being used in practice. The approach developed in chapter 4 overcomes this limitation by using a Poisson instead of a multinomial distribution as likelihood function, modeling absolute numbers of infections instead of relative frequencies. This leads to considerably simpler expressions for the expected pattern frequencies.

7.3.2 Directions of future development

With the available datasets and methodology a range of additional analyses can readily be performed. However, since the method of chapter 4 is computationally costly, it is at the moment not practical to use MCMC methods in order to investigate e.g. heterogeneity in transmission, unless parallelization of the computer program is achieved. Rather, one could follow the much simpler approach of splitting the data according to explanatories of interest (e.g. MOI, geographical location).

A major problem for analysis of datasets with low marker diversity is re-infection with the same marker. This problem was encountered when attempting to analyse the dataset presented in chapter 6 using an immigration-death model. If the assumption that re-infection does not occur is not fulfilled, parameter estimates are biased. In particular, detectability estimates approach zero, because the statistical model interprets non-detection of an absent marker that later appears again as a failure to detect it. Apart from explicit simulation of the process of infection with different markers, and model fitting using MCMC, it is not clear how this problem can be overcome.
Future developments of the immigration-death model presented in this thesis might consider a Bernoulli differential equation [1] to explicitly model an effect of pre-existing infections on new ones.

In the absence of easily available animal models for research on *falciparum* malaria, molecular epidemiological data represents one of the most important sources of information on the within-host dynamics of *P. falciparum*, especially in the context of multiple concurrent infections. Efforts to extract as much from these data as possible should therefore be made. I hope that the present thesis may inspire future research in this direction by showing the range of questions that can be approached through analysis of longitudinal genotyping data, in conjunction with process-based statistical time-series models ...

[1] Such a model of superinfection could be of the form $\frac{dn}{da} = \lambda - \mu n^{\beta}$, where n is the number of infections in a host of age a, λ denotes the FOI, and μ the clearance rate. The positive real-valued parameter β describes the degree of interaction between infections. This type of nonlinear differential equation is analytically solvable.

Bibliography

[1] Sachs J, Malaney P: **The economic and social burden of malaria**. *Nature* 2002, **415**:680–685.

[2] WHO: **World Malaria Report 2009** `http://www.who.int/malaria`.

[3] Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI: **The global distribution of clinical episodes of *Plasmodium falciparum* malaria**. *Nature* 2005, **434**:214–217.

[4] Kiszewski A, Mellinger A, Spielman A, Malaney P, Sachs SE, Sachs J: **A global index representing the stability of malaria transmission**. *Am J Trop Med Hyg* 2004, **70**:486–498.

[5] Warrell DA, Gilles HM: **Essential Malariology**. A Hodder Arnold Publication, 4 edition 2002.

[6] Singh B, Sung LK, Matusop A, Radhakrishnan A, Shamsul SSG, Cox-Singh J, Thomas A, Conway DJ: **A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings**. *Lancet* 2004, **363**:1017–1024.

[7] Cox-Singh J, Singh B: **Knowlesi malaria: newly emergent and of public health importance?** *Trends in Parasitology* 2008, **24**:406–410.

[8] Greenwood B, Marsh K, Snow R: **Why do some African children develop severe malaria?** *Parasitology Today (Personal Ed.)* 1991, **7**:277–281.

[9] Gillies M, DeMeillon B: **The Anophelinae of Africa South of the Sahara (Ethiopian zoogeographical region).** Johannesburg: South African Medical Research Institute 1968.

[10] Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW: **The global distribution and population at risk of malaria: past, present, and future**. *The Lancet Infectious Diseases* 2004, **4**:327–336.

[11] Lengeler C: **Insecticide-treated bed nets and curtains for preventing malaria**. *Cochrane Database of Systematic Reviews (Online)* 2004, :CD000363.

[12] Kouznetsov RL: **Malaria control by application of indoor spraying of residual insecticides in tropical Africa and its impact on community health**. *Tropical Doctor* 1977, **7**:81–91.

[13] Mabaso MLH, Sharp B, Lengeler C: **Historical review of malarial control in southern African with emphasis on the use of indoor residual house-spraying**. *Tropical Medicine & International Health: TM & IH* 2004, **9**:846–856.

[14] Kleinschmidt I, Torrez M, Schwabe C, Benavente L, Seocharan I, Jituboh D, Nseng G, Sharp B: **Factors influencing the effectiveness of malaria control in Bioko Island, equatorial Guinea**. *The American Journal of Tropical Medicine and Hygiene* 2007, **76**:1027–1032.

[15] Carter R, Mendis KN: **Evolutionary and Historical Aspects of the Burden of Malaria**. *Clin. Microbiol. Rev.* 2002, **15**:564–594.

[16] Bruce-Chwatt LJ, Zulueta J, de Zulueta J: **The Rise and Fall of Malaria in Europe: A Historico-Epidemiological Study**. Oxford University Press 1981.

[17] Greenwood BM, Bojang K, Whitty CJM, Targett GAT: **Malaria**. *Lancet* 2005, **365**:1487–1498.

[18] Killeen GF, Seyoum A, Knols BGJ: **Rationalizing historical successes of malaria control in Africa in terms of mosquito resource availability management**. *The American Journal of Tropical Medicine and Hygiene* 2004, **71**:87–93.

[19] Roberts L, Enserink M: **Malaria. Did they really say ... eradication?** *Science (New York, N.Y.)* 2007, **318**:1544–1545.

[20] Tanner M, de Savigny D: **Malaria eradication back on the table**. *Bulletin of the World Health Organization* 2008, **86**:82.

[21] CDC: **Schema of the Life Cycle of Malaria** 2006, `http://www.cdc.gov/malaria/biology/life_cycle.htm`.

[22] Dietz K, Molineaux L, Thomas A: **A malaria model tested in the African savannah**. *Bull World Health Organ* 1974, **50**:347–57.

[23] MacDonald G: **The Epidemiology and control of Malaria**. Oxford 1957.

[24] Gillies M: **Anopheline mosquitos: vector behaviour and bionomics**. Edinburgh: Churchill Livingstone 1988.

[25] Stauffer WM, Cartwright CP, Olson DA, Juni BA, Taylor CM, Bowers SH, Hanson KL, Rosenblatt JE, Boulware DR: **Diagnostic performance of rapid diagnostic tests versus blood smears for malaria in US clinical practice**. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 2009, **49**:908–913.

[26] Snow R, Marsh K: **The consequences of reducing transmission of *Plasmodium falciparum* in Africa**. *Adv Parasitol* 2002, **52**:235–64.

[27] Aikawa M: **Human Cerebral Malaria**. *Am J Trop Med Hyg* 1988, **39**:3–10.

[28] MacPherson GG, Warrell MJ, White NJ, Looareesuwan S, Warrell DA: **Human cerebral malaria. A quantitative ultrastructural analysis of parasitized erythrocyte sequestration**. *The American Journal of Pathology* 1985, **119**:385–401.

[29] Fried M, Duffy PE: **Maternal malaria and parasite adhesion**. *Journal of Molecular Medicine (Berlin, Germany)* 1998, **76**:162–171.

[30] Marsh K: **Malaria–a neglected disease?** *Parasitology* 1992, **104 Suppl**:S53–69.

[31] Snow RW, Omumbo JA, Lowe B, Molyneux CS, Obiero JO, Palmer A, Weber MW, Pinder M, Nahlen B, Obonyo C, Newbold C, Gupta S, Marsh K: **Relation between severe malaria morbidity in children and level of *Plasmodium falciparum* transmission in Africa**. *Lancet* 1997, **349**:1650–1654.

[32] Baird JK: **Age-dependent characteristics of protection v. susceptibility to *Plasmodium falciparum***. *Annals of Tropical Medicine and Parasitology* 1998, **92**:367–390.

[33] Molineaux L, Muir D, Spencer H, Werndorfer W: **The epidemiology of malaria and its measurement**. Edinburgh: Churchill Livingstone 1988.

[34] Rogier C, Trape JF: **[Study of premunition development in holo- and meso-endemic malaria areas in Dielmo and Ndiop (Senegal): preliminary results, 1990-1994[]**. *Médecine Tropicale: Revue Du Corps De Santé Colonial* 1995, **55**:71–76.

[35] Collins WE, Jeffery GM: **A retrospective examination of secondary sporozoite- and trophozoite-induced infections with *Plasmodium falciparum*: development of parasitologic and clinical immunity following secondary infection**. *Am J Trop Med Hyg* 1999, **61**:20–35.

[36] Molineaux L, Traeuble M, Collins WE, Jeffery GM, Dietz K: **Malaria therapy reinoculation data suggest individual variation of an innate immune response and independent acquisition of antiparasitic and antitoxic immunities**. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2002, **96**:205–209.

[37] Sama W, Owusu-Agyei S, Felger I, Dietz K, Smith T: **Age and seasonal variation in the transition rates and detectability of *Plasmodium falciparum* malaria**. *Parasitology* 2006, **132**:13–21.

[38] Hoffman SL, Wistar R, Ballou WR, Hollingdale MR, Wirtz RA, Schneider I, Marwoto HA, Hockmeyer WT: **Immunity to malaria and naturally acquired antibodies to the circumsporozoite protein of *Plasmodium falciparum***. *The New England Journal of Medicine* 1986, **315**:601–606.

[39] O'Meara WP, Collins WE, McKenzie FE: **Parasite prevalence: a static measure of dynamic infections**. *The American Journal of Tropical Medicine and Hygiene* 2007, **77**:246–249.

[40] Ross SR: **Report on the prevention of malaria in Mauritius**. Waterlow and Sons Limited 1903.

[41] Ross SR: **The prevention of malaria**. Dutton 1910.

[42] Koella J: **On the use of mathematical models of malaria transmission**. *Acta Trop* 1991, **49**:1–25.

[43] McKenzie F, Samba E: **The role of mathematical modeling in evidence-based malaria control**. *Am J Trop Med Hyg* 2004, **71**:94–6.

[44] Smith T, Killeen GF, Maire N, Ross A, Molineaux L, Tediosi F, Hutton G, Utzinger J, Dietz K, Tanner M: **Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of *Plasmodium falciparum* malaria: Overview**. *The American Journal of Tropical Medicine and Hygiene* 2006, **75**:1–10.

[45] MacDonald G: **The analysis of malaria parasite rates in infants**. *Tropical Diseases Bulletin* 1950, **47**:915–938.

[46] WHO: **Guidelines for the evaluation of *Plasmodium falciparum* vaccines in populations exposed to natural infections.** 1997.

[47] Draper CC, Voller A, Carpenter RG: **The epidemiologic interpretation of serologic data in malaria**. *The American Journal of Tropical Medicine and Hygiene* 1972, **21**:696–703.

[48] Sama W, Killeen G, Smith T: **Estimating the duration of *Plasmodium falciparum* infection from trials of indoor residual spraying**. *Am J Trop Med Hyg* 2004, **70**:625–34.

[49] Sama W, Dietz K, Smith T: **Distribution of survival times of deliberate *Plasmodium falciparum* infections in tertiary syphilis patients**. *Trans R Soc Trop Med Hyg* 2006.

[50] Mullen K, Prost A: **Decreased microfilarial load and its effect on the calculation of prevalence and the rate of false negatives in the detection of onchocerciasis**. *Int J Epidemiol* 1983, **12**:102–104.

[51] Aron J: **Malaria epidemiology and detectability**. *Trans R Soc Trop Med Hyg* 1982, **76**:595–601.

[52] Falk N, Maire N, Sama W, Owusu-Agyei S, Smith T, Beck H, Felger I: **Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of *Plasmodium falciparum***. *Am J Trop Med Hyg* 2006, **74**:944–950.

[53] Cowman AF, Baldi DL, Healer J, Mills KE, O'Donnell RA, Reed MB, Triglia T, Wickham ME, Crabb BS: **Functional analysis of proteins involved in *Plasmodium falciparum* merozoite invasion of red blood cells**. *FEBS Letters* 2000, **476**:84–88.

[54] Dubbeld MA, Kocken CH, Thomas AW: **Merozoite surface protein 2 of *Plasmodium reichenowi* is a unique mosaic of *Plasmodium falciparum* allelic forms and species-specific elements**. *Molecular and Biochemical Parasitology* 1998, **92**:187–192.

[55] Hill W, Babiker H: **Estimation of numbers of malaria clones in blood samples**. *Proc Biol Sci* 1995, **262**:249–257.

[56] Gatton ML, Cheng Q: **Can estimates of antimalarial efficacy from field studies be improved?** *Trends in Parasitology* 2008, **24**:68–73.

[57] Sama W, Owusu-Agyei S, Felger I, Vounatsou P, Smith T: **An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect**. *Stat Med* 2005, **24**:3269–3288.

[58] Smith T, Felger I, Fraser-Hurt N, Beck H: **Effect of insecticide-treated bed nets on the dynamics of multiple *Plasmodium falciparum* infections**. *Trans R Soc Trop Med Hyg* 1999, **93 Suppl 1**:53–57.

[59] Hastings IM, Smith TA: **MalHaploFreq: a computer programme for estimating malaria haplotype frequencies from blood samples**. *Malar J* 2008, **7**:130.

[60] Smith T, Vounatsou P: **Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models**. *Stat Med* 2003, **22**:1709–1724.

[61] Molineaux L, Diebner H, Eichner M, Collins W, Jeffery G, Dietz K: ***Plasmodium falciparum* parasitaemia described by a new mathematical model**. *Parasitology* 2001, **122**:379–391.

[62] Farnert A, Snounou G, Rooth I, Bjorkman A: **Daily dynamics of *Plasmodium falciparum* subpopulations in asymptomatic children in a holoendemic area**. *Am J Trop Med Hyg* 1997, **56**:538–547.

[63] Owusu-Agyei S, Smith T, Beck H, Amenga-Etego L, Felger I: **Molecular epidemiology of *Plasmodium falciparum* infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana**. *Trop Med Int Health* 2002, **7**:421–428.

[64] Binka F, Morris S, Ross D, Arthur P, Aryeetey M: **Patterns of malaria morbidity and mortality in children in northern Ghana**. *Trans R Soc Trop Med Hyg* 1994, **88**:381–385.

[65] Felger I, Irion A, Steiger S, Beck H: **Genotypes of merozoite surface protein 2 of *Plasmodium falciparum* in Tanzania**. *Trans R Soc Trop Med Hyg* 1999, **93**:3–9.

[66] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A: **Bayesian measures of model complexity and fit**. *J R Stat Soc Ser B* 2002, **64**:583–639.

[67] Lunn DJ, Thomas A, Best N, Spiegelhalter D: **WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility**. *Stat Comput* 2000, **10**:325–337.

[68] Team RDC: **R: A Language and Environment for Statistical Computing**. Vienna, Austria 2008, `http://www.R-project.org`. ISBN 3-900051-07-0.

[69] Smith T, Beck H, Kitua A, Mwankusye S, Felger I, Fraser-Hurt N, Irion A, Alonso P, Teuscher T, Tanner M: **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* 1999, **93**:15–20.

[70] Bruce M, Donnelly C, Packer M, Lagog M, Gibson N, Narara A, Walliker D, Alpers M, Day K: **Age- and species-specific duration of infection in asymptomatic malaria infections in Papua New Guinea**. *Parasitology* 2000, **121**:247–256.

[71] Kwiatkowski D: **Febrile temperatures can synchronize the growth of *Plasmodium falciparum* in vitro**. *J Exp Med* 1989, **169**:357–361.

[72] Kwiatkowski D, Nowak M: **Periodic and chaotic host-parasite interactions in human malaria**. *Proc Natl Acad Sci U S A* 1991, **88**:5111–5113.

[73] Hawking F, Worms M, Gammage K: **24- and 48-hour cycles of malaria parasites in the blood; their purpose, production and control**. *Trans R Soc Trop Med Hyg* 1968, **62**:731–765.

[74] Farnert A, Lebbad M, Faraja L, Rooth I: **Extensive dynamics of *Plasmodium falciparum* densities, stages and genotyping profiles**. *Malar J* 2008, **7**:241.

[75] Magesa SM, Mdira KY, Babiker HA, Alifrangis M, Färnert A, Simonsen PE, Bygbjerg IC, Walliker D, Jakobsen PH: **Diversity of *Plasmodium falciparum* clones infecting children living in a holoendemic area in north-eastern Tanzania**. *Acta Tropica* 2002, **84**:83–92.

[76] Bruce MC, Galinski MR, Barnwell JW, Donnelly CA, Walmsley M, Alpers MP, Walliker D, Day KP: **Genetic diversity and dynamics of *Plasmodium falciparum* and *P. vivax* populations in multiply infected children with asymptomatic malaria infections in Papua New Guinea**. *Parasitology* 2000, **121 (Pt 3)**:257–272.

[77] Martensson A, Ngasala B, Ursing J, Veiga MI, Wiklund L, Membi C, Montgomery SM, Premji Z, Farnert A, Bjorkman A: **Influence of consecutive-day blood sampling on polymerase chain reaction-adjusted parasitological cure rates in an antimalarial-drug trial conducted in Tanzania**. *J Infect Dis* 2007, **195**:597–601.

[78] Daubersies P, Sallenave-Sales S, Magne S, Trape JF, Contamin H, Fandeur T, Rogier C, Mercereau-Puijalon O, Druilhe P: **Rapid turnover of *Plasmodium falciparum* populations in asymptomatic individuals living in a high transmission area**. *The American Journal of Tropical Medicine and Hygiene* 1996, **54**:18–26.

[79] Appawu M, Owusu-Agyei S, Dadzie S, Asoala V, Anto F, Koram K, Rogers W, Nkrumah F, Hoffman SL, Fryauff DJ: **Malaria transmission dynamics at a site in northern Ghana proposed for testing malaria vaccines**. *Tropical Medicine & International Health: TM & IH* 2004, **9**:164–170.

[80] Eyles DE, Young MD: **The duration of untreated or inadequately treated *Plasmodium falciparum* infections in the human host**. *Journal. National Malaria Society* 1951, **10**(4):327–336.

[81] Babiker HA, Abdel-Muhsin AM, Ranford-Cartwright LC, Satti G, Walliker D: **Characteristics of *Plasmodium falciparum* parasites that survive the lengthy dry season in eastern Sudan where malaria transmission is markedly seasonal**. *The American Journal of Tropical Medicine and Hygiene* 1998, **59**(4):582–590.

[82] Bretscher MT, Valsangiacomo F, Owusu-Agyei S, Penny MA, Felger I, Smith T: **Detectability of *Plasmodium falciparum* clones**. *Malaria Journal* 2010, **9**:234.

[83] Carnevale P, Frezil JL, Bosseno MF, Lepont F, Lancien J: **Study of Agressivity of Anopheles-Gambiae A in Relation to Age and Sex of Human Subjects.** *Bull WHO* 1978, **56**:147–154.

[84] Port GR, Boreham PFL, Bryan JH: **The Relationship of Host Size to Feeding by Mosquitoes of the *Anopheles Gambiae Giles* Complex (*Diptera: Culicidae*)**. *Bull Entomol Res* 1980, **70**(01):133–144.

[85] Mayor A, Saute F, Aponte JJ, Almeda J, Gómez-Olivé FX, Dgedge M, Alonso PL: **Plasmodium falciparum multiple infections in Mozambique, its relation to other malariological indices and to prospective risk of malaria morbidity**. *Tropical Medicine & International Health: TM & IH* 2003, **8**:3–11.

[86] Kobbe R, Neuhoff R, Marks F, Adjei S, Langefeld I, von Reden C, Adjei O, Meyer CG, May J: **Seasonal variation and high multiplicity of first Plasmodium falciparum infections in children from a holoendemic area in Ghana, West Africa**. *Tropical Medicine & International Health: TM & IH* 2006, **11**(5):613–619.

[87] Ntoumi F, Contamin H, Rogier C, Bonnefoy S, Trape JF, Mercereau-Puijalon O: **Age-dependent carriage of multiple Plasmodium falciparum merozoite surface antigen-2 alleles in asymptomatic malaria infections**. *The American Journal of Tropical Medicine and Hygiene* 1995, **52**:81–88.

[88] Ofosu-Okyere A, Mackinnon MJ, Sowa MP, Koram KA, Nkrumah F, Osei YD, Hill WG, Wilson MD, Arnot DE: **Novel Plasmodium falciparum clones and rising clone multiplicities are associated with the increase in malaria morbidity in Ghanaian children during the transition into the high transmission season**. *Parasitology* 2001, **123**(Pt 2):113–123.

[89] Arnot D: **Unstable malaria in Sudan: the influence of the dry season. Clone multiplicity of Plasmodium falciparum infections in individuals exposed to variable levels of disease transmission**. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1998, **92**(6):580–585.

[90] Babiker HA, Lines J, Hill WG, Walliker D: **Population structure of Plasmodium falciparum in villages with different malaria endemicity in east Africa**. *The American Journal of Tropical Medicine and Hygiene* 1997, **56**(2):141–147.

[91] Müller DA, Charlwood JD, Felger I, Ferreira C, do Rosario V, Smith T: **Prospective risk of morbidity in relation to multiplicity of infection with Plasmodium falciparum in São Tomé**. *Acta Tropica* 2001, **78**(2):155–162.

[92] Smith T, Felger I, Kitua A, Tanner M, Beck HP: **7. Dynamics of multiple *Plasmodium falciparum* infections in infants in a highly endemic area of Tanzania**. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1999, **93**:35–39.

[93] Szmitko PE, Kohn ML, Simor AE: ***Plasmodium falciparum* malaria occurring 8 years after leaving an endemic area**. *Diagnostic Microbiology and Infectious Disease* 2009, **63**:105–107.

[94] McKenzie FE, Smith DL, O'Meara WP, Riley EM: **Strain theory of malaria: the first 50 years**. *Advances in Parasitology* 2008, **66**:1–46.

[95] Nagelkerke NJ, Chunge RN, Kinoti SN: **Estimation of parasitic infection dynamics when detectability is imperfect**. *Statistics in Medicine* 1990, **9**(10):1211–1219.

[96] Schnabel RB, Koonatz JE, Weiss BE: **A modular system of algorithms for unconstrained minimization**. *ACM Trans. Math. Softw.* 1985, **11**:419–440.

[97] The Apache Software Foundation: **Apache Commons Math Library, Release 2.1** 2010, `http://commons.apache.org/math`.

[98] Collins WE, Jeffery GM: **A retrospective examination of the patterns of recrudescence in patients infected with *Plasmodium falciparum***. *Am J Trop Med Hyg* 1999, **61**:44–48.

[99] Dietz K: **Mathematical models for transmission and control of malaria**. In *Malaria, Principles and Practice of Malariology*, wernsdorfer W.H. & McGregor i. edition, Edinburgh: Churchill Livingstone 1988:1091–1133.

[100] Hethcote H: **The mathematics of infectious diseases**. *Siam Review* 2000, **42**:599–653.

[101] Pasternak ND, Dzikowski R: **PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum***. *Int J Biochem Cell Biol* 2009, **41**(7):1463–1466.

[102] Molineaux L, Dietz K: **Review of intra-host models of malaria**. *Parassitologia* 1999, **41**:221–31.

[103] Guess F, Proschan F: **Mean Residual Life: Theory and Applications**. *Quality Control and Reliability* 1988, :215.

[104] Smith D, Hay S: **Endemicity response timelines for *Plasmodium falciparum* elimination**. *Malaria Journal* 2009, **8**:87.

[105] Guerra CA, Hay SI, Lucioparedes LS, Gikandi PW, Tatem AJ, Noor AM, Snow RW: **Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project**. *Malaria Journal* 2007, **6**:17.

[106] Hay SI, Guerra CA, Gething PW, Patil AP, Tatem AJ, Noor AM, Kabaria CW, Manh BH, Elyazar IRF, Brooker S, Smith DL, Moyeed RA, Snow RW: **A world malaria map: *Plasmodium falciparum* endemicity in 2007**. *PLoS Medicine* 2009, **6**(3):e1000048.

[107] Guerra CA, Howes RE, Patil AP, Gething PW, Boeckel TPV, Temperley WH, Kabaria CW, Tatem AJ, Manh BH, Elyazar IRF, Baird JK, Snow RW, Hay SI: **The international limits and population at risk of *Plasmodium vivax* transmission in 2009**. *PLoS Neglected Tropical Diseases* 2010, **4**(8):e774.

[108] Patil AP, Okiro EA, Gething PW, Guerra CA, Sharma SK, Snow RW, Hay SI: **Defining the relationship between *Plasmodium falciparum* parasite rate and clinical disease: statistical models for disease burden estimation**. *Malaria Journal* 2009, **8**:186.

[109] Schoepflin S, Lin E, Kiniboro B, DaRe JT, Mehlotra RK, Zimmerman PA, Mueller I, Felger I: **Treatment with coartem (artemether-lumefantrine) in Papua New Guinea**. *The American Journal of Tropical Medicine and Hygiene* 2010, **82**(4):529–534.

[110] Nsanzabana C, Hastings IM, Marfurt J, Müller I, Baea K, Rare L, Schapira A, Felger I, Betschart B, Smith TA, Beck H, Genton B: **Quantifying the evolution and impact of antimalarial drug resistance: drug use, spread of resistance, and drug failure over a 12-year period in Papua New Guinea**. *The Journal of Infectious Diseases* 2010, **201**(3):435–443.

[111] Miller LH: **Distribution of mature trophozoites and schizonts of *Plasmodium falciparum* in the organs of *Aotus trivirgatus*, the night monkey**. *The American Journal of Tropical Medicine and Hygiene* 1969, **18**(6):860–865.

[112] Bignami AB: **Osservazioni sulle febbri malariche estive-autunnali**. *La Riforma Medica* 1890, **232**:1334–1335.

[113] Liu S, Mu J, Jiang H, zhuan Su X: **Effects of *Plasmodium falciparum* mixed infections on in vitro antimalarial drug tests and genotyping**. *The American Journal of Tropical Medicine and Hygiene* 2008, **79**(2):178–184.

[114] Ochola LB, Vounatsou P, Smith T, Mabaso MLH, Newton CRJC: **The reliability of diagnostic techniques in the diagnosis and management of malaria in the absence of a gold standard**. *The Lancet Infectious Diseases* 2006, **6**(9):582–588.

[115] **Methods and techniques for clinical trials on antimalarial drug efficacy: genotyping to identify parasite populations.** In *Organisation MfMVWH*, Amsterdam, The Netherlands 2008.

[116] Carvalho BO, Lopes SCP, Nogueira PA, Orlandi PP, Bargieri DY, Blanco YC, Mamoni R, Leite JA, Rodrigues MM, Soares IS, Oliveira TR, Wunderlich G, Lacerda MVG, del Portillo HA, Araújo MOG, Russell B, Suwanarusk R, Snounou G, Rénia L, Costa FTM: **On the cytoadhesion of *Plasmodium vivax*-infected erythrocytes**. *The Journal of Infectious Diseases* 2010, **202**(4):638–647.

[117] Lin E, Kiniboro B, Gray L, Dobbie S, Robinson L, Laumaea A, Schöpflin S, Stanisic D, Betuela I, Blood-Zikursh M, Siba P, Felger I, Schofield L, Zimmerman P, Mueller I: **Differential patterns of infection and disease with *P. falciparum* and *P. vivax* in young Papua New Guinean children**. *PloS One* 2010, **5**(2):e9047.

[118] Havryliuk T, Ferreira MU: **A closer look at multiple-clone *Plasmodium vivax* infections: detection methods, prevalence and consequences**. *Memórias Do Instituto Oswaldo Cruz* 2009, **104**:67–73.

[119] Schoepflin S, Valsangiacomo F, Lin E, Kiniboro B, Mueller I, Felger I: **Comparison of *Plasmodium falciparum* allelic frequency distribution in different endemic settings by high-resolution genotyping**. *Malaria Journal* 2009, **8**:250.

[120] Koepfli C, Mueller I, Marfurt J, Goroti M, Sie A, Oa O, Genton B, Beck H, Felger I: **Evaluation of *Plasmodium vivax* genotyping markers for molecular monitoring in clinical trials**. *The Journal of Infectious Diseases* 2009, **199**(7):1074–1080.

[121] Kirchgatter K, Portillo HAD: **Clinical and molecular aspects of severe malaria**. *Anais Da Academia Brasileira De Ciências* 2005, **77**(3):455–475.

[122] Smith T, Hii JL, Genton B, Müller I, Booth M, Gibson N, Narara A, Alpers MP: **Associations of peak shifts in age–prevalence for human malarias with bednet coverage**. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2001, **95**:1–6.

[123] Sama W: **Statistical Analysis of Within-host Dynamics of *Plasmodium falciparum* infections**. *PhD thesis*, University of Basel, Basel 2006.

[124] Gatton M, Cheng Q: **Modeling the development of acquired clinical immunity to *Plasmodium falciparum* malaria**. *Infect Immun* 2004, **72**:6538–45.

[125] Gysin J, Hommel M, da Silva LP: **Experimental infection of the squirrel monkey (*Saimiri sciureus*) with *Plasmodium falciparum***. *The Journal of Parasitology* 1980, **66**(6):1003–1009.

[126] Contamin H, Behr C, Mercereau-Puijalon O, Michel J: ***Plasmodium falciparum* in the squirrel monkey (*Saimiri sciureus*): infection of non-splenectomised animals as a model for exploring clinical manifestations of malaria**. *Microbes and Infection / Institut Pasteur* 2000, **2**(8):945–954.

[127] Carvalho LJ, Oliveira SG, Alves FA, Brígido MC, Muniz JA, Daniel-Ribeiro CT: ***Aotus infulatus*** **monkey is susceptible to** ***Plasmodium falciparum*** **infection and may constitute an alternative experimental model for malaria**. *Memórias Do Instituto Oswaldo Cruz* 2000, **95**(3):363–365.

[128] Garamszegi LZ: **The evolution of virulence and host specialization in malaria parasites of primates**. *Ecology Letters* 2006, **9**(8):933–940.

[129] Read A, Mackinnon M, Anwar M, Taylor L: **Kin selection models as evolutionary explanations of malaria**. *Adaptive dynamics of infectious diseases: in pursuit of virulence management* 2002, :165–178.

[130] Müller I, Genton B, Rare L, Kiniboro B, Kastens W, Zimmerman P, Kazura J, Alpers M, Smith TA: **Three different Plasmodium species show similar patterns of clinical tolerance of malaria infection**. *Malaria Journal* 2009, **8**:158.

Michael Bretscher - Curriculum Vitae

Names	Michael Thomas Bretscher
Date of birth	June 23, 1977
Nationality	Swiss
Marital status	married
Contact Information	LSHTM Keppel St, Room 280 London, WC1E 7HT United Kingdom *Phone:* +44 20 7927 2300 *E-mail:* mthbretscher@gmail.com

Languages: German (native speaker), English, French.

Research Interests: Epidemiology and physiology of malaria infection, the use of mathematical and statistical models for the analysis of cohort data, design of epidemiological studies, infectious disease dynamics

Education

London School of Hygiene and Tropical Medicine, London, UK

Post doctoral researcher, 2.2011 - present

- Project: Novel statistical methods for the epidemiological interpretation of *Plasmodium falciparum* antibody titers
- Advisor: Dr. Chris Drakeley, Immunology Unit, Immunology and Infection Department, Director of the **Malaria Centre** at LSHTM.
- Funding: personal grant ("prospective researchers") from the Swiss National Science Foundation .

Swiss Tropical and Public Health Institute, Basel, Switzerland

PhD in Epidemiology, 2006-2010

- Thesis: Statistical analysis of *Plasmodium falciparum* infection dynamics
- Advisor: Prof. Thomas Smith, Biostatistics and Computational Sciences Unit

Swiss Federal Institute of Technology, Zürich, Switzerland

MSc in Biology, 2003

- Thesis: Programming and analysis of a population genetic model on the emergence of multi drug resistance in HIV patients during Antiretroviral Therapy (ART)
- Advisor: Prof. Sebastian Bonhoeffer, Theoretical Biology Group

Courses

Swiss School of Public Health PhD Program:

PhD Course, August 31 - September 4, 2009. Observational epidemiological studies: Advanced methods for design and analysis, by Prof. Joel Schwartz, Harvard School of Public Health, USA

PhD Course, October 20-22, 2008. Spatial Statistics for Epidemiologists, by Prof. Dr. Leonhard Held, Institute for Social and Preventive Medicine, Zürich, Switzerland

PhD Course, April 2-4, 2008: Analysis of clustered data: multilevel modelling Course, by Dr. Martin Röösli, Institute for Social and Preventive Medicine, Bern, Switzerland

Technical Skills

- Programming: R, Winbugs, Java, C/C++, UNIX shell scripting, MySQL, SVN, and others.
- Process-based mathematical and statistical modelling.
- Experience in implementation of numerical algorithms for statistical analysis.
- Applications: LaTeX, Ms Office, OpenOffice, and others.

Hobbies playing guitar, travelling

Other

1998

Preparation and realization of a 3-month stay in a village near Bayanga, Central African Republic (CAR), among Bayaka pygmies. Lived in a leaf hut, went on net- and spear hunting trips. Perhaps the origin of an interest in tropical diseases.

2003-2005

Lost a father, travelled the world (two months in New Zealand, eight months through Africa), found a wife, and the way back to science.