Statistical properties of parasite density estimators in malaria and field applications

Imen Hammami

UNIVERSITY OF PARIS DESCARTES

GRADUATE SCHOOL
FRONTIERS IN LIFE SCIENCES

PHD THESIS

To obtain the title of

Doctor of Philosophy

In the subject of : **Biostatistics**

Defended by

Imen Hammami

Statistical properties of parasite density estimators in malaria and field applications

---

June 24^{th}, 2013

Thesis Defense Committee

Pr. Grégory Nuel	*Directeur de thèse*
Dr. André Garcia	*Directeur de thèse*
Pr. Christophe Rogier	*Rapporteur*
Dr. Neal Alexander	*Rapporteur*
Pr. Bernard Prum	*Examinateur*
Pr. René Ecochard	*Examinateur*

Acknowledgments

"Praise the bridge that carried you over."

George Colman

The PhD experience is not one of undisturbed peace and harmony. I can only say that those who had long hair at the beginning of their PhD ended up tearing most of it out. As humans naturally have pride, budding researchers, including me, may see the PhD as an endeavor completed in seclusion. However, this personal effort may be hindered by a series of obstacles and inconveniences and, knowingly, one cannot cope and survive without the help and support from others, insightful scientific discussions and collaborations, and valuable pieces of advice.

Gratitude - this is what I feel at this very moment, reaching the last page of the writing, to all those who made this thesis possible. Though there are no proper words to convey my deep gratitude and respect for all those people, I would still like to give my *sincere thanks* to all of them.

First and foremost, I owe an immense debt of gratitude to my PhD supervisor Grégory Nuel for his guidance, kindness, patience, enthusiastic encouragement, useful critiques of this research work and his assistance in keeping my progress on schedule. Obviously, it is difficult to overstate my gratitude to him. Whatever the reason you offered me this opportunity, thank you for believing in me.

His understanding, encouraging and personal guidance have provided a good basis for the present thesis. I would like to express my deep gratitude to my thesis supervisor André Garcia for his valuable support and constructive recommendations on this work. Without your useful help, the chapter about malaria would be liable for prosecution by the Court of malariologists.

My PhD thesis has been carried out at the "*Mathématiques Appliquées Paris 5*" Laboratory (MAP5), starting November 2009. I am extremely thankful to Pr. Annie Raoult, the Lab director, for hosting me, and to all my colleagues at the MAP5 for their support and friendship. I must also pay homage to Marie-Hélène Gbaguidi, our secretary, for facilitating the paperwork burden with her outgoing personality, tremendous kindness and indelible smile.

I am indebted to the "*Centre National de la Recherche Scientifique*" (CNRS) for the three years of generous funding. My special thanks are extended to my gradute school "*Frontières du Vivant*" (FdV 474) and the "*Bettencourt Schueller Foundation*" for providing me with further support during my research.

I remain grateful to the IRD UMR 216 team in *Cotonou* for providing us with data and for their warm and friendly welcome. A special *thank you* goes out to David Courtin and Gilles Cottrell for preventing the microscopists from hanging themselves.

I would like to express my very great appreciation to my tutors Pr. Bernard Prum and Dr. Jean-François Etard for their valuable and constructive suggestions during the Thesis Advisory Committee meetings. Their willingness to give their time so generously has been very much appreciated.

I am also indebted to the committee members of my thesis defense Pr. René Ecochard, Pr. Bernard Prum, Pr. Christophe Rogier, and Dr. Neal Alexander for accepting to review this work and for being part of the dissertation committee. I am honored by your interest in my research.

I am greatly indebted to the "*Département Santé publique et Biostatistique*" and the "*UFR Mathématiques et Informatique*" at the University of Paris Descartes for giving me the opportunity to teach a variety of mathematics courses, the best and most exhilarating moments for me.

I want to give a "*Big Thank You*", with a small taste of nostalgia, to my PhD mates Djeneba, Christophe and Lilia, who have always kept their sense of humour, cheerfulness and wonderful spirits even under adverse conditions. *Tipsee* you made my days a bit more sunny. *Mes quenelles* from *Paris* to *Tunisia*, thank you for being in my life. My brother and sister, Habib & Raja, no matter the distance, no matter what life brings, you're in my heart and always in my thoughts.

I dedicate this thesis to my parents. My love and gratitude for them can hardly be expressed in words, but still I hope these small words will find their way to their hearts. Thank you for your countless sacrifices, you unconditional love and support, and for always being my comfort and most steadfast shelter. I love and I admire you.

This thesis is also dedicated to the loving memory of *Sawsan Hakki* (1953-2013), a remarkable woman who gave a great deal of herself to her country and who had tragically died in the *Aleppo* University bombings. We will continue to honor your memory, while terribly missing you, and to pray for a future of peace, justice and freedom in *Syria* ... *Inch'Allah*.

Summary

Malaria is a global health problem present in over 109 countries in the world. According to the World Health Organization, malaria affected 219 million people worldwide and caused 660,000 deaths in 2010. Inaccurate estimation of the level of infection may have adverse clinical and therapeutic implications for patients, and for epidemiological endpoint measurements. The level of infection, expressed as the parasite density (PD), is classically defined as the number of asexual parasites relative to a microliter of blood. Microscopy of Giemsa-stained thick blood smears (TBSs) is the gold standard for parasite enumeration. Parasites are counted in a predetermined number of high-power fields (HPFs) or against a fixed number of leukocytes. The project was born out of a need to accurately and consistently assess the PD in epidemiological surveys, which is becoming increasingly important.

PD estimation methods usually involve threshold values; either the number of leukocytes counted or the number of HPFs read. However, the statistical properties of PD estimators generated by these methods have largely been overlooked. We studied the statistical properties (mean error, coefficient of variation, false negative rates) of PD estimators of commonly used threshold-based counting techniques depending on variable threshold values. We also assessed the influence of the thresholds on the cost-effectiveness of PD estimation methods. In addition, we gave more insights on the behavior of measurement errors according to varying threshold values, and on what should be the optimal threshold values that minimize this variability.

Furthermore, data on parasite and leukocyte counts per HPF are of broad scientific value. However, in published studies, most of the information on PD is presented as summary statistics (e.g. PD per microliter, prevalence, absolute/assumed white blood cell counts), but original data sets are not readily available. Besides, the number of parasites and the number of leukocytes per HPF are assumed to be Poisson-distributed. However, count data rarely fit the restrictive assumptions of the Poisson distribution. The violation of these assumptions commonly results in overdispersion. Undetected heterogeneity in parasite and leukocyte data may entail important misleading inferences, when they are related to other explanatory variables (malariometric or environmental), so its detection is essential.

We constituted and published the first dataset on parasite and leukocyte counts per HPF. The data comprise the records of three TBSs of 12-month-old children from a field study of Plasmodium falciparum malaria in *Tori Bossito, Benin*. All HPFs were examined systemically by visually scanning the film horizontally from edge to edge. The numbers of parasites and leukocytes per HPF were recorded. Pearson's test was used to check for overdispersion. Two sources of overdispersion in data are investigated: latent heterogeneity and spatial dependence. We accounted for unobserved heterogeneity in data by considering more flexible models that allow for overdispersion. Of particular interest were the negative binomial model (NB) and mixture models. The dependent structure in data was modeled with hidden Markov

models (HMMs). We found evidence that the Poisson assumptions are inconsistent with parasite and leukocyte distributions. Among simple parametric models, the NB model is the closest to the unknown distribution that generates the data. On the basis of model selection criteria AIC and BIC, the NB-HMMs provide a better fit to data than Poisson mixtures.

An alternative PD estimation method that accounts for heterogeneity and spatial dependence should be seriously considered in epidemiological studies with field-collected parasite and leukocyte data. We devised a reduced reading procedure of the PD that aims to a better operational optimization and a practical assessing of the heterogeneity in the distribution of parasites in TBSs. The motivations behind the design of this alternative protocol are the need to optimize the cost of epidemiological surveys and to reduce the inescapable loss of information. A patent application process has been launched in October, 2012. A prototype development of the counter is in process.

Keywords: Malaria epidemiology, threshold-based counting techniques, PD estimators, mean error, coefficient of variation, false-negative rates, cost-effectiveness, parasite and leukocyte counts per high-power field, Poisson distribution, overdispersion, heterogeneity, negative binomial distribution, mixture models, HMMs, AIC, BIC, patent.

Contents

List of Figures

List of Tables

List of Abbrevations

PD	Parasite Density
TBS	Thick Blood Smear
HPF	High-Power Field
P. falciparum	*Plasmodium falciparum*
WHO	World Health Organization
EM	Expectation-Maximization algorithm
HMM	Hidden Markov Model
NB	Negative Binomial
ME	Mean Error
CV	Coefficient of Variation
FNR	False Negative Rate

Chapter 1

Introduction

"The mere formulation of a problem is often far more essential than its solution, which may be merely a matter of mathematical or experimental skills. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science."

Albert Einstein

Contents

1.1 Background

Malaria is a global health problem present in over 109 countries in the world. According to the World Health Organization [WHO 2011b], malaria affected 219 million people worldwide and caused 660,000 [490,000-836,000] deaths in 2010. Despite these high numbers of malaria assignable fatalities, a reduction of malaria burden has been achieved during the last decade. The estimated incidence rates decreased by 17% worldwide between 2001 and 2010, and the estimated mortality rates decreased by 26% in the same period. This decline is due to unprecedented financial investments by governments and funders in malaria control and prevention.

While malaria burden has abated, history has shown that it is far too soon to cry victory in the battle against the disease. Malaria has been eradicated from 37 countries in 1961 [WHO 1959], but it has resurged with revenge in the following

years as elimination efforts weren't sustained. Malaria history is mostly told through numbers. These numbers, however, are a matter of controversy. Although both sides agree upon the decrease in malaria burden over the last decade, a recent study published in *The Lancet*, reports that the worldwide malaria deaths may be almost twice as high as previously estimated by WHO [Murray 2012]. The study claims that malaria caused 1,238,000 [929,000-1,685,000] deaths worldwide in 2010, and shows more deaths across all ages and regions than the WHO report [WHO 2011b]. The Murray et *al.* study estimated substantially more deaths in adults in Africa (8.11 fold higher than the WHO estimates for children aged five years or older in Africa). The WHO study, however, claims that this large number is unexpected in African countries, since acquired immunity developed at early ages will prevent adults from death [WHO 2012].

The huge discrepancy between Murray and WHO sets of estimates reveal alarming deficiencies at one or more of the following steps: data collection, reporting, analysis or interpretation. Morbidity and mortality statistics are based on clinical records, death certificates, verbal autopsy and governmental returns. However, each one of these direct factors are a potential source of error and may affect the accuracy of estimates. Many malaria cases occur in rural communities of sub-Saharan Africa that are seriously suffering from critical lack of healthcare facilities. Reporting systems in most of these areas remain poor and produce limited and imprecise informations. In addition, uncertainty about malaria statistics call into question the mathematical models used to generate them. Murray et *al.* claims that the WHO only takes into account the effect of vector control and population growth to estimate the mortality burden, and do not include the effect of drug resistance and increased use of effective malaria treatments, in addition to omitting environments and socio-economic factors. As stated in the latest WHO report [WHO 2012], the WHO mortality and morbidity estimation model takes into account the changes in intervention coverage, but ignores the possible changes in climatic conditions from year to year. The latter factor may directly influence malaria endemicity levels.

Another question arises from the differing set of estimates and the wild uncertainty ranges accompanying them as to the accuracy of the clinical records used in the two models. Available tools for malaria diagnosis are used to derive the parasite density estimates included in these models. However, diagnosis methods could be inaccurate and sources of misleading information [Amexo 2004, Reyburn 2004, Zurovac 2006]. To date, there is no standardized way to determine the parasite density. Methodological differences can make comparison and examination of overall trends very difficult. This question is even more relevant given that parasite density data are used at both individual and populational levels, respectively to detect malaria in clinically suspected patients and to assess the epidemiological characteristics of malaria as in previous examples. The importance of accurate parasite density estimation has been recognized for a long time [Christophers 1924, Wilson 1936, Earle 1939, Schwetz 1941, Parrot 1950, Wilson 1950, Miller 1958].

Light microscopy of Giemsa-stained thick blood smears is accepted as the current universal gold standard for the diagnosis of malaria. Great efforts have been devoted

to accurately diagnose the infection and assess the parasite density through microscopic methods [Ross 1910, Raghavan 1966, Trape 1985, Greenwood 1991]. Numerous counting methods have been proposed [Thomson 1911, Sinton 1924, Earle 1932, Boyd 1949, Field 1963, Russel 1963]. According to each estimation method, the blood smears are sampled in a specific way. A common way is to undertake a field-adapted sampling of a given amount of blood on the thick smear, thus suggesting an even distribution of the parasites on the slide. A second way involves making a thick smear with a known, small volume of blood [Warhurst 1996, Planche 2001] and then counting all the parasites on the smear. The parasite density is usually assessed either by counting parasites per high-power field (HPF) or by counting parasites per white blood cells [Wintrobe 1967, Trape 1985, Warhurst 1996]. These estimates usually involve threshold-based counting techniques (ex: if one see less than n parasites in the m first readings then do this, else do that...) which threshold values may vary a lot from a health organism or another. So little methodology, on which the evaluation of these methods depends, have been published. Only some empirical comparisons of parasite density estimation methods on thick blood smears are reflected in the literature [Dowling 1966, Greenwood 1991, Petersen 1996b, Dubey 1999, Planche 2001, Prudhomme O'Meara 2006b, Coleman 2006].

1.2 Goal Statement

The thesis project was born out of a need to accurately and consistently assess the parasite density in epidemiological surveys, which is becoming increasingly important. Current counting methods mislead or could mislead the microscopist, who will make his own subjective and possibly arbitrary understanding of the HPF limits (to avoid overlapping). Moreover, thresholds are fixed in a biased and discriminatory way, most of the time regardless of parasitemia levels and acceptable variability. It is of great importance for epidemiologists, who makes use of parasite density data, to know and assess the influence of these thresholds in the accuracy of their estimates. Health organisms or research teams also need to be aware of the impact the thresholds can have on the cost-effectiveness of the counting procedures (in term of time, and hence, of money). One of the common assumptions underlying these methods is that the distribution of the thickness of the smear, and hence, of parasites and leukocytes on the smear, is homogeneous. These distributions are modeled, most of the time, using the Poisson distribution [Student 1907, Petersen 1996a, Bejon 2006, Hammami 2013]; a hypothesis that wasn't supported by evidence from real data. In addition, variation of parasite density within a slide is expected even when prepared from a homogeneous sample [Alexander 2010]. The sampling variability is a source of interest when studying the efficiency of estimation methods and, then, it should be taken into account through appropriate statistical models. As cited before, parasite density measures are used in epidemiological models [Mwangi 2005, Becher 2005, Chandler 2006, Enosse 2006, Färnert 2009a, Damien 2010, Liljander 2011]. In the light of this, it is of primary

importance to know the consequences of the quality of these measures, when used as a covariate, on models outcomes.

1.3 Solution Statement

In the first part, we studied the statistical properties of parasite density estimators derived from four commonly used threshold-based counting techniques according to varying threshold values. For each estimator, we computed three measures of variability (mean error, coefficient of variation and false negative rates) and we assessed the cost-effectiveness of methods. Firstly, the exact distribution of the parasite density estimator is computed through recursive formulas. Secondly, based on this probability density function, measures of variability are derived. Finally, cost-effectiveness is defined for each method as the required number of HPFs that has to be read until the threshold is reached. The calculations are performed under two assumptions (1) the distribution of the thickness of the smear, and hence of the parasites within the smear, is homogeneous, and (2) the distribution of the parasites in the HPFs is uniform, and thus can be modeled through a Poisson distribution [Petersen 1996a, Kirkwood 2001, Alexander 2010].

An important step fulfilled in the second part of the project was the collection of parasite and leukocytes counts per high-power field in three thick blood smears (entirely examined). These counts allowed to investigate overdispersion in the distribution of parasites and leukocytes in the thick blood smear. We first considered the problem of testing whether our data comes from a single Poisson distribution. The basic null hypothesis of interest is that "variance = mean" (homogeneity hypothesis). We used the Pearson's test to testing the Poisson assumption. When the Poisson assumption is violated, we focus on alternatives that are overdispersed, in the sense that "variance > mean". We used the Kolmogorov-Smirnov (k.s) goodness-of-fit test [Chakravarti 1967] to test the validity of the assumed distribution for the data. In order to estimate models parameters, we performed a direct optimization of the log-likelihood. Model selection criteria are used to determine which of the simple parametric models best fits the data. Secondly, we investigated the first source of overdispersion in count data, which is unobserved heterogeneity. We explored the unobserved heterogeneity among parasite and leukocyte data using mixture models. The motivation behind the use of mixture models is that they can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in data. The objective here is to describe the data as a finite collection of homogeneous populations on thick blood smears. The form of these sub-populations was modeled using Poisson and negative binomial distributions. Thirdly, we considered the second source of overdispersion, which is positive contagion [King 1989] (e.g. high number of parasites in one HPF lead to correspondingly high numbers of parasites in neighboring HPFs or low number of parasites in one HPF drive down counts for other neighboring HPFs). This data-generating process may have important implications for the observed level of dispersion in data. We check for

spatial dependence in data using autocorrelation plots [Box 1976b]. We used hidden Markov models (HMMs) to account for this autocorrelations, since HMMs are an extension of mixture models with spatial dependence taken into consideration. The state-dependent distribution was modeled using Poisson and negative binomial distributions. The proposed mixture models and HMMs were fitted by maximum likelihood using the EM algorithm, and validated by direct numerical maximization.

1.4 Application

An important step, fulfilled at the first beginning of the project, was the field study in Benin. It enlightened my way of understanding the problem of the parasite density estimation. I was able to approach, in practical terms, the parasite density estimation constraints faced by microscopists. I noticed the weakness of existing estimation methods in the matter of variability, left-censoring and cost-effectiveness, which suggested possible improvements by designing more efficient and cost-effective alternative procedures. We attempted to accomplish this by developing, implementing and evaluating a new counting method. The motivations behind this design are (1) the economic and operational optimization of field studies, and (2) the practical assessment of the heterogeneity in the distribution of parasites and leukocytes in the thick blood smears.

This new device is an appropriate protocol for field experience, since it requires neither special equipments, nor operator decisions that might bias the outcome. This methodology is potentially useful for laboratories that routinely perform malaria parasite enumeration. It requires less work load from field operators, and then allow to examine more slides with the same manpower, thus optimizing field operations. Furthermore, this method allows for heterogeneity detection and is proved at least as accurate and precise as existing threshold-based counting techniques.

1.5 Thesis Outline

The present chapter briefly describes what the thesis is about and how the design for this research project has been growing throughout the last three years. An outline of the major themes and questions, the main points to be made about each one and the statistical evidence of each one are presented.

Chapter 2 reviews the past and current burden of malaria, and defines the context where the problem lies. This chapter reflects my own understanding of malaria burden, as a non-expert, and does not give an account to the complex *Plasmodium* biology, which goes far beyond the scope of this chapter.

Chapter 3 highlights the importance of the parasite density estimation. It is also devoted to a summary and discussion of the commonly used threshold-based counting techniques.

Chapter 4 studies the statistical properties (mean error, coefficient of variation, false negative rates) of parasite density estimators derived from these methods and

depending on variable threshold values. This part of the dissertation also assesses the influence of the thresholds on the cost-effectiveness of parasite density estimation methods. In addition, it gives more insights on the behavior of measurement errors according to varying threshold values, and on what should be the optimal threshold values that minimize this variability. This chapter was the subject of Article 1.

Chapter 5 describes the EM algorithm for mixture models and HMMs, where observations are Poisson- and negative binomial-distributed.

Chapter 6 describes the first open-source dataset on parasite density per HPF. The problem considered in this chapter is to test whether parasites and leukocytes are spread evenly throughout the film. Unobserved heterogeneity in the data is accounted for by considering more flexible models that allow for overdispersion. This chapter was the subject of Article 2.

Chapter 7 summarizes and discusses the key points from the previous chapters and provides suggestions and insights for their possible improvements.

1.6 Scientific Publications & Communications

Scientific Publications

Article 1 Imen Hammami, Grégory Nuel, André Garcia. **Statistical properties of parasite density estimators in malaria.** *PLoS ONE*, **8:**e51987, 2013.

Article 2 Imen Hammami, André Garcia, Grégory Nuel. **Evidence for overdispersion in the distribution of malaria parasites and leukocytes in thick blood smears.** *Malaria Journal*, **12:**398, 2013.

International Communications

- Statistical properties of parasite density estimators in malaria.
  ICSA Applied Statistics Symposium 2011, New York, USA.
- Statistical properties of parasite density estimators in malaria and field applications. *43ième Journées de Statistique de la SFDS 2011, Tunis,Tunisia.*
- Evidence for overdispersion in the distribution of malaria parasites and leukocytes in thick blood smears.
  XXVIth International Biometric Conference 2012, Kobe, Japan.
- A new method to reliably count parasites in thick blood smears.
  Infectious Disease Week 2012, San Diego, USA.

Others:

- *ESOF - Euroscience Open Forum 2010, Torino, Italy.*
- *Paris Interdisciplinary PhD Symposium 2011, Paris, France.*

Patent

In order to protect the intellectual property related to the method briefly mentioned in (1.4), a patent application process has been launched in October, 2012.

A prototype development of the counter is in process. Public (written or oral) disclosure of the invention prior to the filing of the original patent application may invalidate the patent. For this reason, details and specifications of the invention are not included in this manuscript.

1.7 Interdisciplinarity

Even if this project is first a statistical one, it has been developed in close collaboration with the University of Paris Descartes and the UMR 216 IRD team, which is specialized in malaria. UMR 216 provided field data, valuable feedback and practical field experiment of the alternative estimation methodology. This collaboration added a real interdisciplinary dimension to the project. Besides, this project has been developed in close collaboration with the "*Faculté des Sciences de la Santé*" (FSS), University of Abomey-Calavi, *Cotonou*, *Benin*. This partnership have been essential, since it provided expertise on malaria epidemiology, operational experience of the data collection, availability of real data sets and field experiments.

The project covers several disciplines while being solidly founded within the field of biostatistics. It has the particularity to start from the field, to go through innovative statistical developments, before returning to the field. Besides, it focuses on the problem of malaria, which remains, to date, a major public health problem in the developing countries. From a statistical point of view, it deals with a problem as appealing as critical, particularly regarding the quality of the threshold-based estimators commonly used in parasite density estimations. Finally, it should be noted that this project naturally fits one of the priority axes of research of the University of Paris Descartes, namely the "*Institute for the Development and the International solidarity*".

Chapter 2

Malaria

"Worse than the sun in March.
This praise doth nourish agues."

Shakespeare, *Henry IV*

Contents

This chapter is dedicated to the general characteristics of malaria. They are described in broad strokes without going deep into the complex biology of malaria parasites. Nevertheless, this brief overview is sufficient in itself, whatever else might be added, to draw the attention to the longstanding issues surrounding this potentially lethal disease, and to define the context where the problem lies. The World Health Organization and its partner, Roll Back Malaria, offer exhaustive web-based portals that give access to a wide range of informations for people who are eager to learn more about malaria.

The word malaria (from Italian origin) hides a part of its history. The contraction of *"mala aria"* means *"bad air"*. We owe this name to the Italian physician

Francesco Torti (1658-1741). Occurrences of this word can be found in numerous Shakespearean plays. In Shakespeare's time, people thought that the sun raise up the bad air in marshy areas that caused symptoms of *ague* (what we now call malaria). This thought continued on until 1898, when Ronald Ross (1857-1932) discovered that malaria was a mosquito-vectored disease.

2.1 Introduction

Malaria, a widespread and potentially fatal infectious disease, has wreak havoc on our world for much of human history. Malaria history can be depicted through unprecedented findings that has been, most of time, rewarded with remarkable Nobel prizes. Since the discovery of the parasite causing malaria in 1880, researches have been carried out in science and medicine for hundreds of years, and have expanded, considerably, from preventive and treatment strategies to include a better understanding of its biology. Within a short time frame (from 1880 to 1899), basic knowledge of malaria has been produced at a fast pace.

Nevertheless, more than a century after the first discoveries, the scourge is still present and malaria situation is still showing a daunting figure. Malaria, once triumphantly held to be eradicable, causes approximately one million deaths each year. Many of the antimalarial drugs are losing effectiveness, as the parasite evolves high levels of drug resistance. Besides, no effective malaria vaccine has yet been developed. The world's first potential malaria vaccine proved only 16.8% effective over the four-year period, calling into question whether it can be a useful weapon in the fight against the deadly disease.

This unblemished record of failure leaves humans bewildered and depressed. As malaria remains a major public health problem, understanding its history is a key to address some important questions concerning the present situation.

2.2 History of malaria

Malaria preciously concealed some of its parasitological secrets, so that a complete century was needed to score the first (temporary) victory of humanity against the disease in 1961. A brief overview of the recent history of malaria by the end of the nineteenth century show important breakthroughs that enlarged the understanding of the disease, and provided important weapons to fight malaria-causing parasite. The roots of this process of scientific discovery began in the 1880s.

Earlier theories postulated that malaria was caused by bad air ("*mala aria*") from marshy areas. However, the hypothesis of a bacterial origin of malaria became increasingly attractive after the discoveries of Louis Pasteur that most infectious diseases are caused by microbial germs, known as the "germ theory".

However, as deaths due to malaria were frequent in the army by 1880, Charles Louis Alphonse Laveran (1845-1922), a French army surgeon, studied the disease's clinical aspects and its anatomic pathology. While he examined the blood of a

patient who had been febrile for 15 days, Laveran saw "... *on the edges of a pigmented spherical body, filiform elements which move with great vivacity, displacing the neighboring red blood cells*". He also saw the exflagellation of a male gametocyte, a phase in the life cycle of malaria parasites which usually occurs in the stomach of the Anopheles mosquito. These findings convinced him that he had discovered the malaria agent, which is a protozoan parasite. Hence, Laveran was the first to notice parasites in the blood of a patient suffering from malaria in 1880. During the following years, Laveran looked for the parasite in the environment surrounding the human host (air, water, soil of marshy areas ...). His efforts were unsuccessful, which makes him suspect that the parasite could develop inside the body of the mosquito. However, the quest to prove this hypothesis took him years. Laveran's findings were generally met with skepticism, especially among Louis Pasteur' disciples, the defenders of a bacterial cause hypothesis. In 1884, Laveran shared his discovery with Pasteur who was immediately convinced (Roux, 1915). Few years later, the parasitic origin of malaria was accepted. Laveran was awarded the Nobel Prize in Physiology or Medicine in 1907 for his discovery.

In the years 1886-1892, Camillo Golgi (1843-1926), an Italian neurophysiologist, provided fundamental contributions to the study of malaria. He elucidated the cycle of the malaria agent in red blood cells, and distinguished two forms of the disease. He found correlation between febrile episodes and the release of parasites into the blood stream. Golgi was awarded a Nobel Prize in Physiology or Medicine in 1906.

At the beginning, Laveran had believed that there was only one species, *Oscillaria malariae*. Other species were discovered during the following years. In 1890, *P. vivax* and *P. malariae* was revealed by the Italian investigators Giovanni Battista Grassi (1854-1925), and Raimondo Filetti. In 1897, William Henry Welch (1850-1934), reviewed the discovery made by Laveran and described the malignant tertian malaria parasite *P. falciparum*. *P. ovale* was discovered by John William Watson Stephens (1865-1946) in 1922, while *P. knowlesi* was first described by Robert Knowles (1883-1936) and Biraj Mohan Das Gupta (1885-1956) in 1931 in macaques. It was not until 1957 that Garnham (1901-1994) et al. suggested that *P. knowlesi* could also cause malaria in humans.

In 1878, Patrick Manson (1844-1922), the father of tropical medicine, formulated his theory of mosquito transmission. Manson claimed that a parasite that causes human disease could be spread by a mosquito. He had discovered that the filariae of sufferers from *Lymphatic filariasis*[1], commonly known as *elephantiasis*, were ingested by mosquitoes. Along the same lines of the latter discovery, Manson had suspectd the mosquito to be a vector for malaria, and that the exflagellation, described by Laveran, could not take place within the bloodstream, and requires moisture and a lower temperature outside of the human body, such as within the stomach of mosquitoes.

Manson's hypothesis inspired Ross and dissipated his early doubts about the

1. a disease characterized by the thickening of the skin and underlying tissues, and can result in an altered lymphatic system and the abnormal enlargement of body parts, causing pain and severe disability.

existence of the parasite. On the basis of the Laveran's discovery, Ross studied the trasmission of protozoan from the mosquito to the host. In 1898, Ross identified the mosquito species Anopheles as the carrier of malaria. His studies led to ground breaking discovery of the life cycle of malaria parasite, which earned him the Nobel Prize in Physiology or Medicine in 1902.

Following Ross's discoveries, a series of innovative breakthroughs pioneered new routes and set new goals in malaria research. Robert Koch (1843-1910) favored the mosquito-malaria theory and confirmed Ronald Ross's discovery of malaria parasite's life cycle. He was awarded the Nobel Prize in Physiology or Medicine in 1905. In 1899, Giovanni Battista Grassi, Amico Bignami (1862-1929) and Giuseppe Bastianelli (1862-1959) demonstrated that the Anopheles mosquito carries the plasmodium of malaria in its digestive tract, and they determined the complete sporogonic cycle of *P. falciparum*, *P. vivax*, and *P. malariae*.

The very first initiatives in malaria control concerned war areas. A chapter in the *The Prevention of Malaria* book by Ronald Ross, entitled *The Prevention of Malaria in War* [Melville 1910] presents malaria as the most stubborn enemy of American soldiers during wars. During World War I and World War II, important financial investment devoted to malaria control have been made to supply soldiers with existing anti-malarial agents and to develop alternative drugs.

Peruvian Indians were the first to use the original antimalarial agent called *quinine* in the 17th century. The drug is extracted from the bark of the cinchona tree in Peru mountains. During the Spanish colonization of the Americas, the Spanish discovered the miraculous cure and used it to protect soldiers in malaria-prone countries. Until World War I, *quinine* was the only effective treatment for malaria. However, the increased need of *quinine* after the colonization of malarious countries caused inconsistent supplies of this natural anti-malaria drug from South America. Hence, it was critical, at that point of time, to develop a synthetic alternative to this drug. By 1944, artificial syntheses of *quinine* were developed, but none of them have been as economically viable as the natural anti-malarial drug. Due to a number of limitations, including drug resistance, ineffectiveness against *Plasmodium* gametocytes and side effects, the *quinine* was not a permanent solution to cure malaria [White 1999a]. Hence, alternative drugs (such as *chloroquine* and *primaquine*) replaced *quinine* during World War II. More recently, *artemisinin*, which had been used for more than two thousand years in traditional Chinese medicine in the treatment of many diseases including malaria, has become the treatment of choice for malaria in 2006.

Another turning point in the fight against malaria is the discovery of the *insecticide Dichloro-diphenyl-trichloroethane* (DDT) in 1942 by Paul Hermann Müller (1899-1965), the Nobel Prize Laureate in Physiology or Medicine in 1948. DDT may be viewed, through history, as a double-edged sword. At first, it was hailed as a miracle weapon, which has be proven to be very effective against malaria-carrying mosquitoes. This potent pesticide was cheap and long-lasting. This had made its use appealing in World War II. DDT had also initiated The Global Malaria Eradication Program in 1955 that showed (temporary) promising results. The program

was based on two fundamental activities: (1) the treatment of infected individuals with *chloroquine* and (2) the use of DDT for mosquito control [WHO 1959]. These global efforts to eradicate malaria successfully eliminated the disease from 37 of 143 endemic countries by 1961. Despite initial success, elimination efforts weren't sustained in the following years, which caused the resurgence of malaria. This is in a great part due to the emergence and widespread of DDT resistance. In countries committed to the use of DDT, governments were so caught up in the maelstrom of the unequal battle against malaria, that they ignored the toxicity of DDT and failed to sustain an efficient application of control measures.

At the beginning of the twentieth century, the humanity had access to important informations about malaria (agent, vector, *Plasmodium* life cycle, prevention and treatment). One century after, the humanity stands at such a point in the history of malaria with an overwhelming evidence of failure. The bottom line is that the scientific community does not know enough about the very complex malaria disease. They are back at square one in knowing how to eliminate the parasite due to anti-malarial drug resistance. Hence, square two would be a heck of a long way off. The contradiction between the avalanche of discoveries between 1880 and 1899, and this record of failure leads us back to the initial question whether it can be a useful weapon in the fight against the disease. The most pessimistic voices emerging from the scientific community describe the vaccine research as a scientific boondoggle. They claim that investing more in malaria control and prevention is crucial and can be sufficient to eliminate malaria. Effective control measures, however, rely on accurate estimation of malaria endemicity to bring to completion malaria "eradication". Here too, there is a problem of imprecise statistics.

2.3 Epidemiology of malaria and Plasmodium species

The ravages of malaria have been etched into human history. It is claimed that the devastating disease has killed half of people who have ever lived [Finkel 2007]. As mentioned in the previous section, the complexity of the disease is, in part, responsible for the failure in control. In this section, we shall look at the question of the heterogeneity in the distribution of malaria burden and species. This heterogeneity adds more complexity and undermines the overall situation.

A combination of specific climate features directly influences the distribution and the seasonality of malaria, and allow the existence of an endemic malaria transmission. The tropical climate characteristics (high temperatures, humidity and abundant rainfall year-round) favors the presence of malaria in tropical regions. The standing water spots after rainfalls provide mosquitos with a suitable environment, in which they can breed and mature [Jamieson 2006]. In drier regions, mapping rainfall episodes allow to predict effectively, and quite accurately, outbreaks of malaria [Abeku 2007]. However, the climate-disease model, on its own, is insufficient to account for the heterogeneous distribution of malaria endemicity.

An effective global strategy for malaria control needs accurate estimates of

malaria endemicity. In order to better understand endemicity, and to support malaria control planning, many projects provided a detailed mapping of malaria risk and endemicity in Africa. These mapping are based on accurate estimates of the burden of malaria at regional or district level. In 1999, the MARA/ARMA project has provided the first continental maps of malaria distribution in Africa [Craig 1999]. More recently, Malaria Atlas Project focussed on analysing climate and weather information required to accurately predict the spread of malaria and to provide contemporary and robust tools to assess malaria burden [Guerra 2007]. This effort led to the publication of a map of *P. falciparum* endemicity in 2010 (Figure 2.1) that shows the global spatial limits of the disease.

Using the term malaria as a whole hides a more complex species heterogeneity. Five malaria species are known to cause malaria in humans: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and *P. knowlesi*. Prevalences of the five *Plasmodium* species are heterogeneously distributed.

P. falciparum has its origins in West Africa and *P. vivax* appears in West and Central Africa. Human migrations led to the global spread of the disease and malaria becomes the world's deadliest disease. *P. falciparum* and *P. vivax* are the most common forms of malaria. *P. falciparum* is the most life-threatening species of malaria. This species is present accros much of tropical and subtropical areas, but it is much more prevalent in sub-Saharan Africa. *P. falciparum* showed resistance to many antimalarial drugs. As a consequence, *P. falciparum* reemerged with a vengeance in some areas, where it was thought to have been eradicated [Sharma 1996, Rab 2001, Faulde 2007, Wangai 2011].

P. vivax is the most geographically widespread of the human malaria species. It accounts for 100-300 million cases in much regions of South-East Asia, The Americas and The Middle East, where the wide majority of malaria burden is caused by this species.

P. ovale is generally not fatal. Its spread is restricted to tropical Africa, New Guinea, and the Philippines. Symptoms are similar to those of *P. vivax*. *P. ovale* and *P. vivax* can remain in a dormant stage in the liver, without causing illness. If untreated, they can cause relapses (malaria attacks) months or years after the first infection.

The sporadic presence of *P. malariae* has been reported in Africa, India, South America and Western Pacific. The diagnosis of this species is difficult due to the presence of very few parasites in the blood. If untreated, latent *P. malariae* infections may be present for many years. *P. vivax* and *P. malariae* were the most widespread forms of malaria in the past, but *P. malariae* has lost its predominance, and it is now less common than the other forms.

P. knowlesi, similar morphologically to *P. malariae*, has been identified by molecular methods in patients in Malaysia, the Philippines, Thailand, and Myanmar. *P. knowlesi* causes malaria in macaques. It has not yet been proven to be transmitted from humans to mosquitoes. A monkey reservoir may be required to infect mosquitoes. High levels of *P. knowlesi* infections may lead to organ failure or death.

Estimates by WHO of the number of cases and deaths from malaria from 2000

to 2010 were published in the World Malaria Report 2011 [WHO 2011b]. In 2010, the WHO estimated that 219 million people worldwide are affected by malaria and that 660,000 people died from the disease. Malaria affects over 109 countries in the world. The disease is presently endemic in countries along the Equator, in The Americas, in South East Asia regions, however, the vast majority of estimated cases (80%) and deaths (91%) occur in sub-Saharan Africa, where *P. falciparum* is by far the deadliest of human malarias (85% of deaths). Children under 5 years old accounts for the majority of deaths (86%).

Precise statistics are unavailable. A recent study, published in *The Lancet*, reports that the worldwide malaria deaths may be almost twice as high as previously estimated by WHO [Murray 2012]. The Murray et al. systematic analysis of 1980-2010 global malaria mortality brings discredit to the WHO estimates. The study claims that malaria caused 1,238,000 [929,000-1,685,000] deaths worldwide in 2010 in comparison to the WHO estimates of 660,000 [490,000-836,000]. Moreover, the study shows that mortality is higher across all ages and regions than the WHO report (1.3 fold higher for children under five years old in Africa, 8.11 fold higher for children aged five years or older in Africa, and 1.8 fold higher for all ages outside of Africa).

Murray claims that the WHO only takes into account the effect of vector control and population growth to estimate the mortality burden, and do not include the effect of drug resistance and increased use of ACT, in addition to environments and socio-economic factors. In the latest report [WHO 2012], WHO states that the two sets of estimates are not significantly different since the ranges overlap (for deaths in people under 5 years old in Africa, deaths in people 5 years old outside Africa, and deaths in people aged five or older outside Africa) with the exception of deaths in people aged five or older in Africa. According to WHO, the large number of deaths in people aged five or older relative to those under five years old estimated by Murray et al. is unexpected in African countries, since acquired immunity developed at early ages will prevent adults from death. A study published in 2011 by [Cibulskis 2011], based on confirmed microscopic diagnosis, shows that the adult-to-child death ratios in much lower. However, Murray et al. maintain, on the basis of vital registration, verbal autopsy and hospital data, that an important number of deaths occur in people aged 15 years or older in sub-Saharan Africa contrary to what have been stated. Both studies, however, agree that investments made by governments and funders, although affected by the global crisis, have substantially decreased the burden of malaria in the last five years. If these efforts are maintained, malaria mortality will fall below 100,000 in 2020.

Due to the dramatic consequences these substantial discrepancies can have on political decisions, it may be worthwhile to question the reasons of this heterogeneity. Huge differences between Murray and WHO sets of estimates reveal alarming deficiencies in current data collection, reporting, analysis and interpretation. Morbidity and mortality statistics are based on clinical records, death certificates, verbal autopsy and governmental returns. However, each one of these direct factors are a potential source of error and may affect the accuracy of estimation. Many malaria

cases occur in rural communities of sub-Saharan Africa that are seriously suffering from critical lack of healthcare amenities. Reporting systems in most of these areas remain poor and produce limited and imprecise informations. Verbal autopsy is also an imprecise estimator of malaria mortality, since it is subjective and unable to distinguish severe malaria from other febrile illnesses. Available tools for malaria diagnosis and parasite density estimation methods can also be sources of misleading information. Indeed, diagnosis methods could obviously be inaccurate, and there is no standardized way to determine the parasite density. Methodological differences can make comparison and examination of overall trends very difficult. These issues are discussed in more details in Chapter 3. Finally, much uncertainty exists about malaria statistics, which call into question the mathematical models used to provide them. Important resources are invested in malaria control and prevention campaigns, but there are no reliable tools to assess their effectiveness.

2.4 Life cycle of Plasmodium

The agent responsible of the disturbing number of deaths, although subject of controversy, is the *Plasmodium* protozoan. Understanding the parasite life cycles is of primary importance, since it predicts the parasite's involvement in disease, and gives informations about the disease pathogenesis and clinical signs. The parasite life cycle also provides information of epidemiological significance and, then, facilitates the development of control and prevention strategies. A detailed understanding of parasite life cycles is also needed before any discussing about a potential malaria vaccine, as a vaccine could act at one or many stages during the parasite life cycle.

The life cycle of *Plasmodium* involves several stages, including sporozoites (the infectious form injected by the mosquito), merozoites (the stage invading the erythrocytes), trophozoites (the form multiplying in erythrocytes), schizonts (found (1) in the liver; when sporozoites are mature, (2) within erythrocytes; when trophozoites mature and divide), and gametocytes (sexual stages).

The mosquito injectes parasites (sporozoites) into the human host. Parasites travel into the bloodstream to the liver, where they mature and release another form of the parasite (merozoite). The merozoites leave the liver, and invade red blood cells (RBCs). In RBCs, merozoites reproduce and develop into trophozoites and schizonts, which, in turn, produce more merozoites. New merozoites burst out and seek for new RBCs to infect. A small proportion of asexual parasites differentiate in the human bloodstream into sexual erythrocytic stages (gametocytes), which are infectious to mosquitoes. Parasite transition from host (the human) to vector (the mosquito) is mediated by gametocytes. The gametocytes, which are the transmissible parasite form, are taken up into the mosquito when it feeds. Soon after the blood meal, gametocytes rapidly undergo sexual reproduction in the midgut of the mosquito, and create sporozoite forms, which are infectious to humans. Hence, the life cycle of the parasite is completed (Figure 2.2).

As the parasite invades RBCs, malaria can be transmitted through organ trans-

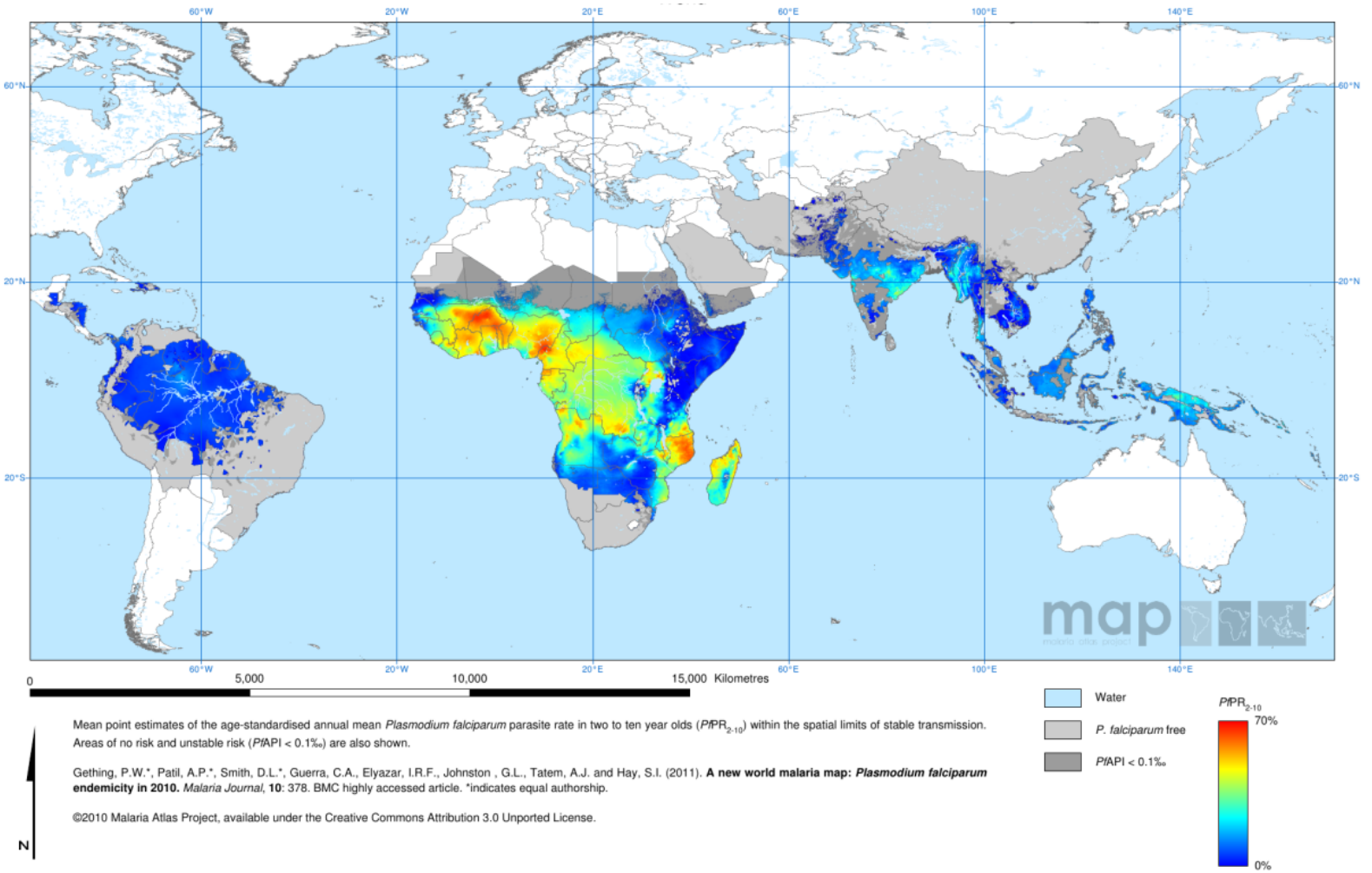

Figure 2.1: ***The spatial distribution of P. falciparum malaria endemicity map in 2010 globally.***
This map shows estimated levels of *P. falciparum* malaria endemicity [Gething 2011]. The mapped variable is the age-standardised P. falciparum Parasite Rate ($PfPR_{2-10}$) [Smith 2007], which is an index of malaria transmission intensity based on the estimated proportion of the two to 10 year olds infected with *P. falciparum* in the general population in 2010. Unstable transmission areas (where risk is very low) are located in Asia (91%), the Americas (5%) and Africa (4%). Stable transmission areas (where risk is high) are located in Africa (52% of the global total), Central, South and East Asia (46%) and a smaller proportion in the Americas (2%) The highest levels of *P. falciparum* transmission risk are present in Africa (99% of the total area), where $PfPR_{2-10} \geq 40\%$ for 95% of the population. *source:* Malaria Atlas Project (map)

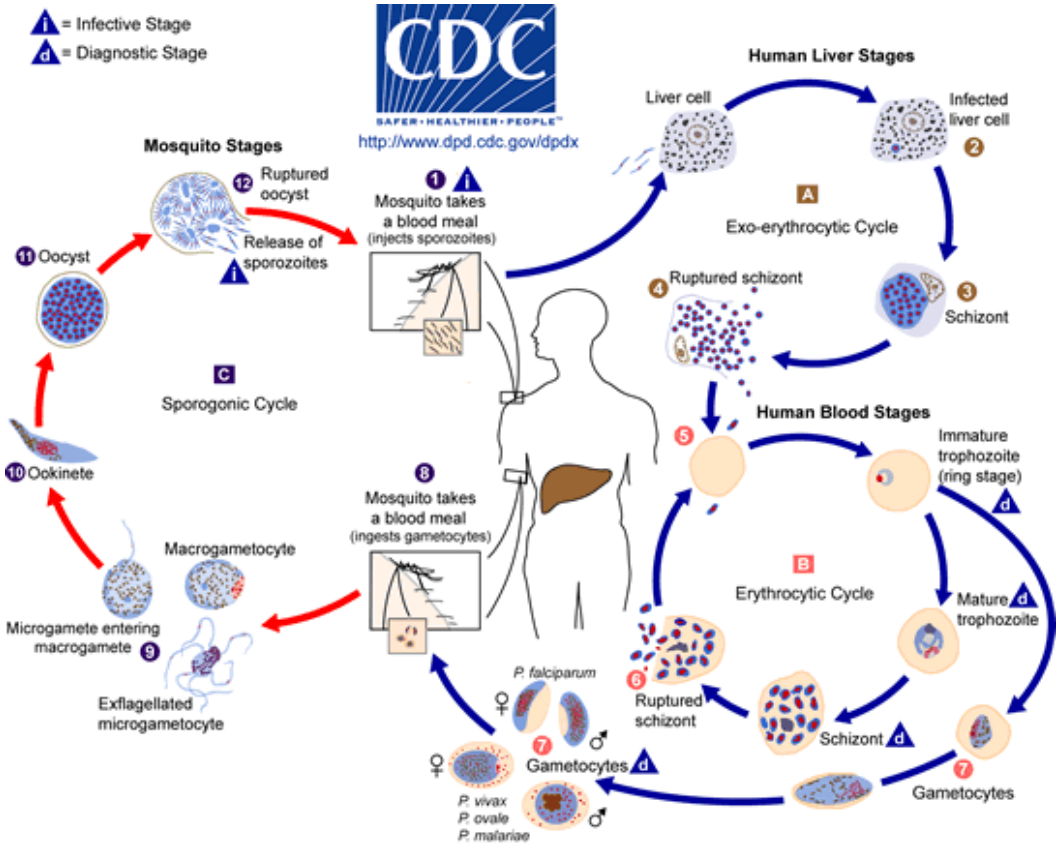

Figure 2.2: ***Life cycle of Plasmodium falciparum.***
Three primary stages can be distinguished according to whether parasites are in the liver, the bloodstream or the mosquito's stomach. [A] Human liver stages (*exo-erythrocytic cycle*) (1) The female anopheles mosquito injects parasites (sporozoites) into the human after a blood meal. Sporozoites travel into the bloodstream to the liver. (2) Sporozoites invade the liver cells (hepatocytes). (3) Sporozoites mature into schizonts. (4) Schizonts rupture and release merozoites. [B] Human blood stages (*erythrocytic cycle*) (5) Merozoites leave the liver and invade red blood cells, in which they reproduce and develop into trophozoites. (6) Trophozoites mature into schizonts, which in turn rupture and release merozoites. (7) A proportion of parasites differentiate into gametocytes (sexual forms). [C] Mosquito stages (*sporogonic cycle*) (8) Gametocytes are taken up by a mosquito when it feeds. (9-12) Gametocytes undergo sexual reproduction in the midgut of the mosquito and develop into sporozoites, which migrate to the salivary glands. (1) The mosquito injects sporozoites through saliva into another human. *source:* Centers for Disease Control and Prevention (CDC)

plant, shared use of syringes, and blood transfusion. Infected mothers transmit parasites to their child during pregnancy before or during delivery (birth), which is known as "congenital malaria".

2.5 Signs and symptoms

The reproduction of *Plasmodium* parasites in the liver, and their spread into the bloodstream, produce typical malaria symptoms. As shown in Figure 2.2, parasites invade RBCs in the bloodstream. Some of them stay and multiply in the liver and periodically send more merozoites into the bloodstream. This action comes with repeated bouts of symptoms. The patient develops a high fever. The frequency of febrile and afebrile episodes depends on the species causing malaria.

People suffering from uncomplicated malaria will complain (in addition to fever) of headache, nausea, shaking chills, sweating and weakness. Anemia is common in patients with malaria, in part due to the destruction of the RBCs. In countries with limited malaria transmission, these symptoms may be attributed to more common infections, or simply to cold or influenza. On the contrary, in countries with high risk of malaria, the symptoms are, often, indiscriminately recognized and treated as malaria without evidence-based clinical diagnosis.

Severe malaria happens when *P. falciparum* infections are compounded by serious failure of the body's major organs or by metabolic disorders. The manifestations of severe malaria include (in addition to fever) severe anemia, kidney failure, pulmonary edema, cerebral malaria, coma, or death.

The high risk populations for malaria are pregnant women and children under 5 years old. Symptoms in children may be nonspecific, leading to delays in diagnosis. Pregnant women are particularly vulnerable to the burden of malaria, since malaria can make worse a pre-existing anemia and causes prematurity, spontaneous abortion or stillbirth. Moreover, some genetically determined parasites are sequestred in the placenta causing what is classically called "placental malaria" infection (PMI) or pregnancy-associated malaria (PAM). This PMI is primarily caused by *P. falciparum* and is responsible of low birth weight (LBW), which is associated with a higher risk of mortality during the first months of live [Jelliffen 1968, Guyatt 2004, Aribodor 2009, Tiono 2009]. Moreover, children born of mother with PMI are more vulnerable to malaria infections during their first months of life [Le Hesran 1997, Le Port 2011a].

In tropical areas, multiple infections (helminths, dengue,tuberculosis, filariasis, typhoid, schistosomiasis, leptospirosis, HIV...) is the general rule and not the exception. Co-existence of several infectious agents within the same host has consequences on immunity against those infections. In such a situation, the host immune system interacts with the different infectious agents leading, in some cases, to modified clinical manifestation (classical clinical signs can be exacerbated or moderated). In the particular case of HIV, co-infection with malaria causes increased mortality. HIV-infected individuals show increased vulnerability to malaria infec-

tion [Cohen 2005, Patnaik 2005, Kamya 2006]. Acute malaria is associated with an increase in HIV viral load [Kublin 2005], and a more rapid decline in CD4 cell count [2] [Mermin 2006]. These interactions adversely impact the outcome of both diseases, and may facilitate both diseases spread.

Once clinical symptoms appear, the development of the disease depends in part on the host's immunity and on the treatment received. Immunity specific to malaria infection is a particularly complex aspect of the disease that goes far beyond the scope of this thesis. It will be briefly addressed in the following section.

2.6 Immunity against malaria

Following infection, the clinical manifestations of malaria may be acute in non-immune patients, and the infection may progress to severe disease and death. With more infections, anti-disease immunity develops and eliminates clinical symptoms, and also decreases the risk of severe disease. The human body's defenses against infection with malaria fall, in general terms, into innate (non-specific) resistance and acquired (specific) immunity.

The innate resistance is constitutively present in the host and does not depend on any previous infection. Innate immunity is activated immediately upon infection. It is termed non-specific because the protective response is the same regardless of the initial infection. This is in contrast to the adaptive immunity which is slower, responds specifically, and generates an immunological memory. Acquired or adaptive immunity against malaria develops after infection with multiple strains of malaria [Färnert 2009b]. Its protective efficacy depends on several factors as the host, his environment and the number of infections.

In man, acquired immunity against plasmodia is a mixture of anti-disease immunity, anti-parasite immunity and premunition (or sterilizing immunity). Anti-disease immunity confers protection against clinical disease and affects the disease morbidity. Anti-parasite immunity confers protection against high parasitemia and affects the density of parasites. Premunition is due to persistent latent infection. It protects against new infections by maintaining a low-level and asymptomatic parasitemia [Sergent 1935, Carter 2002, Doolan 2009].

People who have lived for years in areas with malaria may develop enough immunity to protect them from malarial illness, but not from malarial infection. Finding malaria parasites in the bloodstream does not necessarily mean that the person has active malaria. This partial immunity may wane if the person leaves the malarious area. In this case, the acquired immunity turns ineffective, and once again the individual becomes vulnerable to the full impact of a malarial infection.

Asymptomatic parasitemia is another aspect of the complex interactions between the parasite and the host's immunity. Asymptomatic carriers are people infected but not ill. This non sterilizing immunity (or premunition) is very frequently en-

2. A type of white blood involved in protecting against viral, fungal, and protozoal infections, and orchestrating the immune response

countered in endemic areas. Asymptomatic carriers do not show any clinical signs. However, they act as a parasite reservoir (as they continuously feed mosquitoes). Hence, patients with asymptomatic malaria contribute to the spread of the disease in the population. Taking into account both clinical and asymptomatic infections will help to establish effective tools for malaria elimination.

2.7 Diagnosis of malaria

Tools available for the diagnosis of malaria are used at both individual level to detect malaria in clinically suspected patients, and populational level to assessment the epidemiological characteristics of malaria.

Clinical malaria is defined as the presence of acute fever ($> 38°C$) (or other symptoms) and malaria parasites. Health authorities in malaria endemic countries have recommended for long decades to consider every fever case as malaria. Hence, a presumptive treatment of every febrile child or suspected malaria cases has to be directly initiated. Presumptive treatment can be explained by the fear of rapid exacerbation of malaria, especially in children. This instruction is until recently strictly applied in most African health settings, and the diagnosis of malaria in children is most often only based on the presence of fever. Moreover, even if a biological diagnosis is used and proved negative, in the great majority of cases a doctor confronted with a young febrile child do not hesitate to prescribe an anti-malarial drug.

However, effective treatment of clinical malaria requires prompt and precise laboratory diagnosis. Hence, when malaria is suspected, the clinical diagnosis (based on the patient history, symptoms and clinical findings) must always be confirmed by a laboratory diagnosis. Laboratory diagnosis of malaria involves a direct identification of malaria parasite or the presence of antigens in the blood of the patient.

At a population level, public health or research teams also need accurate diagnostic tools to assess the *Plasmodium* endemicity or to control the efficiency of a preventive action. Here, public health authorities or research programs can be interested in both clinical or asymptomatic infections.

2.7.1 Microscopic diagnosis: blood films

Microscopic examination of thick and thin blood smears is the "gold standard" for laboratory confirmation of malaria [Christophers 1951, Colbourne 1971, Draper 1971, Collier 1983, Trape 1985, Kilian 2000, Bejon 2006]. This method is used in both clinical and asymptomatic infections to assess the existence of an infection, to determine the parasite density and to identify the parasitic species causing the infection. It has the advantage of being reliable, technically simple and economic [3]. However, the efficiency of this method is closely tied to the quality of the smear, the parasite density, the reading technique and the microscopist skills.

3. In the endemic countries, one slide costs about 12 to 40 US cents.

A drop of the patient's blood is collected by fingerprick, or from a larger venous blood specimen. A small volume of blood is then spread on a glass slide (blood smear), dipped in *Giemsa* stain [4]. Blood smears are examined under a microscope at a 1000-fold magnification. Each of the four major parasite species can be distinguished by its physical features and by the appearance of the red blood cells that they have infected. The staining process slightly colourises the red blood cells (RBCs) but highlights Plasmodium species parasites, white blood cells (WBC), and platelets or artefacts. To prevent false diagnosis, the "stained" objects have to be analysed very carefully to determine if they are parasites or not. For this purpose, two sorts of blood film are examined: thick and thin blood smears.

Thick blood smears are used to detect infection, and to estimate parasite concentration. In thick blood smears, RBCs are hemolyzed [5]. As a consequence, only leukocytes and malaria parasites are detectable. However, due to the hemolysis and slow drying (30 minutes), the appearance of the parasite are distorted, which makes difficult the identification of species. A large volume of blood is examined (approximately 5-10 μl in thick smear against 1 μl in thin smear). An experienced microscopist can detect parasite levels as few as 5 parasites/μl of blood [Guerrant 2006]. Hence, picking up low levels of infection is easier on the thick film, which are about eleven times more sensitive than the thin film.

Thin films allow species identification, because the parasite's appearance is best preserved, which is important for giving the patient appropriate treatment. Here, the drop of blood is spread across a large area of the slide. It is dried for 10 minutes, and fixed in methanol after drying. Usually, both thin and thick films are made on the same slide. Hence, while fixing the thin smear, all care should be taken to avoid exposure of the thick smear to methanol. Both thick and thin smears shoud be used to make a definitive diagnosis [Warhurst 1996].

Microscopy can be time-consuming (requiring at least 60 minutes from time of sample collection to diagnosis). Another drawback of this method is that very low intensity infections can sometimes be missed, given a low number of parasites in the blood. Besides, diagnosis of species may be tedious as early trophozoites of all Plasmodium species look identical. Species identification is always based on several trophozoites. As mentioned in section 2.5, *P. malariae* and *P. knowlesi* look very similar under the microscope. *P. knowlesi* parasitemia increases rapidly, and causes more severe disease than *P. malariae*. Hence, infections must be recognized and treated promptly. Modern methods, such as PCR (see 2.7.4) or antigen tests (see 2.7.3), should be used to distinguish between the two species, especially in South-east Asia, where the two infections are prevalent [Murray 2008, McCutchan 2008].

An important point of controversy about microscopy concerns the reliability of the parasite density estimation using thick blood smears. Since there is no standardized way to count parasites and blood cells in thick blood films, the accuracy of the parasite density estimation will vary depending on the method used. This issue

4. *Giemsa* is a reagent that stains malaria parasites and allow for detection and recognition of Plasmodium species.

5. TBSs are not fixed with methanol.

is thoroughly discussed in Chapter 3.

2.7.2 Quantitative Buffy Coat (QBC)

The Quantitative Buffy Coat (QBC) is a laboratory test that detects infection with malaria or other blood parasites based on acridine orange staining of centrifuged blood samples. The Buffy Coat, also known as leukocyte concentrate, is the fraction of an anticoagulated blood sample containing the majority of the white blood cells and platelets after centrifugation of blood specimens.

Some studies indicate that infected red blood cells (RBCs) containing older trophozoites of *P. vivax*, *P. ovale*, and possibly *P. malariae* tend to concentrate above the RBC layer. The parasitized erythrocytes are concentrated in a small part of the RBCs column (immediately below the Buffy Coat). They are pressed against the wall of the tube, where they can be viewed by ultraviolet light microscopy. This helps to rapidly scan malaria parasites [Krishna 2003].

Concern over the ability of the QBC method in identification of species has been noted [Pinto 2001]. Hence, QBC cannot be considered as an acceptable alternative to microscopy for routine laboratory diagnosis. Additionally, special equipments are required, which make the QBC relatively expensive. One more disadvantage of QBC technique is that a permanent record of the test cannot be saved [Krishna 2003].

2.7.3 Antigen tests

In the presence of febrile episode, the Rapid Diagnostic Test (RDT) (also called Antigen-Capture Assay or Dipstick) allow to quickly establish the diagnosis of malaria infection by detecting specific malaria antigens in a patient's finger-stick or venous blood. RDT only requires a drop of blood. The test takes 15-20 minutes, and the results are read visually on the dipstick, as the presence or absence of colored stripes. These advantages make the RDT suitable for field experience. RDTs can be coupled with microscopic examination of thick blood smears.

Many healthcare settings, where an appropriate microscopy expertise is not available, either save blood samples for malaria microscopy until a qualified person is available to perform the test, or send the blood samples to commercial or reference laboratories, which result in long delays in diagnosis. In this case, the use of RDT is particularly appealing, since it allows a prompt diagnosis. TBSs are used later to quantify the parasite density. Yet not every fever is caused by malaria, RDTs were systematically performed for each case of fever for some years. However, none of the existing rapid tests are currently as sensitive[6] as thick blood films, nor as cheap [WHO 2011a]. In addition, RDT results are only qualitative. Hence, RDT can not replace on its own malaria microscopy.

6. The threshold of detection is about 100 parasites/μl by RDT against only 5 parasites/μl using TBS.

2.7.4 Molecular methods

Although thin blood smears are usually used to differentiate malaria parasite species, this task may become tedious when the morphologic characteristics of species are very similar (for P. vivax and P. ovale), or when the parasite morphology is distorted due to the bad quality of the smear or to drug treatment. Molecular diagnostic tests, such as polymerase chain reaction (PCR), can be used to bypass this difficulty. Moreover, PCR can detect very low levels of parasitemia, where thick blood smears may fail.

Developed in 1983 by Kary Mullis, Nobel Prize laureate, PCR is a technique in molecular genetics that allows the analysis of any short sequence of DNA (or RNA) even in samples containing very small quantities of DNA (or RNA). PCR is used to amplify (reproduce) selected sections of DNA (or RNA) for analysis. Thousands to millions of copies can be made of the DNA in a few hours [Bartlett 2003].

Although microscopy remains the gold standard diagnostic test for malaria in clinical settings, PCR-based assays can be 10 to 100-fold more sensitive than microscopy [Milne 1994, Hermsen 2001], especially in the setting of low parasitemia [Coleman 2006] or subclinical infections [Roshanravan 2003]. The PCR test has also been found useful in unraveling the diagnosis of malaria in cases of undiagnosed fever. Although this technique may be more sensitive than smear microscopy, it is of limited utility for malaria diagnosis standard healthcare settings. PCR is infinitely more expensive. Moreover, PCR results are usually not available rapidly enough to inform clinical decision-making in real-time. PCR is most of the time a efficient tool used by research teams.

2.8 Treatment

The lack of precise malaria diagnosis remains an important obstacle to the treatment, together with other relevant factors including drug resistance and availability and sustainability of drugs. These factors substantially increase the permanent challenge of malaria treatment in tropical areas.

In this section, we not give an account of the exhaustive list of drugs available and the different therapeutic schedules proposed. The aim of this section, however, is to present the recent evolution in malaria treatment that contributed to the improvement of the overall situation. This evolution has been required as drug resistance emerged. The widespread and indiscriminate use of antimalarials exerted a strong selective pressure on malaria parasites to develop high levels of resistance. This resistance is reflected by the ability of the parasite strains to survive and to multiply, despite the administration of proper doses of antimalarial drugs. Therefore, almost all available drugs have been compromised by the high adaptability of plasmodia. The limited number of effective drugs and the emergence of multi-resistant strains consolidate the need for new antimalarials.

The treatment of malaria depends on the assessment of the severity of the disease by the clinician. If uncomplicated malaria is diagnosed, the patient is given oral

drugs. In the case of severe malaria, the patient must be immediately admitted.

The treatment of uncomplicated malaria aims to block the aggravation of the disease and to completely eradicate the infection from the body. Patients must be followed for long enough for treatment outcomes to occur and until full recovery. Public health treatment strategies aim to reduce the transmission of infections, and to avert the emergence and spread of resistance to antimalarial drugs. Nowadays and since the beginning of the 21^{st} century, depending on the countries, the first line treatment used worldwide for *P. falciparum* infections is the combination of artemisinins with other antimalarials, which is known as *Artemisinin*-Combination Therapy (ACT) [Kokwaro 2009, WHO 2010a, WHO 2010b]. This is done to reduce the risk of resistance against *artemisinin*. Indeed, resistance can be prevented or slowed down, by combining antimalarials with different modes of action and, then, different resistance mechanisms. Probability of developing resistance to both drugs is the product of the two probabilities. Although the theory underlying combination treatment is well known in treating tuberculosis, leprosy and HIV [7] infections, it has recently been applied to malaria [8], and many malaria-endemic countries switched antimalarial drug policy to ACTs [Curtis 1986, White 1999a, White 1999b, Mutabingwa 2005, Garner 2005, Huho 2012]. Infection with *P. vivax*, *P. ovale* or *P. malariae* is usually treated on an outpatient basis. The treatment involves the treatment of blood stages using chloroquine or ACT, and the clearance of liver forms using *primaquine* [WHO 2010b, Waters 2012].

The top task of severe malaria treatment is to prevent death and neurological disabilities. Parenteral administration of antimalarial drugs is required for severe malaria treatment. *Quinine* remained the mainstay of malaria treatment until 1920, but *artesunate* has been shown to be superior to quinine [PrayGod 2008, Mathew 2010, Dondorp 2010, Achan 2011]. Severe malaria is most often caused by *P. falciparum*. In severe *P. falciparum* malaria, intravenous or intramuscular *artesunate* is recommended (for adults). *Quinine* is an acceptable alternative if parenteral *artesunate* is not available. Treatment of severe malaria requires additional measures, including management of high fevers and resulting subsequent seizures, and monitoring for respiratory depression, hypoglycemia, and hypokalemia [Sarkar 2010].

As previously mentioned, pregnant women are particularly vulnerable to malaria, since pregnancy reduces women's immunity. Since the beginning of the 21^{st} century, things have changed. For a very long time, the prevention of malaria during pregnancy was based on chloroquine together with the recommendation of using insecticide impregnated bed net. Currently, the protection of pregnant women in endemic areas rests upon an intermittent preventive treatment (IPTp), which consists in two curative doses of drug given after the first trimester. The drug usually used is a combination of *Sulfadoxine-Pyrimethamine*. However, due to high level of resistance, alternative drugs are needed. Pregnancy narrows the scope of alterna-

7. Combination therapy is made available to HIV/AIDS patients for the first time in 1996, leading to a dramatic decline in AIDS-related deaths.

8. In early 2004, the World Health Organization recommended that countries adopt ACTs.

tives among available drugs. However, many researches are currently undertaken to solve this question and mefloquine could be an alternative that is already available [WHO 2007b, Menéndez 2007, Briand 2009]. Here too, insecticide impregnated bed net use is strongly recommended.

Due to drug resistance, the prevention of malaria is argued to be more cost-effective than the treatment of the disease in the long run.

2.9 Prevention

Control strategy for malaria involves three living beings and their environment: human (the host), plasmodia (the agent), and anopheles mosquito (the vector), which implies a complex chain of measures that often complement one another.

Humans on the move not only transmit the disease, but spread antimalarials' resistance as well. Mosquitoes are also moving, they highly adaptable and have shown resistance to insecticides (see 2.9.1). The parasite, which hides in humans and mosquitoes, is also highly adaptable, and has also developed resistance to antimalarial drugs (see 2.8). Hence, an effective malaria control would target human first, control mosquitoes next, and keep trying to tackle the parasite with development of effective drugs and vaccines.

Sustainable malaria control requires a comprehensive set of solutions including the distribution and spraying of insecticide, mosquito nets, availability of antimalarial drugs and education in endemic countries. Research into vaccines is also crucial in the effort to eradicate the disease.

2.9.1 Vector control

Vector control remains the most generally effective measure to prevent malaria transmission. Its effect on morbidity and mortality is still under debate. Two commonly used and effective approaches in the vector control of mosquitoes are Insecticide Treated Nets (ITNs) and Indoor Residual Spraying (IRS). These methods vary considerably in their applicability, cost and the sustainability of their results.

IRS is the organized and timely spraying of insecticides on the inside walls of houses or dwellings in malaria-affected areas. The first pesticide used for IRS was DDT [CDC 2010]. Controversy concerns the health effects of IRS. Awareness of the negative environmental impactâ of the abusive use of DDT increased in the 1960s, when DDT became one of the largest used pesticides. Overspraying of DDT on crops contributed to the emergence and spread of DDT resistance in Anopheles mosquitoes and, ultimately, it was banned for agricultural use in many countries in the 1970s [van den Berg 2009]. The deterioration of IRS programs in some countries renewed interest in anti-larval and personal protection measures for reduction of malaria transmission.

ITNs, or bednets, are widely believed to be an effective way of controlling malaria. They are treated with insecticides that directly kill or inactivate mosquitoes and drive them away before they find the holes. Hence, even a treated net with holes

can provide a good protection. ITNs offer more than 70% protection compared with no net. Treated nets are twice as effective as untreated nets [Raghavendra 2011].

Global malaria control strategies has enabled endemic countries to greatly increase the access to prevention measures (ITN and IRS). ITN ownership in sub-Saharan Africa rose from 3% in 2000 to 50% in 2011. The majority of owners (96%) claimed that they make effective use of ITNs. The proportion of households protected by IRS in this area increased from less than 5% in 2005 to 11% in 2010 [WHO 2010a]. IRS requires the acceptance of the population of spraying insecticides once or twice a year and the preservation of sprayed surfaces. In contrast, ITNs requires the continuous use of the treated nets. Hence, IRS is more suitable for the rapid protection of a population, and ITNs are more suitable for progressive introduction and incorporation into sustainable population habits.

2.9.2 Vaccination

During its complex, multi-stage life cycle, malaria parasite not only expresses a great variety of proteins at different stages, but these proteins also keep changing often. This complexity makes the development of a malaria vaccine a very difficult task. Given this, there is currently no commercially available malaria vaccine, despite many decades of intense research and development effort [Geels 2011].

Malaria vaccines candidates traditionally target the different stages of the parasite's life cycle (pre-erythrocytic stage, asexual and sexual stages). Contrary to most malaria vaccines that target one of the three stages of the parasite's life cycle, SPf66 was based on both pre-erythrocytic and asexual blood stages. SPf66 was massively tested in human field trials, in both low and high disease transmission areas, in the 1990s, but evidence of efficacy was not enough to develop the vaccine [Graves 2006c]. Some vaccine candidates, who target the blood-stage, has been proven to be insufficient to meet the desired efficacy [Graves 2006a]. One of the most promising vaccines, who target the pre-erythrocytic stage, is undoubtedly the RTS,S vaccine [Graves 2006b].

RTS,S is the most advanced vaccine candidate against *P. falciparum*. During a phase II trial in Kenya, this vaccine has shown 53% efficacy in reducing all episodes of clinical malaria in infants aged 5-17 months, the duration of follow-up varied according to the time of recruitment, between 4.5 and 10.5 months (mean, 7.9) [Bejon 2008]. While 53% percent protection is not very effective[9] even that much protection potentially can be translated into tens of millions of cases of malaria in children averted annually which would save millions of lives over a decade.

Initial results from larger ongoing phase III revealed that RTS,S decreased malaria by a half in young children, and by one-third in infants over 12 months [The RTS S Clinical Trials Partnership 2011]. However, new findings on long-term follow-up of earlier phase II study showed that the efficacy of RTS,S vaccine over the 4-year period was 16.8%. Efficacy declined over time and with increasing malaria

9. most vaccines are not released until they do better than 90 percent.

exposure [10] [Olotu 2013].

2.9.3 Chemoprophylaxis

Malaria can be severe in a non-immune individual. Travelers from non-malarious area to a malarious area should be protected. As there is no vaccine available for protection against malaria, despite intensive research for decades, an alternative method that offers a fairly reliable protection against malaria is needed. Use of antimalarial drugs to prevent the development of malaria is known as *chemoprophylaxis*.

Chloroquine has been used extensively where the parasite was sensitive [Jacquerioz 2009]. This molecule is no more indicated [Sidhu 2002]. Alternative strategies include (1) mefloquine (*Lariam*), (2) doxycycline (available generically), and (3) the combination of atovaquone and proguanil hydrochloride (*Malarone*) [Jacquerioz 2009]. Since (1) is classically associated with higher rates of neuropsychiatric adverse effects [Jacquerioz 2009], (2) and (3) are the best tolerated.

Any malaria prophylaxis must be taken before, during, and especially after traveling to a malarious area. Proguanil, mefloquine, and doxycycline are only effective once the parasite has entered the erythrocytic stage (see Figure 2.2), and therefore have no effect until the liver stage is complete. Hence, these prophylactics must continue to be taken for four weeks after leaving the endemic area. This is in contrast with *Malarone* and primaquine prophylactics that target the blood stage as well as the initial liver stage. Thus, the user only have to take the medicine for 7 days after traveling rather than 4 weeks.

Medications for malaria prophylaxis do not provide a complete protection. Hence, travellers to malaria endemic areas should take measures to prevent mosquito bites in addition to good compliance with chemoprophylaxis. Malaria prophylaxis is not recommended for people living in malarious area, and it is not prescribed as a remedy to prevent re-infections as well.

Malaria treatment, control and prevention measures dramatically impute on the budget allocated to health care in malaria endemic countries that remain trapped in a downward spiral of poverty.

2.10 The socio-economic burden of malaria

Malaria costs lives, and money. The socio-economic impact of malaria is tied to a vicious spiral of poverty-malaria-poverty. Any discussion about the two causalities would be incomplete without defining poverty. A simplistic and nuanced understanding of poverty draws poverty as insufficient income or consumption. However, poverty has also non-monetary dimensions, as insufficient outcomes in education, healthcare infrastructure, environment quality, research and control, political and civil rights and many others. A click at the poverty site shows malaria as one of

10. Vaccine efficacy was 43.6% in the first year but was -0.4% in the fourth year.

the top seven problems of poverty. Malaria death counter (per hour), posted in this website, will undoubtedly leave you puzzled.

Poverty causes malaria. This paradigm leaps to the conclusion that malaria is a disease of the poor. The anophele mosquito, of course, does not distinguish between the rich and the poor. However, the geography and the environment, in which poor communities are living, are particularly appealing for mosquitoes. 58% of malaria deaths occurs in the poorest 20% of the world's population [Ricci 2012]. Risk for malaria increases in children from poor householders [Krefis 2010]. Correlations between febrile episodes in children and low householders income have been observed in Sub-Saharan Africa [Filmer 2005]. Poor householders cannot afford malaria prevention tools such as bednets, good-quality drugs, doctors' fees and health facilities access. Low education and lack of awareness about the scourge of malaria disarm people in their struggle against this disease of poverty. Wars, political turmoil and upheavals have also been marred by malaria [Melville 1910].

Malaria causes poverty. From 1965 to 1999, growth rate in endemic countries was 1.3% lower than in non-endemic countries. Direct costs include prevention and treatment expenditures by personal protection measures and government health authorities. In Africa, these direct costs have been estimated to be 12 billions dollars per year, which maintains and furthers poverty. Indirect cost involves labor time loss due to illness or death, which accounts for 75% of the total malaria cost per household. Absenteeism is a thorn in the side of growth. Intangible costs inevitably exacerbate the detrimental economic impact of malaria. They includes low birth weights, suffering, physical disabilities (anemia, neuro-disability, cognitive deficits), deschooling and social exclusion. The combination of the factors cited below results in deep-seated inequalities between endemic countries and non-endemic countries.

On a more positive note, governments, today, are more and more investing in malaria R&D, which is highly cost-effective. Investing in malaria R&D reduces malaria incidence and yields positive economic benefits. Saved costs may be reinvested to fund other health initiatives and to reduce poverty.

2.11 Discussion

Malaria as a disease was unveiled by several remarkable Nobel Prize Laureates, who discovered separate parts of its pathology at different points in time. Laveran had made discerning observations on the parasitic cause of malaria. Manson toyed with the idea that the disease might be carried by mosquitoes. However, it was Ronald Ross who attempted to gather together these fragments of innovative thoughts into one final concept. Since Ross's discovery, the situation of people suffering from malaria has significantly changed, but still not enough.

After the disappointing results from RTS,S malaria vaccine trials, irrepressible optimism have again given way to a grimmer reality: the struggle against malaria persists. Since immunity given by a natural malaria infection wanes in the absence of continued exposure, chances that a vaccine confers lasting immunity are slim

to none. However, there remains a glimmer of hope: country-level elimination of malaria may be completely achieved with sustainable malaria control programs.

Important measures for reducing the burden of malaria morbidity and mortality include more sensitive diagnostic tools, effective use of antimalarial drugs, improved personal protection and mosquito control. Moving from malaria control towards permanent reduction to zero of all malaria cases, is a complex and difficult process that will only occur if there is sustained leadership, innovation, financial commitment, political will and concerted community efforts.

The role and importance of community involvement is often over-looked in control programs. Many malaria programs focus on providing the communities with vector control interventions (indoor spraying and bednets). ITNs and IRSs are offered to the community for free, and without expecting any action on the part of the community. Awareness campaigns often consists in providing information at healthcare utilities and through the media. However, rural communities' accessibility to pertinent information still remain largely unmet because of lack of facilities and/or low levels of education. These campaigns usually emphasis on preventive measures at individual level by motivating inhabitants of rural areas to use bednets and to protect themselves from mosquito bites, but, they omit to focus on the direct impact of individual behavior on the overall effectiveness of the local malaria control interventions. Due to the lack of rural community members' involvement, these practices often results in a culture of dependency, in which the affected individuals in rural areas completely rely on formal control policies to protect them from mosquitoes. Although rural community members' usually do not have enough resources or knowledge, they can actively, and efficiently, participate in the control of malaria in different ways.

During a field study in Tori Bossito (Benin), I was able to notice that, in some villages, a variety of risks could be easily avoided by local inhabitants with little or no resources. This made me remember when I first announced to my father, who is engineer at the Ministry of Environment and Sustainable Development of Tunisia, and who participated to the campaigns for the eradication of malaria in Tunisia in 1979 [Chadli 1986], that I decided to work on malaria for my phd, with field applications in Benin, he just said: "*... People do not need mathematical models to solve the problem of malaria. They first need environmental sanitation*". For example, small puddles of standing water, often gathered in and around rural villages in the rainy season, can easily be filled in by nearby householders. This intervention do not require any material or technical resources. Although these puddles are not perceived as a thread by local inhabitants, they do provide breeding sites to thousands of vector mosquitoes, which not only increase the risk of malaria, but also participate to the nuisance factor. Another example is the inappropriate use of bednets. Most of bednets in rural areas are in poor conditions, or used inappropriately. Deteriorated bednets are a poor physical barrier against blood-seeking mosquitoes. Without such local commitment to strengthening malaria control interventions efficacy and sustainability, malaria may never be eliminated, and the picture would be far more bleak, with millions of deaths over years.

While these weapons have proven to be potent in the fight against malaria, one of the major technical problems facing malaria eradication in certain areas is the development of drug resistance. Drug resistant cases of *P. falciparum* are the most alarming, since this species causes the most fatal and medically severe form of the disease. At this point it has no known cure. In the fact of such a tragic dilemma: what hope can there be? Vaccine would be the backbone of public health interventions against malaria, especially in poor countries, where it would help contain or even eradicate this global killer. Despite global efforts to wipe out this disease, no effective vaccine currently exists. The need for a vaccine is great, but biology is complex and economics are disadvantageous.

Although *P. falciparum* is the most dangerous type of malaria, all *Plasmodium* species are potentially life threatening, especially when managed inappropriately. In the case of missed or delayed diagnosis, the harm could be devastating. A patient may go from mild through complicated to severe disease. Prompt, reliable and accurate diagnosis of malaria is one of the cornerstones of effective disease management. Studies on the field show the rates of agreement between microscopical and serological diagnosis of malaria are surprisingly low [Mitiku 2003]. Estimation of the parasite density on thick blood smears is a task of major importance. The accuracy of parasite density estimation depends on the reading method, the time spent in reading and the microscopist skills. A standardized tool, which is able to perform diagnosis with the same ground criteria uniformly everywhere, is therefore needed.

Chapter 3

Parasite Density Estimation

"You can't control what you can't measure."

Tom DeMarco, 1986

Contents

In this chapter, we review the commonly used parasite density estimation methods. Emphasis is laid on the importance of the parasite density estimation according to clinical malaria diagnosis and epidemiological studies.

3.1 Introduction

In Chapter 2, we pointed out the enormous burden of malaria. It is therefore very essential that every case of malaria be assessed thoroughly. Positive and negative diagnosis is usually considered sufficient for the assessment of therapeutic outcome. However, assessing the burden of malaria in terms of morbidity and level of infection require measuring the parasite density (PD) of malaria infection as a primary endpoint. This is an important factor in assessing the efficacy of a vaccine candidate or a drug in clinical trials. Despite some inherent limitations [Bejon 2006], thick blood smears (TBSs) are the most established, and widely-used technique for PD quantification in the blood [WHO 1999].

3.2 Field experience

At the first beginning of my PhD, the malaria issue was completely foreign to me because, first of all, I belong to a country in which malaria has been completely eradicated since 1979, secondly, at that moment, the only African country I had ever been in was *Tunisia*. For these reasons, I could not have expected a better initiation than a field experience in *Benin* from the word 'go'.

The main objectif of this field experience was to establish a first contact with the teams that collect data in the field. I started out by mostly observing to better understand their techniques and to identify their potential difficulties, constraints and needs. Through this field study, I took a closer look at the TBSs reading methods, and I proudly read ones! But, as time went on, different aspects of the problem, that we wouldn't expect at the beginning, came into light.

A TBS is considered acceptable if it is well made and of even thickness. Errors in TBS preparation may affect the examination, and then the outcome for the patient. Routine examination of a thick film is based on the examination of a predetermined number of "good" high-power fields (HPFs) (see Figure 3.1), which varies by program. A slide is declared negative if no parasites have been seen in these fields. Microscopists are asked to select a readable and a thick area of the film, away from feather and lateral edges, that is well stained, free of staining precipitate and well populated with white blood cells (WBCs).

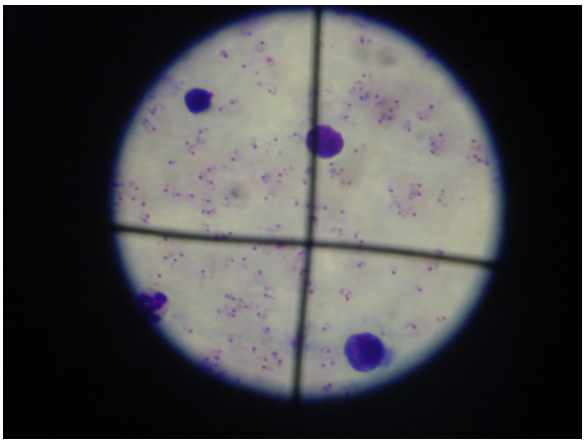

Figure 3.1: ***High power microscopic field.***
The circular area corresponds to the part of the TBS viewed through the lens of the microscope, namely the high-power field (HPF). The picture depicts two types of stained objects: numerous small specks which represent ring-form *P. falciparum* trophozoites lysed from the RBC's, as well as a number of huge purple dots representing the white blood cells.

In addition, TBSs must be examined in a specific way for consistency. They are examined following the pattern of movement shown in Figure 3.2; that is, starting at the "x" mark, the film should be carefully examined, field by field, by moving to each contiguous field along the edge of the thick film, then moving the slide inwards by one field (without overlapping with the previous field), returning in a vertical movement and so on. For efficient examination, the microscopist continuously focus and refocus using the fine adjustment throughout examination of each field.

However, in practice, these instructions are not being fully complied with. In addition, even under optimal conditions, the distribution of blood thickness within the TBS will never be completely homogeneous [Dowling 1966, Kilian 2000], and then the choice of the "good" readable area for examination could influence the

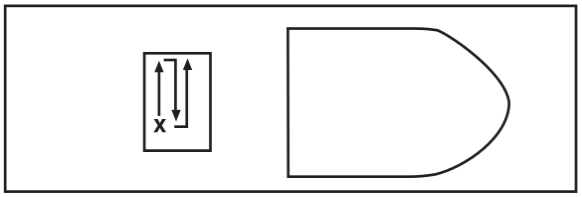

Figure 3.2: ***Examination of a thick blood film.***
Thick (left) and thin (right) films are normally made on the same slide. For consistency, TBS must be examined following the pattern of movement shown in this reading diagram.
source: [WHO 2010a]

PD estimation. Additionally, the microscopist may make his own subjective and possibly arbitrary understanding of the HPF limits to avoid overlapping. Moreover, examining 200 HPFs takes approximately 15 to 20 minutes for an experienced microscopist. The choice of the predetermined number of HPFs (or WBCs) that have to be read (or seen) is fixed in a biased and discriminatory way, most of the time regardless of parasitemia levels or acceptable variability. The choice of this number, however, may have a critical impact on the overall cost of the epidemiological survey. Finally, malaria microscopy is a skilled exercise requiring great care at each step of the procedure (from the sample collection to the examination) and precise visual and differential skills. The microscopy results are then as reliable as the microscopist performing the examination. Hence, allowing operator decision making adds further bias to the microscopy outcomes and increases discrepancies between malaria slide readings.

Despite my unfortunate and accidental stay with a "giant" lizard (three years later and I am still not over it), my field experience in *Benin* influenced my work in a very positive, scientifically sound way. It allowed me to approach in practical terms the PD estimation constraints faced by microscopists during the examination of TBSs. It enlightened my way of understanding the importance of PD estimation. We shall dwell on this last point in the following section. Furthermore, this field experience enabled me to notice the weakness of existing methods in the matter of left-censoring and cost-effectiveness, points that we shall take up in Chapter 4. This elements provided additional food for thought for this work and suggested possible improvements by designing more efficient and cost-effective alternative protocols. Finally, I was able to gain a broad perspective of the extent of malaria burden in rural areas of the *Benin* and the challenges facing elimination.

3.3 The importance of PD estimation

Accurate estimation of the PD is an important endpoint in epidemiological studies and clinical trials, both as a direct measure of the level of infection in a population and when defining parasitemia thresholds to diagnose malaria in case of fever episodes. Malaria PD estimates are also used to assess the development of naturally

acquired immunity [Rogier 1993] and in malaria vaccine investigations [Alonso 1994, Petersen 1996a, Small 2010]. Therefore, inaccurate estimation of PD can lead to patient mismanagement, failure of clinical trials of drugs and vaccine candidates and public health misinformation [Dini 2003, Bates 2004, Prudhomme O'Meara 2006a].

However, two approaches must be distinguished. The first approach concerns clinical malaria diagnosis and two problems can be pointed out depending on whether the question concerns an individual or a population. At the individual level, recent studies have highlighted the massive problem of misdiagnosis in malaria endemic countries [Amexo 2004, Reyburn 2004, Zurovac 2006]. From a clinical point of view the question is to determine whether a person presenting fever suffers from malaria or not and, in that sense, the main problem is a false negative result. Here, if a measure is falsely negative, the patient will be miscategorized and incorrectly treated, and measurement errors can lead to poor patient outcome. Nevertheless, although the level of infection is considered as a controversial sign of potential severity [WHO 2000], treatment and medical supervision must be immediately started even if the PD is not accurately determined.

From an epidemiological point of view (e.g. to determine the incidence or prevalence of clinical malaria in an area or in a population under close medical surveillance) clinical malaria is often considered as any case of fever or fever-related symptoms (headache, vomiting, subjective sensation of fever) associated with a *P. falciparum* parasite/leukocyte ratio higher than an age-dependent pyrogenic threshold of PD previously identified in the patient [Rogier 1996, Milet 2010]. In this case, a feverish individual harboring a PD under his age-specific threshold is not considered as a malaria case and will be monitored by the medical team involved in the study. In such a situation the accuracy of PD determination is obviously of great importance, not only for the patient but also for the outcomes and the conclusions of the study.

A second approach concerns the assessment of PD in epidemiological studies, when PD is used as the variable of interest or as a covariate, independently of clinical disease. For example, genetic epidemiology studies often focused on a mean level of *P. falciparum* infection during a follow-up period [Garcia 1998, Timmann 2007, Sakuntabhai 2008, Milet 2010]. Great care must also be taken in the analysis of PD estimates when PD is related to other explanatory variables, malariometric (e.g. parasite ratio, gametocyte ratio, mixed infection) or not (age, environmental or behavioral factors, medicine intake in clinical trials), when using statistical models as logistic regression and linear mixed effect models. In these cases as well as in population studies using a pyrogenic threshold to define clinical malaria, inaccurate estimates of PD might influence the parameters of associations between drug efficacy and the incidence of clinical malaria episodes in field trials [Kilian 2000], or between risk factors in epidemiological studies.

Several methods for PD estimation by microscopic examination of TBSs are used [Dowling 1966, Trape 1985, Kremsner 1988, Planche 2001, Garcia 2004, WHO 2010a, Le Port 2011b]. In most of these methods, the number of asexual parasites seen is usually counted either relative to a given number of HPFs, or according

to a fixed number of WBCs. To express PD as the number of parasite per microliter of blood, parasite counts are converted to parasites per microliter using appropriate multiplicative factors depending on whether parasites are counted relative to WBCs or HPFs. In the following, the "given" (of predetermined) number of WBCs or HPFs will be referred to as the threshold value. PD estimation methods involving threshold values are called threshold-based counting techniques.

3.4 Threshold-based counting techniques

Threshold type and value may vary from one method to another. In the first case, parasites are counted relative to a given number of HPFs. The conversion to counts per microliter relies on the assumption that the volume of blood per HPF is approximately 0.002 μl [Dowling 1966, Bruce-Chwatt 1985, Warrell 2002]. Parasites are usually counted in 100 [Earle 1932, WHO 1961, Dowling 1966] or 200 [Molineaux 1980, Trape 1985] consecutive HPFs. Since the volume of blood corresponding to 200 HPF of a thick film is 0.4 μl, examining 200 HPFs was considered as the best compromise between the need to reduce the risk of missing parasites and the need to minimize the reading time [Trape 1985]. A variant of this method adapts the reading effort to how numerous the parasites are. If less than p parasites are counted in the n first HPFs, then an additional number of fields m has to be read. Generally, a predetermined number of HPFs has to be examined before deeming a slide to be negative.

The relative quantity of blood examined in a predetermined number of HPFs is unknown. In addition, usually the thickest part of the slide is examined in the first type of threshold-based counting methods. However, the volume of blood examined and the thickness of the smear depend on the experience of the microscopist. The exact volume of blood examined is then required to give a precise estimation of the PD, since it allows a direct conversion of counts per a fixed number of HPFs to counts per μl. To bypass this difficulty, some methods suggest to count parasites in a fixed volume of blood.

The *Lambaréné* method [Kremsner 1988] of counting is applied by dropping and evenly distributing $10\mu l$ of blood on a $10 \times 18 - mm$ area of a microscope slide. The number of parasites is counted per 3, 5, 10, 50, or 100 HPFs. Then, the count is multiplied by an appropriate multiplication factor to yield to the total number of parasites per microliter of blood. The value of this multiplication factor depends on the microscope magnification ($\times 1,000$ usually) and the area of the microscope slide, and it is most of the time $400 - 800$ for counts per HPFs. A variant of this method was briefly mentioned in [Planche 2001]. The *Lambaréné* method has the advantage of direct quantification of the PD, being rapid, and sensitive. This method showed similar results to those of quantifying the parasitaemia according to the standard number of leukocytes and may also be preferable to the first type of methods in terms of accuracy and reliability [Planche 2001].

In the second case, parasites are counted against a fixed number of WBCs.

The number of parasites is counted on one tally counter and the number of WBCs is counted on a second one. The conversion to counts per microliter depends on an assumed mean of 8,000 WBCs per microliter of blood [Bruce-Chwatt 1958, Greenwood 1987]. The average value of 8,000 WBCs per μl is accepted as reasonably accurate by *The World Health Organization* (WHO) [WHO 2010a]. The number of parasite is usually counted against 200 WBCs and multiplied by 40 to give the number of parasites per μl [Chippaux 1991, Greenwood 1991, Warhurst 1996, Prudhomme O'Meara 2006a]. In more adaptative methods, the WBC threshold depends on how numerous the parasites are. In these methods, if parasites are not numerous in the first readings, then an additional number of WBCs should be counted. Hence, the lower the number of parasites counted, the higher the number of WBCs that should be counted. According to the WHO recommendations [WHO 2010a], parasites are counted until 200 WBCs are seen. If less than 100 are found, then counting should be continued up to 500 leukocytes. During a research program conducted in the *Tori Bossito* area in *Southern Benin* [Le Port 2011b], the PD was determined by simultaneously counting parasites and leukocytes. The counting stops when either 500 WBCs or 500 parasites are seen whichever comes first.

The accuracy of the mean number of 8,000 WBCs per μl has been investigated in many studies. Leukopenia (decreased WBC count) and leukocytosis (increased WBC count) may confound population studies that estimate parasite densities on the basis of the assumed WBC count. Using of assumed WBC count rather the absolute WBC count may lead to over-estimation, or under-estimation of the PD in malaria infections [Jeremiah 2007]. This number is shown to be lower in adults, e.g. 7,000 μl in [Wintrobe 1967], and between 5000 and 6500 per μl in [Blistein 1950, Acker 1967, Rougemont 1991]. This number is higher in children, e.g. 10,000 per μl in children between two and four years in [Cartwright 1968], the same average is recorded in children less than five years in [Adu-Gyasi 2012]. Hence, there is a significant correlation between the WBC count and the age. WHO recommended an age-group system for malaria studies [WHO 1963]. Regional-based WBC counts are suggested to improve the accuracy of PD estimation in epidemiological surveys [Trape 1985, Adu-Gyasi 2012]. However, most malaria-endemic countries in sub-Saharan Africa may not be able to express the PD on the basis of the "real" WBC count [Olliaro 2011]. The use of the assumed WBC count is considered as reasonable for use in PD estimation [WHO 2010a]. It has been demonstrated to be fairly accurate as it counterbalances the loss of parasites after the dehaemoglobinization and staining of thick films in [Bruce-Chwatt 1958, Dowling 1966].

Some counting methods based on the "real" WBC count have been proposed by research teams in order to optimize the assessment of PD. Among them, *Garcia et al.* (2004) proposed the following methods, which proceeded in two independent steps. Firstly, only WBCs were counted in 30 HPFs and expressed as the number of WBCs by HPFs. Secondly, parasites were counted in 10, 50 or 200 HPFs, and expressed as the number of parasites by HPFs. The number of HPFs read depended on how numerous the parasites were and the result was computed by dividing the

mean number of parasites by the mean number of leucocytes and expressed as a number of parasites per 100 leucocytes. A TBS has been declared negative when no parasite was detected in 200 fields. This method has been used for a genome wide association study using the data of a cohort of children followed-up in *Niakhar* area in *Senegal* [Milet 2010].

3.5 Discussion

The accuracy and efficiency of conventional malaria microscopy of TBSs have been investigated in the scientific litterature [Dowling 1966, Trape 1985, Bland 1986, Payne 1988, Greenwood 1991, Clendennen 1995, Mulder 1998, Dubey 1999, Dini 2003, Prudhomme O'Meara 2005, Bejon 2006, Prudhomme O'Meara 2006a, WHO 2007a, Alexander 2010]. These studies have shown that many factors may influence the reliability of estimation methods, including the microscopist skills, the sample preparation, the method features (number and type of thresholds), the loss of parasites during the staining and the dehaemoglobinization of TBSs, assumptions made on the volume of blood per HPF and the assumed number of WBCs per μl. Errors in PD estimation have major consequences for the patient management, the results of epidemiological surveys, the public health initiatives, and the effectiveness of clinical trials of drugs and vaccine candidates [Dini 2003, Bates 2004, Prudhomme O'Meara 2006a].

In an attempt to understand how the thresholds involved in parasite enumeration methods contribute to the magnitude of discrepancies in PD determination, their impact in variability measures generated by commonly used threshold-based counting techniques are studied in Chapter 4. Furthermore, the accuracy and efficiency of PD estimation methods in malaria need to be improved through the enhancement of operational training and the standardization of laboratory diagnosis. The latter issue is discussed in Chapter 6.

Chapter 4

Statistical Properties of Parasite Density Estimators

"It is really just as bad technique to make a measurement more accurately than is necessary as it is to make it not accurately enough."

Arthur David Ritchie, 1923

Contents

4.1 Introduction

Epidemiological interpretations must rely on solid evidence. The reproducibility for parasite density (PD) data is of major interest. However, all the methods used to determine the PD (see section 3.4), potentially induce variability. To deal with this potential inaccurate estimation of PD, research teams tend to analyze more slides and subjects. By taking duplicate readings or larger sample sizes, we can statistically improve our knowledge of the PD being measured. Then, we can decrease the variability in microscope slide readings and improve the accuracy and reproducibility of the measurements. However, one of the problems the research teams have to deal with is that during large scale studies the number of thick blood smears performed can be greater than $10,000$. Then, the repetition of the microscope slide examination

leads to an important cost overrun in terms of both money and time. One may wonder whether such practices have a significant interest for the final results. With low parasitemias, it is probably worth the effort of reading more slides. But in some situations it is not needed, for example, with large parasitemias levels.

To our knowledge, none of the studies of variability have dealt with the sampling error generated by the threshold-based counting techniques or evaluated the impact of the existing threshold values in endpoint measurements. In addition, the accuracy and consistency of these methods have largely been overlooked. Furthermore, there is no general agreement on the optimal method for estimating the PD according to threshold values. Further experimental evidence is needed to determine which parasite counting technique is most accurate, reproducible, and efficient. The aim of this chapter is to explore the variability of four frequently used threshold-based counting methods of determination of PD. For each of these methods, we assessed the consequences that a modification of the threshold can have on variability.

4.2 Materials and Methods

4.2.1 Threshold-based counting techniques

PD estimates accuracy vary significantly depending on the methodology from which they are derived. The estimation method differs from one healthcare organisation to another. Here, we are interested in four basic types of threshold-based counting techniques commonly used in epidemiological surveys. In these methods, parasites are usually counted either relative to a given number of high-power fields (HPFs), or according to a fixed number of white blood cells (WBCs). References to these (and other) methods can be found in Chapter 3.

The first method consists in counting parasites in 200 consecutive HPFs [Molineaux 1980, Trape 1985]. This method will be referred to as Method A. Here, the number of HPFs read is the threshold value. We investigate the influence of this number on the reliability of the PD estimation.

The second method consists in counting parasites against 200 WBCs [Chippaux 1991, Greenwood 1991, Warhurst 1996, Prudhomme O'Meara 2006a]. This method will be referred to as Method B. In this method, only one threshold value is specified, which is the number of leukocytes seen ℓ. We are interested in how the value of ℓ affects measures of variability.

The third method considered is the one recommended by *The World Health Organization* [WHO 2010a]. In this method, parasites are counted until 200 WBCs have been seen. If 100 parasites or more are found, the number of parasites per 200 WBCs is then recorded. Else, counting should be continued up to 500 WBCs. This method will be referred to as Method C. Three parameters are specified : the required number of parasites p, the required number of leukocytes in the first step ℓ_1, and the required number of leukocytes in the second step ℓ_2. Modeling, estimating and validating multidimensional distribution functions cast difficult problems, both conceptual and technical. For that reason, it is more convenient to fix $\ell_2 = 500$ and

to study the method's performance by varying the two parameters p and ℓ_1. Hence, we obtain the influence of adding an extra threshold value on the final estimation.

The last method considered was used during a research program conducted in the *Tori Bossito* area in *Southern Benin*. In this program, the PD is determined by simultaneously counting parasites and leukocytes. The counting stops when either 500 WBCs or 500 parasites are seen whichever comes first [Le Port 2011b]. This method will be referred to as method D. Two parameters are specified : the required number of parasites p and the required number of leukocytes ℓ. We analyze the performance of the method with respect to effectiveness and efficiency for different values of parameters p and ℓ.

Unlike methods A and B, methods C and D are adaptative methods. In these methods, counting stops when parasites are found in sufficient number. Hence, their cost is reduced for high parasitemias.

4.2.2 Measures of variability

The source and scale of measurement error depends on several parameters, such as sample preparation, staining process, counting technique, microscopist performance, etc. However, variation of the PD within a slide is expected even when prepared from a homogeneous sample [Alexander 2010]. The sampling variability is a source of interest when studying the efficiency of estimation methods. It refers to the different values which a given function of the data takes when it is computed for two or more samples drawn from the same population. In this chapter, we are interested in the sampling errors and biases induced by threshold-based counting techniques and more particularly in the impact of threshold values in endpoint measurements.

Let θ be the parameter that denotes the real value of the PD per microliter of blood and let $\hat{\theta}$ be its estimate. Since $\hat{\theta}$ is a random variable, it can never been said with certainty that this estimate is close to the true value of θ. For that reason, we consider its statistical properties, that is, its probability distribution $P(\hat{\theta})$, or some restricted aspects thereof. Here, we focuss on variability measures : mean error (ME), coefficient of variation (CV) and false negative rate (FNR).

4.2.2.1 Mean Error

In order to define the variability measures, we need to introduce the concept of mathematical expectation. The expected value of the estimator $\hat{\theta}$ denoted as $\mathbb{E}(\hat{\theta})$ is an average taken over all possible values of $\hat{\theta}$. Suppose $\hat{\theta}$ takes value s_1 with probability $p_1 = P(\hat{\theta} = s_1)$, value s_2 with probability $p_2 = P(\hat{\theta} = s_2)$, and so on, up to value s_n with probability $p_n = P(\hat{\theta} = s_n)$. Then the expectation of $\hat{\theta}$ is defined as

$$\mathbb{E}(\hat{\theta}) = \sum_k s_k p_k.$$

The sampling bias occurs when the true value (in the population) differs from the observed value (in the study) due to a flaw in the sample selection process. An estimator *bias* is the difference between the estimator's expected value $\mathbb{E}(\hat{\theta})$ and the true value of the estimated parameter θ. Hence, in computing the bias induced by different counting techniques, we used bias $= \sum_k s_k p_k - \theta$. An estimator with zero bias is called *unbiased*. Mean error is the bias expressed as a percentage of θ, i.e. ME $= \frac{\text{bias}(\widehat{\theta})}{\theta}$. It provides a measure of the magnitude of the bias and allows comparing different methods.

4.2.2.2 Coefficient of variation

A measure of the sampling error is the standard deviation which is the square root of its variance. Standard deviation is a measure of dispersion from the mean, or the expected value and it is commonly used to compute confidence intervals in statistical inferences. The reported margin of error is typically about twice the standard deviation (1.96), the radius of a 95 percent confidence interval. Sampling variability can also be expressed relative to the estimate itself through the coefficient of variation (CV), which is defined as the ratio of the standard deviation σ to the true value θ. In computing the CV induced by different estimation methods, we used

$$\text{CV} = \frac{\sqrt{\sum_k (s_k - \mathbb{E}(\hat{\theta}))^2 p_k}}{\theta}.$$

Then, CV is expressed as a percentage of θ.

4.2.2.3 False negative rate

The last measure of variability we study is the false negative rate (FNR), also known as Type II error or β error, which is the error of failing to reject a false null hypothesis. The false negative rate indicates the probability of a counting method to estimate PD as null when it is not, i.e. $P(\hat{\theta} = 0 \mid \theta > 0)$.

4.2.2.4 Cost

In addition to the variability, we are also interested in the cost-effectiveness of each method. We define the method's cost as the number of HPFs that has to be read to reach the threshold value. Once it is reached, we stop the examination of the smear. Depending on the method being used, the cost is based on the number of parasites (or WBCs) required to stop the reading of the thick blood smear. Let T denote the required number of HPFs to stop the counting. We first compute $P(T = t)$. Then, we express the Method cost as $\mathbb{E}(T)$.

4.2.3 Methodology

The following methodology is used to compute the three measures of variability (ME, CV and FNR) and to assess cost-effectiveness of methods. Firstly, the ex-

act distribution of $\hat{\theta}$ is computed through recursive formulas. Secondly, based on this probability density function, measures of variability are derived. Finally, cost-effectiveness is defined for each method as the required number of HPFs that has to be read until the threshold is reached. A `C++` program is used to implement these recursive formulas. The calculations are performed under two assumptions :

A1. The distribution of the thickness of the smear, and hence of the parasites within the smear, is homogeneous.

A2. The distribution of the parasites in the HPFs is uniform, and thus can be modeled through a Poisson distribution [Petersen 1996a, Kirkwood 2001, Alexander 2010].

Notations

Let X_i be a random variable that represents the number of parasites in the *i-th* HPF. Suppose that X_i are independent and identically distributed (*i.i.d.*).
Under an assumption of uniformity, the number of parasites per field can be modeled using Poisson distribution (assumption A2). If the expected number of parasites per HPF is λ_p, then $X_i \sim \mathcal{P}(\lambda_p)$. Thus $\mathbb{E}(X_i) = V(X_i) = \lambda_p$.

Let Y_i be a random variable that represents the number of leukocytes in the *i-th* HPF. Suppose that Y_i are independent and identically distributed (*i.i.d.*). Leukocytes are supposed evenly distributed over the thick smear. Therefore, the number of leukocytes per field can be modeled using the Poisson distribution. If the expected number of parasites per HPF is λ_ℓ, then $Y_i \sim \mathcal{P}(\lambda_\ell)$. Thus, $\mathbb{E}(Y_i) = V(Y_i) = \lambda_\ell$.

- Let ϕ denote the parasite density per WBC.
- Let S_t be the sum of parasites in t consecutive HPFs.
  Then, $S_t = \sum_{i=1}^{t} X_i \sim \mathcal{P}(t\lambda_p)$.
- Let R_t be the sum of leukocytes in t consecutive HPFs.
  Then, $R_t = \sum_{i=1}^{t} Y_i \sim \mathcal{P}(t\lambda_\ell)$.
- Let U_p be the minimum number of HPFs required to obtain p parasites.
  U_p can be expressed in terms of X_i as follows

$$U_p = \arg\min_t \{\sum_{i=1}^{t} X_i \geq p\}$$

  Probability of U_p is given by

$$\begin{aligned} P(U_p = t) &= P(S_t \geq p,\ S_{t-1} < p) \\ &= \sum_{s=0}^{p-1} P(X_t \geq p - s) \cdot P(S_{t-1} = s) \end{aligned} \tag{4.1}$$

- Let V_ℓ be the minimum number of HPFs required to obtain ℓ leukocytes.
  V_ℓ can be expressed in terms of Y_i as follows

$$V_\ell = \arg\min_t \{\sum_{i=1}^{t} Y_i \geq \ell\}$$

The probability mass function of V_ℓ is given by

$$\begin{aligned} P(V_\ell = t) &= P(R_t \geq \ell,\ R_{t-1} < \ell) \\ &= \sum_{r=0}^{\ell-1} P(Y_t \geq \ell - r) \cdot P(R_{t-1} = r) \end{aligned} \tag{4.2}$$

4.2.3.1 PD estimation

For method A, natural estimator of θ is used and the exacts formulas of ME, CV and FNR are given. However, the estimation of θ is not straightforward for the remaining methods (B, C, D). Hence, recurrence formulas are used to derive variability measures.

Let $\widehat{\theta}_A$ be the estimator of θ for Method A. Let n be the number of HPF read. Since $X_i \sim \mathcal{P}(\lambda_p)$ and X_i are iid, we have $S_n \sim \mathcal{P}(n\lambda_p)$. The number of parasite per field λ_p is then estimated by $\widehat{\lambda}_p$ where $\widehat{\lambda}_p = \frac{S_n}{n}$. Assuming the average amount of blood in each field as 0.002 μl [Dowling 1966], the PD is estimated by $\widehat{\theta}_A = \widehat{\lambda}_p \times 500$. Since $\mathbb{E}\left[\widehat{\theta}_A\right] = \theta$, $\widehat{\theta}_A$ is *unbiased*. Thus, the ME is null.

In order to evaluate the efficiency of this estimator, the variance is to be compared against the Fisher Information I(θ). The variance of this unbiased estimator is bounded by the inverse of the I(θ); namely the *Cramer-Rao* Bound (CRB). We show that the variance of the proposed estimation technique reaches the Cramer-Rao lower bound as follows

$$var(\widehat{\lambda}_p) \geq \frac{1}{nI(\lambda_p)}$$

where

$$I(\lambda_p) = -\mathbb{E}_{\lambda_p}\left[\frac{\partial^2 \log L(X, \lambda_p)}{\partial^2 \lambda_p}\right]$$

The log likelihood function is defined by

$$\begin{aligned} \mathcal{L}(x_i, \lambda_p) &= \log L(x_i, \lambda_p) \\ &= \log p(x_i, \lambda_p) \\ &= -\lambda_p - \log(x_i!) + x_i \log(\lambda_p) \\ \frac{\partial \log \mathcal{L}(x_i, \lambda_p)}{\partial \theta} &= -1 + \frac{x_i}{\lambda_p} \\ \frac{\partial^2 \log \mathcal{L}(x_i, \lambda_p)}{\partial^2 \theta} &= -\frac{x_i}{\lambda_p^2} \\ \mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(x_i, \lambda_p)}{\partial^2 \theta}\right] &= -\frac{1}{\lambda_p} \end{aligned}$$

Then, $I(\lambda_p) = \frac{1}{\lambda_p}$, which gives CRB $= \frac{\lambda_p}{n}$.

The variance of the estimator is defined by

$$\begin{aligned} var(\widehat{\lambda_p}) &= var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) \\ &= \frac{1}{n} var(X_i) \\ &= \frac{\lambda_p}{n} \end{aligned}$$

Hence, $var(\widehat{\lambda_p})$ reaches the *CRB*. Hence, $\widehat{\theta}_A$ is an *efficient* estimator of θ.

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to θ, which is equal to $\frac{1}{\sqrt{n\lambda_p}}$.

In practice, false negatives occur when diagnosing by mistake PD as null after reading n HPFs, i.e. $P(S_n = 0)$, which gives $FNR = e^{-n\lambda_p}$.

For Method B, ϕ is estimated by $\widehat{\phi}_B = \frac{S_{V_\ell}}{R_{V_\ell}}$. Then, θ is estimated by $\widehat{\theta}_B = \widehat{\phi}_B \times 8,000$.

To derive statistical properties of the PD estimate, we first need to compute the probability of seeing k parasites (resp. r leukocytes) in V_ℓ HPFs. These probabilities can be expressed as follows

$$\begin{aligned} P(S_{V_\ell} = k) &= \sum_t P(S_t = k) \cdot P(V_\ell = t) \\ P(R_{V_\ell} = r) &= \sum_t P(R_t = r) \cdot P(V_\ell = t) \end{aligned} \tag{4.3}$$

S_t and R_t are Poisson-distributed and the probability of V_ℓ is computed according to Equation (4.2).

The probability density function of $\widehat{\phi}_B$ is

$$P(\widehat{\phi}_B = s) = \sum_{\frac{k}{r}=s} P(S_{V_\ell} = k) \cdot P(R_{V_\ell} = r)$$

Let p be the required number of parasites for Method C. Let V_{ℓ_1} be the minimum number of HPF required to obtain ℓ_1 leukocytes and V_{ℓ_2} be the minimum number of HPF required to obtain ℓ_2 leukocytes. Let U_{ℓ_1,ℓ_2} be the minimum number of HPFs required to obtain p parasites. The probability mass function of U_{ℓ_1,ℓ_2} is as follows

$$P(U_{\ell_1,\ell_2} = t) = P(V_{\ell_1} = t) \cdot P(S_{V_{\ell_1}} \geq p) + P(V_{\ell_2} = t) \cdot P(S_{V_{\ell_1}} < p)$$

Probabilities of V_{ℓ_1} and V_{ℓ_2} are computed according to Equation (4.2).
Probability of $S_{V_{\ell_1}}$ is computed according to Equation (4.3).
Then, θ is estimated as follows $\widehat{\theta}_C = \widehat{\phi}_C \times 8,000$, where $\widehat{\phi}_C = \frac{S_{U_{\ell_1,\ell_2}}}{R_{U_{\ell_1,\ell_2}}}$.

The probability density function of $\widehat{\phi}_C$ is

$$P(\widehat{\phi}_C = s) = \sum_{\frac{k}{r}=s} P(S_{U_{\ell_1,\ell_2}} = k) \cdot P(R_{U_{\ell_1,\ell_2}} = r)$$

For Method D, $U_{\ell,p}$ denotes the minimum number of HPF required to obtain either ℓ leukocytes or p parasites, i.e. $U_{\ell,p} = \min(V_\ell, U_p)$. The probability mass function of $U_{\ell,p}$ is as follows

$$P(U_{\ell,p} = t) \quad = \quad P(V_\ell = t) \cdot P(U_p > t) + P(U_p = t) \cdot P(V_\ell > t)$$

Probabilities of U_p and V_ℓ are computed according to Equation (4.1) and Equation (4.2).

Then, θ is estimated by $\widehat{\theta}_D = \widehat{\phi}_D \times 8,000$, where $\widehat{\phi}_D = \frac{S_{U_{\ell,p}}}{R_{U_{\ell,p}}}$.

The probability density function of $\widehat{\phi}_D$ is

$$P(\widehat{\phi}_D = s) = \sum_{\frac{k}{r}=s} P(S_{U_{\ell,p}} = k) \cdot P(R_{U_{\ell,p}} = r)$$

4.2.3.2 Validation study

Simulations are used to study the accuracy of our mathematical models, and to validate the theoretical results derived from estimators' probability functions. For the purpose of simulations, we predefine a data-generating model of θ. Given λ_p and λ_ℓ, PD data are sequentially generated for each HPF. In that way, random samples of θ are generated under the Poisson assumption. Then, we investigate properties of sample means, variances and FNR. We use the statistical software package `R` to perform $10,000$ simulations. In each simulation step, we generate $1,000$ random drawings of θ and we save the sample ME, CV, FNR and cost in a vector. In that way, we were able to investigate the results of all simulation steps. We compare the simulated results to the theoretical ones. Simulations are computationally expensive. Then, it is burdensome to have to perform $10,000$ simulations to estimate each PD value according to methods A, B, C and D. Hence, computing the exact distribution of θ is a most useful alternative.

4.2.3.3 Colormaps

The recursive formulas described above are used to compute the exact distribution of variables. Each computation takes as input : λ_ℓ the number of leukocytes per field, λ_p the number of parasites per field and the threshold values (the number of HPFs or the number of WBCs or the number of parasites). The outputs are ME, CV, FNR and cost values. This approach is computationally expensive due to recursive formulas that precisely compute probability of getting r WBCs (resp. k parasites) according to each counting technique. These probabilities are used to compute $P(\widehat{\theta})$. Statistical properties of PD estimators are then derived. These

data sets are gridded into colormaps where the values taken by a variable (ME, CV, FNR, cost) in a two-dimensional table (X,Y) are represented as colors. Each rectangle in this grid is a pixel (or a color sample). This program sets each pixel to a color index according to its coordinates. Each pixel has an X and Y position where the X coordinate is the PD value and the Y coordinate is the threshold value. The X axis spans the range of 0 to 20,000 parasite per μl (400 values). The Y axis ranges from 0 to 500 (500 values). Hence, we used a resolution of 400×500 pixels. Contour lines are overlaid over the colormaps. A contour line connects points where the function has constant value. Linear interpolation is used in generating contour data. A higher resolution is needed to achieve a smoother mapping and to avoid artifacts (jagged contours), which arise due to interpolation.

We believe that if data mappings are addressed simultaneously in a single framework, the resulting approach will facilitate visual comparisons of methods. At that point, we consider the problem of scale. The methods use either different numbers of arguments or different types of arguments. We got rid of this by expressing variability measures for all methods as functions of parasity density and WBCs count. To do so, we convert threshold values used in each method into WBCs count. For Method A, we assume an average of $8,000$ WBCs per microliter of blood [WHO 2010a] and an average of 0.002 microliter of blood in each field [Dowling 1966]. The number of HPFs read n is then multiplied by $\lambda_\ell = 16$ WBCs to give the number of WBCs counted in n HPFs. For Method B, the thresold value is the number of WBCs counted ℓ. For Method C, we consider the case where $\ell_1 = \frac{p}{2}$ and we fix $\ell_2 = 500$. For Method D, we consider equal numbers of parasites p and leukocytes ℓ that have to be seen to stop the counting, hence $\ell = p$. Theses decisions are based on common use and can be considered reasonable assumptions. Moreover, this two-dimensional representation has the conceptual advantage of reducing the number of arguments and ensures a common approach for assessing methods performance.

4.3 Results

In the following, parasitemia was categorized as either low (PD $\leqslant$ 100 parasites/μl), intermediate (100 $<$ PD $<$ 10,000 parasites/μl) or high (PD $\geqslant$ 10,000 parasites/μl).

Method C involves three threshold values, which leads to a multidimensional problem. For this reason, we chose to express the parasite count in the first step as the half of the leukocyte count. We fixed the leukocyte count in the second step to 500 WBCs.

4.3.1 Impact of thresholds on variability measures

4.3.1.1 Mean error

The mean value of the Method A estimator equals the true value of the PD. Therefore this estimator is called an *unbiased* estimator. In addition to having the lowest variance among unbiased estimators (so called Minimum Variance Unbiased Estimator), this estimator also satisfies the *Cramér-Rao* bound, which is an absolute lower bound on variance for statistics of a variable and thus is an *unbiased efficient* estimator.

As shown in Figure 4.1, the ME of Method B only depends on threshold values. In Method B, parasites are counted until a fixed number of WBCs are seen and the number of parasites seen is not involved in the stopping rule of this counting process. Hence, the ME is independent from the PD. The colormap of the ME shows that the mean error decreases as threshold values increase. For instance, counting parasites until 400 WBCs instead of 200 WBCs decreases the bias by 0.5% of the PD. This can help choose the threshold value that allows to decrease the bias to a reasonable value.

For Method C, three parts can be distinguished. For PD $\leqslant$ 4,000 parasites/μl, contour lines are increasing functions of the PD and the thresholds. The darkest part on the map represents a constant ME. Due to the limited number of parasites in this area, counting is carried out until 500 WBCs are seen. Hence, Method C has similar behavior to Method B for WBCs= 500. By counting up to 500 WBCs, the mean error is fixed to 0.4%. For 1,000 $<$ PD $<$ 10,000 parasites/μl, ME values are represented by a set of bell-shaped density curves with a peak reached at PD = 4,000 parasites/μl. 4,000 is half the standard number of WBCs per μl. For Method C, the number of parasites counted is half the number of leukocytes. For PD $\leqslant$ 4,000 parasites/μl, increasing the PD increases the leukocyte count for a fixed ME value. If the microscopist wants to estimate for constant ME a higher PD, he needs to count more leukocytes. A higher threshold value is then required due to the small number of parasites present in this area. For PD $>$ 4,000, lower leukocyte counts are needed to maintain a constant ME value. A steady state will be reached afterwards whereby the ME is density independent. Due to the abundance of parasites, the ME only depends on the WBCs count. This steady-state region starts at PD = 6,650 parasites/μl for ME = 0.5%.

Note that the same ME level may be reached by more than one threshold value (eg. two contour lines for ME = 10%).

For PD $\leqslant$ 6,000 parasites/μl, the ME generated by Method D is density independent (see Figure 4.1). In this interval, leukocytes are more abundant than parasites. Hence, parasites are counted until a predetermined number of leukocytes is reached. We notice that increasing the leukocyte count will not significantly reduce the bias. For instance, counting parasites until 400 WBCs are seen, instead of 200, decreases the bias only by 0.5% of the PD. For 6,000 $<$ PD $<$ 10,000 parasites/μl, contour lines reach their minimum at PD = 8,000 parasites/μl. In this area, the number of leukocytes per field (λ_ℓ) and the number of parasites per field (λ_p) are very close. Lower threshold values are needed to maintain a constant ME. For high parasitemia, parasites are more numerous than leukocytes. The ME is therefore density dependent. If the microscopist wants to estimate for constant ME a higher PD, he needs to count more parasites.

4.3.1.2 Coefficient of variation

The CV of Method A is the inverse square root of λ_p times the number of fields (see Materials & Methods). If the counting does not exceed 12 HPFs (i.e. WBCs $\leqslant$ 200), CV values are higher than 9.94% of the real PD for low and intermediate parasitemias (see Figure 4.2). Above 20,000 parasites/μl, CV values are less than 3%. Counting up to 31 HPFs (i.e. WBCs $\approx$ 500) instead of 12 HPFs (i.e. WBCs $\approx$ 200), decreases the CV by approximately 10% of the PD.

For Method B, CV values lie midway between 10% and 20% of the PD for high parasitemias when WBCs $\geqslant$ 100 (see Figure 4.2). Notice that increasing the number of WBCs counted will not significantly decrease the CV for high parasitemias.

For Method C, the vertical lines indicate that the CV only depends on PD for low parasitemias. Due to the small number of parasites, CV levels are obtained by counting parasites until 500 WBCs are seen. Notice that CV values exactly match those obtained by Method B with WBCs = 500. Figure 4.2 shows bell-shaped patterns for higher densities (for 1,000 $<$ PD $<$ 10,000 parasites/μl) with a peak reached at PD = 4,000 parasites/μl. Along the same line as Method B, a constant CV level may be reached by more than one threshold value for PD $\geqslant$ 13,500 parasites/μl.

For Method D, the negative slope of the contour lines captures the indirect relationship between the threshold and the densities for PD $<$ 8,000 parasites/μl. For a fixed CV level, threshold values decline with density. In this interval, the counting stops when the fixed number of leukocytes (i.e: the threshold value) is obtained. The minimum is reached at PD = 8,000 parasites/μl. For PD $>$ 8,000 parasites/μl, positively sloped CV curves reflect the direct relationship between the threshold and the PD. In this area, parasites are more abundant than leukocytes. Therefore, the counting stops when the fixed number of parasites (i.e: the threshold value) is reached. If the microscopist wants to estimate with the same level of precision (i.e: for constant CV) a higher PD, he needs to count more parasites. A

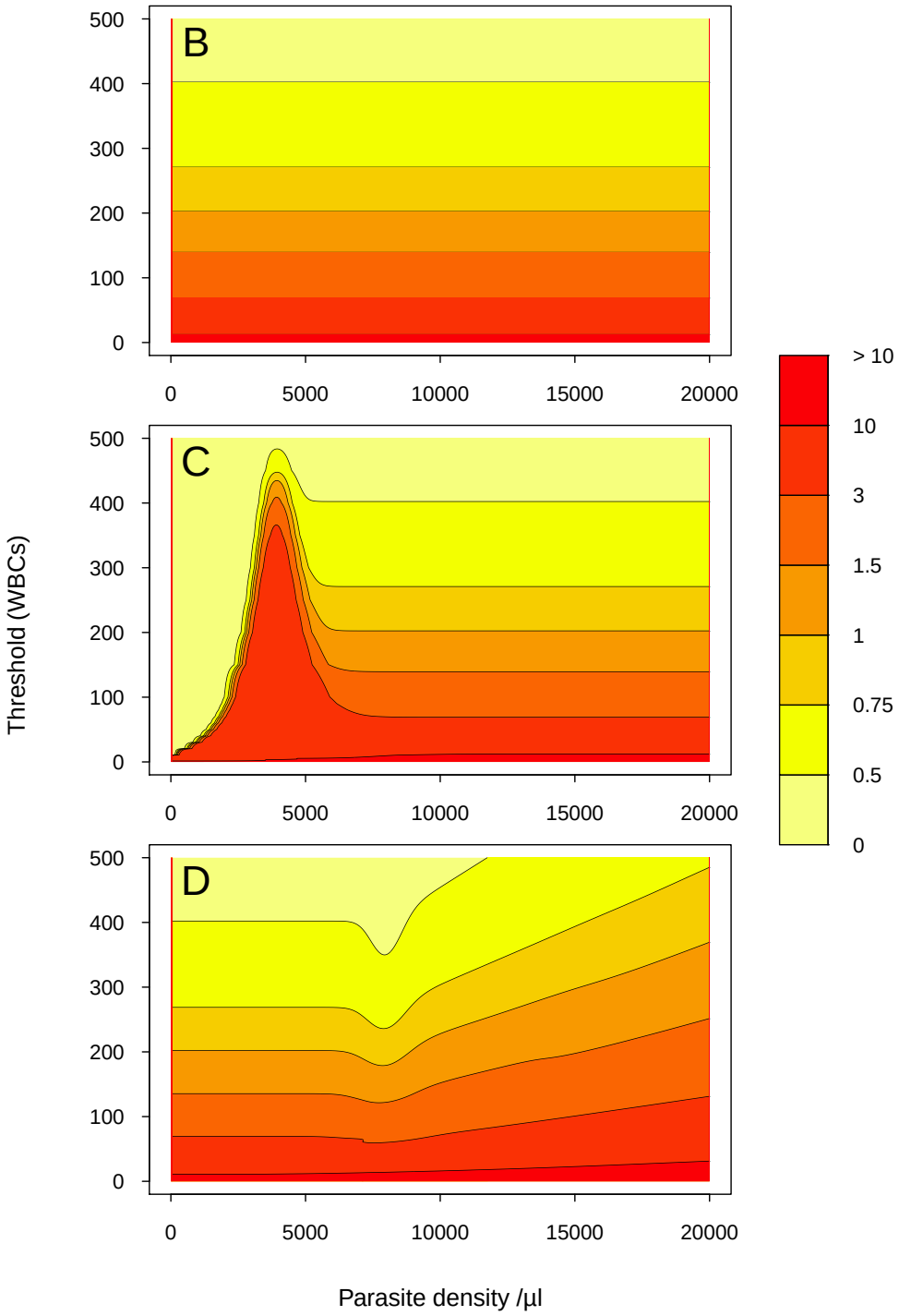

Figure 4.1: ***Mean error colormap.***
The colormap is drawn given a two-dimensional array of ME values. To allow for direct point-to-point numerical and visual comparison, we express the ME as a function of the parasite density (on the x-axis) and the WBC count (on the y-axis) in each of the four methods. Parasite density values are generated starting with 0, at increments of 50, and ending with 20,000. Threshold values (WBCs) are generated starting with 0, at increments of 1, and ending with 500. Then, each pixel is assigned a value that represents the ME-level. A color scale grading was applied to show levels. 7 degree intervals are depicted using a red-to-yellow colorspace with increasing intensity. We contour the ME at 0.5, 0.75, 1, 1.5, 3 and 10. The gaps between each pair of neighboring contour lines is filled with a color.

higher threshold value is then required.

4.3.1.3 False negative rates

For Method A, FNR decreases exponentially with increasing number of fields (n) and increasing number of parasites per field (λ_p) (see Materials & Methods). If the counting does not exceed 20 HPFs (i.e. WBCs $\leqslant 320$), the probability of misdiagnosis is high for low parasitemia levels (see Figure 4.3). For intermediate densities, this probability is less than 1% when the threshold is above 30 HPFs. For high parasitemias, false negatives occur much less frequently ($< 0.001\%$). Despite unbiasedness and efficiency, this estimator generates a high number of false negatives when the problem is difficult (low parasitemia).

Figure 4.3 shows that the FNRs of Method B vary from 5% to 80% for low parasitemia levels. For intermediate densities, this probability is less than 5%. False negatives do not occur for high parasitemia levels.

For Method C, the FNRs are threshold independent for PD $\leqslant 200$ parasites/μl and WBCs > 20. The number of false negatives arises from counting up to 500 WBCs. For $200 <$ PD $\leqslant 2,000$ parasites/μl and $10 \leqslant$ WBCs $\leqslant 20$, FNR values varies from 0.001% to 0.5%. For PD $\leqslant 2,000$ parasites/μl and WBCs < 10, FNR values are higher than 0.5%. False negatives do not occur for high parasitemia levels (PD $\geqslant 10,000$ parasites/μl).

For low parasitemias, we point out striking similarities between the FNRs in Method D and the FNRs in Method B. Due to the scarcity of parasites in this area, estimates are based on the leukocyte count in Method D.

4.3.1.4 Cost-effectiveness

Method A does not adapt to the variation of PD from one individual to another and costs a fixed HPFs number for all PD values.

The cost of Method B is an increasing linear function of the threshold values (see Figure 4.4). The cost here is independent from PD. This can be explained by the homogeneous distribution of leukocytes within the fields. Since we assumed a fixed number of leukocytes per field (λ_ℓ), the number of fields needed will indeed be independent of the PD.

For low parasitemia levels and WBCs > 5, the darkest color in the cost colormap of Method C indicates a constant cost of approximately 31 HPFs, which corresponds to the number of fields needed to reach 500 WBCs. For intermediate parasitemia levels, the cost varies depending on both the threshold value and how numerous the parasites are. The cost is independent from PD for high parasitemia levels. The number of fields needed is the ratio of WBCs to λ_ℓ.

Method D is highly adapted to parasitemia levels in terms of cost. For PD $\leqslant$ 8,000 parasites/μl, the cost is density independent. For PD $> 8,000$ parasites/μl, the cost decreases with density for a fixed threshold value. In this interval, parasites are more abundant than leukocytes. A lower number of fields is then needed to reach a predetermined threshold value.

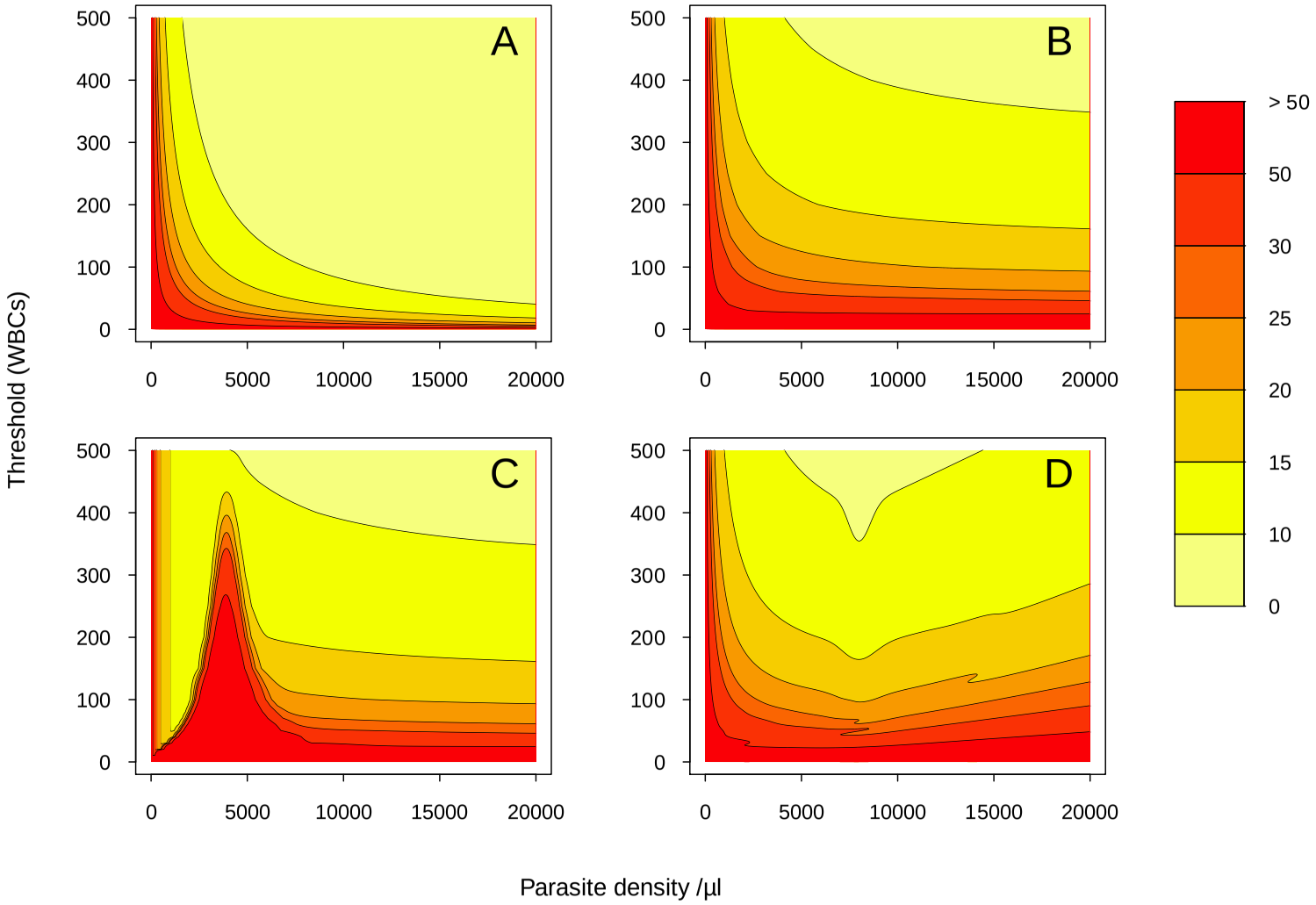

Figure 4.2: ***Coefficient of variation colormap.***

The colormap is drawn given a two-dimensional array of CV values. To allow for direct point-to-point numerical and visual comparison, we express the CV as a function of the parasite density (on the x-axis) and the WBC count (on the y-axis) in each of the four methods. Parasite density values are generated starting with 0, at increments of 50, and ending with $20,000$. Threshold values (WBCs) are generated starting with 0, at increments of 1, and ending with 500. Then, each pixel is assigned a value that represents the CV-level. A color scale grading was applied to show levels. 7 degree intervals are depicted using a red-to-yellow colorspace with increasing intensity. We contour the CV at 10, 15, 20, 25, 30 and 50. The gaps between each pair of neighboring contour lines is filled with a color.

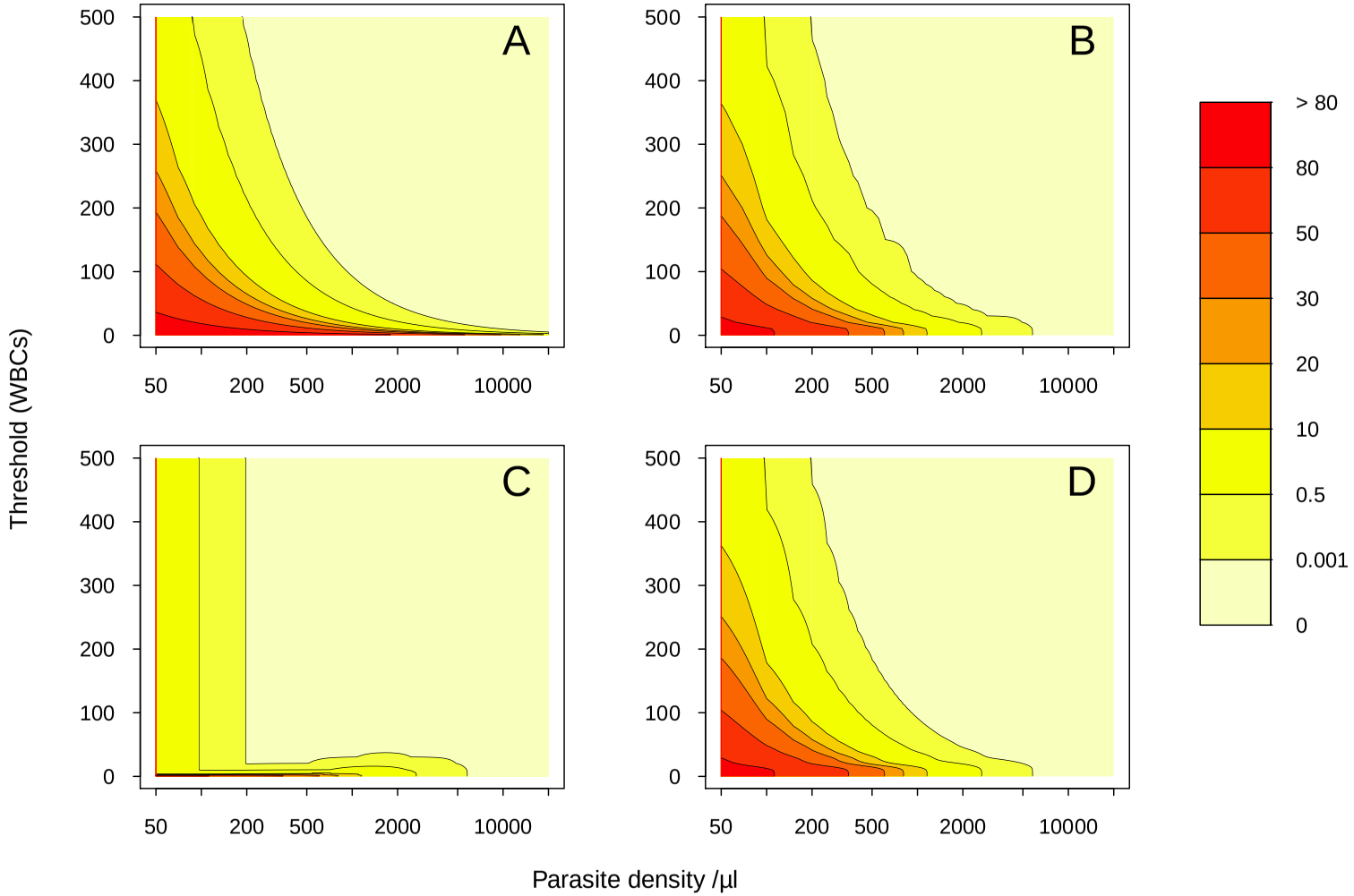

Figure 4.3: ***False negative rates colormap.***

The colormap is drawn given a two-dimensional array of FNR values. To allow for direct point-to-point numerical and visual comparison, we express the FNR as a function of the parasite density (on the x-axis) and the WBC count (on the y-axis) in each of the four methods. Parasite density values are generated starting with 0, at increments of 50, and ending with $20,000$. Threshold values (WBCs) are generated starting with 0, at increments of 1, and ending with 500. Then, each pixel is assigned a value that represents the FNR-level. A color scale grading was applied to show levels. 8 degree intervals are depicted using a red-to-yellow colorspace with increasing intensity. We contour the CV at 0.001, 0.5, 10, 20, 30, 50 and 80. The gaps between each pair of neighboring contour lines is filled with a color. A logarithmic scale is used on the x-axis and a linear scale is used on the y-axis.

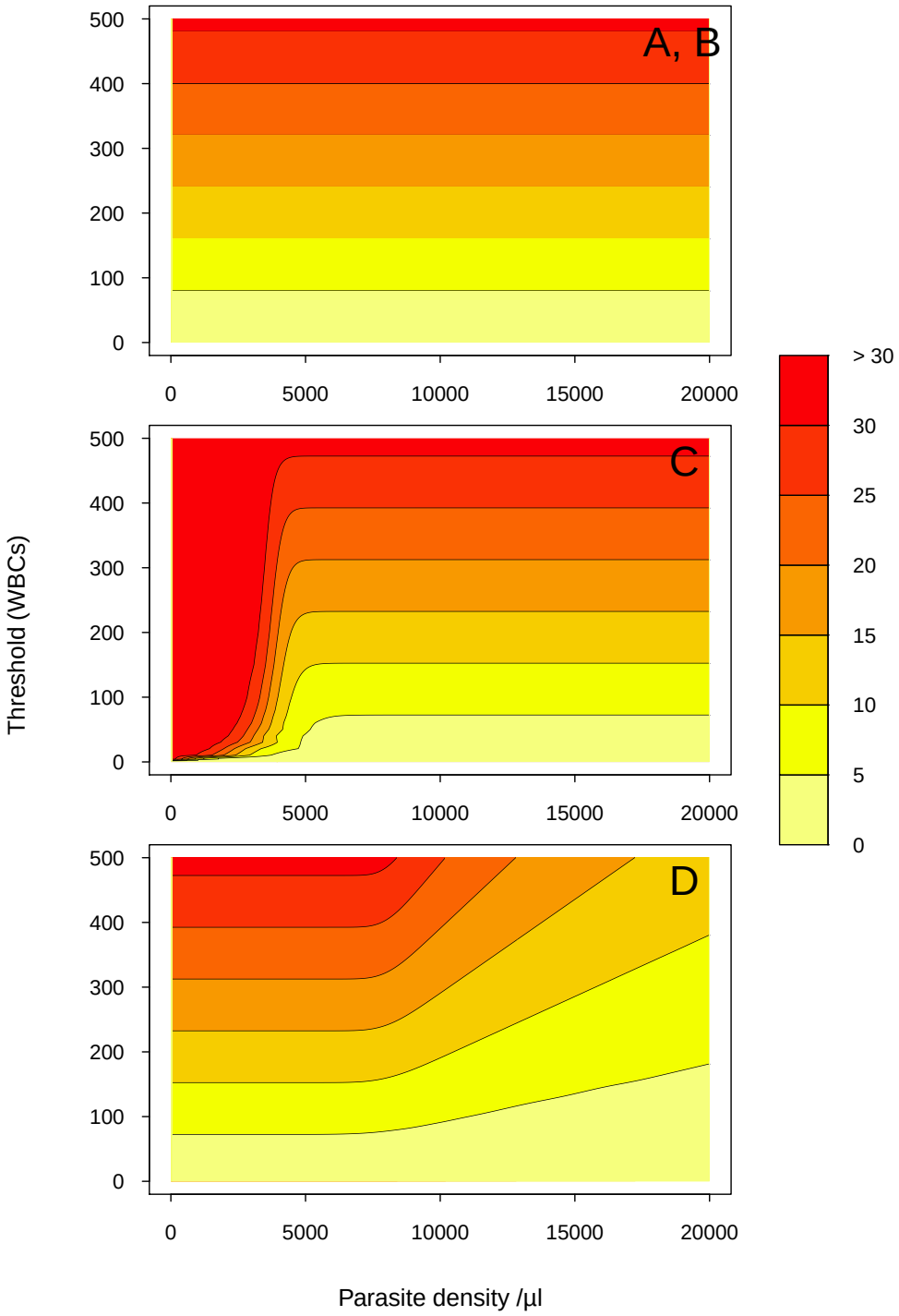

Figure 4.4: ***Cost-effectiveness colormap.***
The colormap is drawn given a two-dimensional array of cost values. To allow for direct point-to-point numerical and visual comparison, we express the cost as a function of the parasite density (on the x-axis) and the WBC count (on the y-axis) in each of the four methods. Parasite density values are generated starting with 0, at increments of 50, and ending with 20,000. Threshold values (WBCs) are generated starting with 0, at increments of 1, and ending with 500. Then, each pixel is assigned a value that represents the cost-level. A color scale grading was applied to show levels. 7 degree intervals are depicted using a red-to-yellow colorspace with increasing intensity. We contour the cost at 5, 10, 15, 20, 25 and 30. The gaps between each pair of neighboring contour lines is filled with a color.

4.3.2 Methods comparison for three parasitemia levels

To explore similarities and differences in method behaviors, we look more closely at the statistical properties of PD estimates. We choose three cut-offs for low (100 parasites/μl), intermediate (1,000 parasites/μl) and high (10,000 parasites/μl) parasitemias.

As Method A was shown to be unbiased, it was excluded from the ME analysis. As shown in Figure 4.5, Method B and Method D seem to have similar behaviors in terms of ME for low and intermediate parasitemias insofar as the two estimates are based on the leukocyte count in this density interval. For high parasitemias, Method B and Method C give the same results. The parasite count in Method C does not influence the accuracy of the method as long as parasites are numerous. For this reason, the two methods basically behave the same way.

To understand how the threshold values influence the variability of PD estimates, we plotted the CV according to threshold values. Figure 4.5 shows that the CV is highly sensitive to any variation of low thresholds ($\leqslant 100$). However, we see very few variations of the CV as threshold values increase (> 100). Both Methods B and D generate very close CV values for low and intermediate parasitemia levels. This result is expected since the number of WBCs seen is greater than the parasite number in the considered PD intervals. Hence, the two methods have the same stopping rules. For high parasitemias, Method B and Method C generate similar variability whereas Method A is significantly more precise than the other methods (B, C, D). However, Method A generates higher FNR for intermediate parasitemia than other methods when the count does not exceed 20 HPFs. For high PD levels, false negatives do not occur when the count exceeds 5 HPFs.

Figure 4.5 point out the high level of accuracy and precision performance of Method C for low and intermediate parasitemias. Thus, adding a supplementary stopping rule to the counting process and taking into account the parasite counts have enhanced the method performance, which raises questions regarding the repercussions in terms of cost-effectiveness. As shown in Figure 4.5, Method C is more expensive and time-consuming for low and intermediate parasitemias and requires constant cost (31.5 HPFs). As leukocytes are more present than parasites in the considered PD intervals the counting will be carried out until 500 WBCs are seen. Method A and B costs are density independent and increase linearly with threshold values. Method D outperformed the three other methods in terms of cost for high parasitemia levels.

4.3.3 Variability of measurements at equal cost-effectiveness

The duality between variability and cost illustrated in the previous section prompted a more detailed analysis of method performance differences at equal cost-effectiveness. In order to do this, we represent the variability measures as a function of cost. As shown in Figure 4.6, Methods B and D behave the same in terms of ME and CV for low and intermediate PD levels. For Method C, ME, CV and FNR

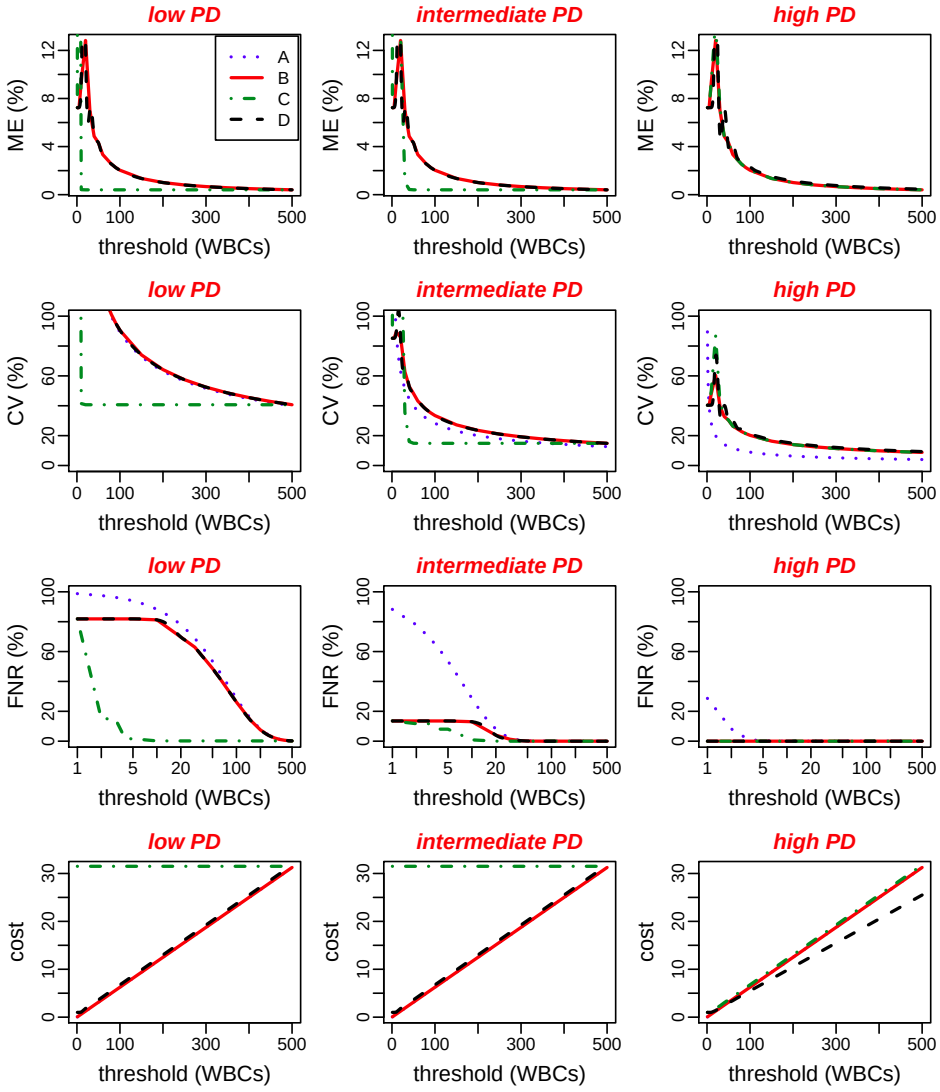

Figure 4.5: **Statistical properties of PD estimators cut-offs according to threshold values for three PD levels: low** (100 parasites/μl), **intermediate** (1,000 parasites/μl) **and high** (10,000 parasites/μl).
Variability measures (ME, CV, FNR) and cost are expressed as functions of the WBCs count (threshold) for the four methods (A, B, C, D). This graph gives the required number of WBCs for each method according to an expected amount of variability or cost, and favours a direct comparison between methods in terms of WBCs count. A logarithmic scale is used on the x-axis for FNR.

are density independent for low and intermediate parasitemias. For high PD levels, Methods B, C and D present similar results for ME, CV and FNR. Method A has the lowest CV values but generates higher FNR.

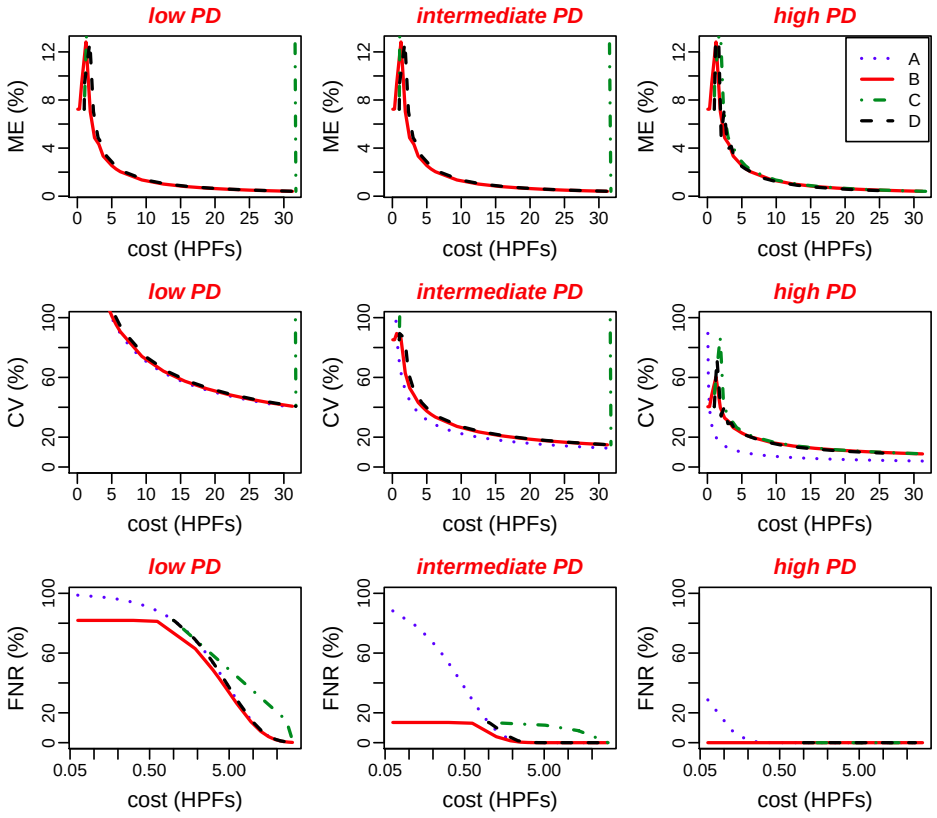

Figure 4.6: **Statistical properties of PD estimators cut-offs according to methods cost for three PD levels**: **low** (100 parasites/μl), **intermediate** (1,000 parasites/μl) **and high** (10,000 parasites/μl).
Variability measures (ME, CV, FNR) are expressed as functions of the cost (the number of HPFs needed to stop the counting) for the four methods (A, B, C, D). This graph gives the cost for each method according to an expected amount of variability, and favours a direct comparison between methods in terms of cost. A logarithmic scale is used on the x-axis for FNR.

4.3.4 Methods comparison for standards threshold values

To identify both similarities and differences between the commonly used threshold-based counting techniques, we estimate ME, CV, FNR and cost as a percentage of PD according to three parasitemia levels (low, intermediate, high) for commonly used threshold values. We used 200 HPFs for Method A, 200 WBCs for Method B, 100 parasites and 200 WBCs for Method C and 500 WBCs or 500 parasites for Method D.

Table 4.1 shows that Method A is the most efficient method in terms of accuracy

Low parasitemia $\theta = 100$ parasites/μl				
Method	A	B	C	D
ME (%θ)	0.00	1.01	0.40	0.40
CV (%θ)	15.81	64.20	40.68	40.68
FNR (%)	0.00	7.56	0.18	0.18
Cost	200	12.50	31.75	31.75
Intermediate parasitemia $\theta = 1,000$ parasites/μl				
Method	A	B	C	D
ME (%θ)	0.00	1.01	0.40	0.40
CV (%θ)	5.00	23.53	14.85	14.85
FNR (%)	0.00	0.00	0.00	0.00
Cost	200	12.50	31.75	31.75
High parasitemia $\theta = 10,000$ parasites/μl				
Method	A	B	C	D
ME (%θ)	0.00	1.01	1.01	0.45
CV (%θ)	1.58	14.03	14.03	9.31
FNR (%)	0.00	0.00	0.00	0.00
Cost	200	12.50	13	25.50

Table 4.1: **Threshold-based counting techniques comparison for low** (100 parasites/μl), **intermediate** (1,000 parasites/μl) **and high** (10,000 parasites/μl) **parasitemias**.

Measures of variability (ME, CV, FNR) and cost-effectiveness of methods are compared for fixed threshold values : 200 HPFs for Method A, 200 WBCs for Method B, 100 parasites and 200 WBCs for Method C, and 500 WBCs or 500 parasites for Method D.

(ME) and precision (CV), but has an important cost (200 HPF). Conversely, Method B is less accurate and precise than Method A while much more cost-effective (12.5 HPFs). Method C and Method D present similar properties for low and intermediate parasitemia levels. In fact, Method C behaves as Method D with a fixed leukocyte count (500 WBCs) due to the scanty presence of parasites. Hence, the mean error is density independent in these PD intervals. For high PD levels, Methods B and C behave the same. Due to the abundance of parasites, the enumeration is stopped when 200 WBCs are seen in both counting procedures. However, Method D is better suited to high parasitemia levels in terms of accuracy and precision compared to Methods B and C but results in up to a 2-fold increase in costs.

4.4 Discussion

To the best of our knowledge, this is the first study of threshold-based counting technique performance using the theoretical properties of PD estimators. We considered four commonly used threshold-based counting techniques, and assessed the performances of these methods according to threshold values. These thresholds may be fixed or variable. We showed that adaptative methods are more efficient than the ones involving fixed threshold values. To define the theoretical properties of the estimators we hypothesized that the distribution of parasites within HPFs follows a Poisson distribution. We demonstrated that Method A estimator is unbiased and efficient. However, this estimator generates a high number of false negatives, especially for low parasitemia levels when the counting does not exceed few HPFs. Moreover, Method A is time-consuming. We showed that the ME of Method B is independent from PD, and only depends on the threshold value. This helps to handle the amount of bias with an appropriate choice of the WBC threshold value. We showed that adding a new parameter to the stopping rules (the number of parasites seen) implies more accuracy and precision without increasing the method's cost for low and intermediate parasitemias. Method B and Method D have similar behaviors for low and intermediate parasitemia levels while Method D is more accurate and precise in the considered PD intervals. For high parasitemia levels, Method B and Method C have similar behaviors and are more accurate and precise than Method D. However, for high parasitemias, Method D outperformed the three other methods in terms of cost. For each method, different threshold values may be fixed, which raises questions regarding the accuracy and reproducibility of these parasite counting techniques.

The importance of parasite density data reproducibility stems from the need for epidemiological interpretations to be based on solid evidence. However, variation of parasite density within a slide is expected even when prepared from a homogeneous sample [Alexander 2010]. The source and scale of measurement error (sample preparation, staining process, counting technique, microscopist performance) have been investigated. The notion of inter-rater reliability is a source of concern in this context. It refers to a metric for raters' consistency that measures the degree of

agreement among raters. Many techniques were developed to measure inter-rater reliability. Some reports deal with the variability in the methods for detecting and counting parasites in thick smears. They attempt to evaluate the inter-rater reliability of malaria microscopy in epidemiological studies by looking at the variation of results due to the microscopist's reading. The variability of these methods has been assessed using statistical approaches [Dowling 1966, Trape 1985, Bland 1986, Payne 1988, Greenwood 1991, Clendennen 1995, Mulder 1998, Dubey 1999, Prudhomme O'Meara 2005, Alexander 2010]. These methods used several criteria to assess the inter-rater reliability and to quantify the degree of agreement between malaria slide density readings. For continuous data, Analysis of Variance (ANOVA) is the method of choice. Bland & Altman (1986) [Bland 1986] plotted the differences in log-transformed data versus average in mean counts. They expanded on this idea by plotting the difference of each point, the mean difference, and the confidence limits on the vertical axis against the average of the two ratings on the horizontal axis. The resulting Bland & Altman plot [Bland 1986] demonstrates not only the overall degree of agreement, but also whether the agreement is related to the underlying value of the item. For instance, two raters might nearly agree in estimating the size of small items, but disagree about larger ones. Alexander et al. (2010) [Alexander 2010] assessed agreement between replicate slide readings of malaria parasite density using as criterion the repeatability, that is to say the value below which the absolute difference between results may be expected to lie with a 95% probability [Braun-Munzinger 1992]. This metric is linked to Bland & Altman limits of agreement [Bland 1986]. It is half the distance between the upper and lower limits of agreements. For nominal data, the *kappa* coefficient of *Cohen* [Fleiss 1973] and its many variants and the *Scott*'s *pi* [Scott 1955] are the preferred statistics.

However, very few studies have examined the threshold-based counting techniques or evaluated the impact of the sampling error in endpoint measurements. In Nigeria, Dowling & Shute (1966) [Dowling 1966] showed that only 43% of infections in adults were detected by examining 200 fields, 61% by examining 600 fields and 70% by examining 1,000 fields. In the Garki Project, Molineaux & Gramiccia (1980) [Molineaux 1980] showed that the prevalence observed by the examination of 400 HPFs compared to 200 HPFs was increased by 10% for *P. falciparum*, by 24% for *P. malariae* and by 21% for *P. ovale*. Trape (1985) [Trape 1985] compared the results of the examination of 100 and 200 fields of the thick film in 245 schoolchildren aged 6 to 16 from Linzolo (Congo). He concluded that the systemic examination of 200 oil immersion fields of the thick smear is the best compromise between the need for precision and rapidity. Prudhomme O'Meara et al. (2006) [Prudhomme O'Meara 2006b] showed empirically that counting beyond 200 WBCs may not significantly improve parasite density measurements.

In addition, the accuracy and consistency of these methods have been generally overlooked. There is no general agreement on the optimal method for estimating parasite density according to threshold values. Further experimental evidence is needed to determine which parasite counting technique is most accurate, repro-

ducible, and efficient. Ultimately, the question is: to which extent would threshold values (specifically the number of WBCs counted and HPFs seen) influence the variability in parasite density estimates? However, there remains the issue of homogeneity. The distribution of the thickness of the smear and hence the distribution of parasites within the smear is not completely homogeneous [Alexander 2010]. Therefore, a proportion of the variability may be explained by this homogeneity factor.

To understand how the thresholds involved in parasite enumeration methods contribute to the magnitude of discrepancies in density determination, we studied their impact in variability measures generated by commonly used threshold-based counting techniques. We showed that estimators perform quite differently according to threshold values, and that an overall performance measure probably hides a lot of complexity in the behavior of each estimator. Another important aspect of this study is that we observed how estimators perform at different parasitemia levels, and how much the choice of threshold values may influence the performance of estimators relative to each parasitemia level.

In summary, while all four estimators had some deficiencies, Method D outperformed all the other estimators for accuracy, precision measures and cost-effectiveness, and should therefore be seriously considered in future studies of comparative performance of PD estimators with field-collected data. In this chapter, we explored the duality between cost-effectiveness and precision implied by estimation methods. An open question remains: To what extent is it possible to reduce methods' cost while staying accurate and precise in estimation measures?

In further support of the arguments cited in this chapter, empirical validation of the theoretical results is needed through a rereading experience conducted in the field. And toward a better understanding of threshold effects, we are interested in the study of the consequences of the quality of these estimators in models classically used and starting from these measures (mixed effects linear and logistic regression, generalized linear models, etc).

Chapter 5

EM for Mixtures and HMMs

"I'm not very good with numbers."

Arthur Dempster

Contents

The purpose of this chapter is to introduce the tools that are used in Chapter 6 to fit the distribution of parasites and leukocytes per HPF. The EM algorithm is presented with applications to mixture models and Hidden Markov Models (HMMs). A thorough introduction to HMMs with many applications can be found in [Zucchini 2009].

Notation

$X = X_{1:T}$	Observed variables.
X_t	The value of X at time t.
$X_{t:t'}$	Vector of observations $(X_t, ..., X_{t'})$.
$S = S_{1:T}$	Latent (unobserved) variables.
$\Theta^{(k)}$	The estimate of the parameters at iteration k.
$\log P(X \mid \Theta)$	The marginal log-likelihood.
$P(S \mid X, \Theta)$	The posterior distribution.
$\log P(X, S \mid \Theta)$	The complete data log-likelihood (CDLL).
$Q(\Theta \mid \Theta^{(k)})$	The expected CDLL $\sum_S P(S \mid X, \Theta^{(k)}) \log P(X, S \mid \Theta)$.

5.1 Introduction

The expectation-maximization (EM) algorithm is a numerical method for performing maximum likelihood estimation in missing data problems [Dempster 1977].

For a statistical model which is specified through a set of observed data X, a set of unobserved latent data S, and a vector of unknown parameters Θ, along with a likelihood function $L(\Theta \mid X, S) = P(X, S \mid \Theta)$, the maximum likelihood estimate (MLE) of Θ is determined by maximizing the marginal likelihood of the observed data

$$L(\Theta \mid X) = P(X \mid \Theta) = \sum_S P(X, S \mid \Theta) \tag{5.1}$$

hence

$$\hat{\Theta}_{MLE} = \arg\max_{\Theta} L(\Theta \mid X)$$

Maximizing $L(\Theta \mid X)$ can be quite tedious because it contains a sum over a large number of S configurations. The EM algorithm allows to circumvent this problem. The algorithm estimates parameters of model Θ that maximize the incomplete data log-likelihood, $\log P(X \mid \Theta)$, by iteratively maximizing the expectation of the complete data log-likelihood, $\log P(X, S \mid \Theta)$. The expected CDLL, with respect to the conditional distribution of S given X, is defined in the EM as an auxiliary function, Q, of current parameter set $\Theta^{(k)}$ and new parameter set Θ given by

$$\begin{aligned} Q(\Theta \mid \Theta^{(k)}) &= \mathbb{E}_{S|X,\Theta^{(k)}}[\log L(\Theta \mid X, S)] \\ &= \sum_S P(S \mid X, \Theta^{(k)}) \log P(X, S \mid \Theta) \end{aligned} \tag{5.2}$$

If the CDLL is factorizable, optimizing the Q-function could be much easier than optimizing the log-likelihood.

Each iteration consists of an expectation (E) step and a maximization (M) step. After choosing starting values for Θ, the algorithm proceeds as follows :

- **E-step** : Compute $Q(\Theta \mid \Theta^{(k)})$, which gives the conditional expectations of the unobserved data given the observations and given the current estimate of Θ.
- **M-step** : Maximize $Q(\Theta \mid \Theta^{(k)})$, using instead of the unobserved values their conditional expectations from the E-step, that is, solve the optimization problem

$$\Theta^{(k+1)} = \arg\max_{\Theta} Q(\Theta \mid \Theta^{(k)})$$

These two steps are repeated until convergence. A fixed stopping rule determines in advance the desired accuracy of the estimation. For instance, the algorithm may be stopped when $\sum_i \mid \theta_i^{(k)} - \theta_i^{(k-1)} \mid < c$, where c is the convergence criterion. Defining this stopping rule and the starting values for Θ are crucial. The algorithm is

conceptually simple and easy to implement. Under mild conditions (e.g. exponential families in [Dempster 1977]), each iteration k of the algorithm is guaranteed to increase the log-likelihood $L(\Theta^{(k)} \mid X)$, and $\Theta^{(k)}$ is guaranteed to converge to a $\hat{\Theta}_{MLE}$.

5.2 EM Algorithm for Mixture Models

5.2.1 Mixture models

We assume that X belongs to a heterogeneous population consisting of m homogeneous subpopulations. We assume that, for $t \in [\![1\,;T]\!]$ and $i \in [\![1\,;m]\!]$, X_t is distributed in the i^{th} component with the probability $p_i(X_t \mid \theta_i)$. Let δ_i be the proportion of the i^{th} component, such that $\sum_{i=1}^{m} \delta_i = 1$. Hence, the marginal probability is

$$P(X_t \mid \Theta) = \sum_{i=1}^{m} \delta_i p_i(X_t \mid \theta_i)$$

The marginal mean of the independent mixture is given by

$$\mathbb{E}(X_t) = \sum_{i=1}^{m} \delta_i \mathbb{E}\left(X_t \mid \theta_i\right)$$

To compute the unconditional variance of the mixture, we use the law of total variance

$$V(X_t) = \mathbb{E}[V(X_t \mid \Theta)] + V[\mathbb{E}(X_t \mid \Theta)]$$

The expected value of conditional variances is given by

$$\mathbb{E}[V(X_t \mid \Theta)] = \sum_{i=1}^{m} \delta_i V\left(X_t \mid \theta_i\right)$$

The variance of the conditional means is given by

$$V[\mathbb{E}(X_t \mid \Theta)] = \sum_{i=1}^{m} \delta_i \mathbb{E}\left(X_t \mid \theta_i\right)^2 - \left(\sum_{i=1}^{m} \delta_i \mathbb{E}(X_t \mid \theta_i)\right)^2$$

In the case of a two-component mixture model with weights δ_i, means μ_i and variances σ_i^2, the total mean and variance will be

$$\mathbb{E}(X_t) = \delta_1\mu_1 + \delta_2\mu_2 \tag{5.3}$$

$$V(X_t) = \delta_1\sigma_1^2 + \delta_2\sigma_2^2 + \delta_1\delta_2(\mu_1 - \mu_2)^2 \tag{5.4}$$

We will show later that the variance of the mixture model is greater than its expectation, which allows to account for overdispersion in data.

The incomplete-data log-likelihood expression is given by

$$\log L(\Theta \mid X) = \log \prod_{t=1}^{T} P(X_t \mid \Theta) = \sum_{t=1}^{T} \log \left(\sum_{i=1}^{m} \delta_i p_i(X_t \mid \theta_i) \right)$$

The incomplete-data log likelihood may be difficult to maximize. The numerical difficulty is due to the sum inside the log. However, if we assume that observations X are incomplete and that they are generated by an unobserved process S, the likelihood expression can be dramatically simplified, which motivates the use of the EM algorithm. Before we proceed to the computation of $Q(\Theta \mid \Theta^{(k)})$ in (5.2), we first need to derive the distribution of the complete data and the distribution of the unobserved data. We have

$$\begin{aligned} \log P(X, S \mid \Theta) &= \sum_{t=1}^{T} \log \left(P(X_t = x_t \mid s_t) P(S_t = s_t \mid \Theta) \right) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{m} \mathbf{1}_{\{s_t = i\}} \log \left(\delta_i p_i(x_t \mid \theta_i) \right) \end{aligned} \tag{5.5}$$

and

$$P(S \mid X, \Theta^{(k)}) = \prod_{t=1}^{T} P(s_t \mid x_t, \Theta^{(k)}) \tag{5.6}$$

where

$$P(s_t = i \mid x_t, \Theta^{(k)}) = \frac{\delta_i^k p_i(x_t \mid \theta_i^{(k)})}{p(x_t \mid \Theta^{(k)})} = \frac{\delta_i^k p_i(x_t \mid \theta_i^{(k)})}{\sum_{j=1}^{m} \delta_j^k p_j(x_t \mid \theta_j^{(k)})}$$

5.2.2 The E-step

In the context of finite mixtures, the Q-function can be rewritten from (5.2) using (5.5) and (5.6) as

$$\begin{aligned} Q(\Theta \mid \Theta^{(k)}) &= \sum_{S} P(S \mid X, \Theta^{(k)}) \log P(X, S \mid \Theta) \\ &= \sum_{s} \sum_{t=1}^{T} \sum_{i=1}^{m} \mathbf{1}_{\{s_t = i\}} \log \left(\delta_i p_i(x_t \mid \theta_i) \right) P(s_t \mid x_t, \Theta^{(k)}) \\ &= \sum_{i=1}^{m} \sum_{t=1}^{T} \log(\delta_i p_i(x_t \mid \theta_i)) P(s_t = i \mid x_t, \Theta^{(k)}) \\ &= \sum_{i=1}^{m} \sum_{t=1}^{T} \log(\delta_i) P(s_t = i \mid x_t, \Theta^{(k)}) \\ &\quad + \sum_{i=1}^{m} \sum_{t=1}^{T} \log(p_i(x_t \mid \theta_i)) P(s_t = i \mid x_t, \Theta^{(k)}) \end{aligned} \tag{5.7}$$

5.2.3 The M-step

Equation (5.7) can be decomposed in two parts. We maximize the first part with respect to δ_i, the second part with respect to θ_i (λ_i for the Poisson distribution, r_i and π_i for the negative binomial (NB) distribution).

We use a Lagrange multiplier to find the expression of δ_i, since $\sum_{i=1}^{m} \delta_i = 1$. Maximizing the Q-function in (5.7), subject to the constraint $\sum_{i=1}^{m} \delta_i = 1$, comes down to solving the following equation

$$\frac{\partial}{\partial \delta_i} \left[\sum_{i=1}^{m} \sum_{t=1}^{T} \log(\delta_i) P(s_t = i \mid x_t, \Theta^{(k)}) + \lambda \left(\sum_{i=1}^{m} \delta_i - 1 \right) \right] = 0$$

Then

$$\sum_{t=1}^{T} \frac{1}{\delta_i} P(s_t = i \mid x_t, \Theta^{(k)}) + \lambda = 0 \tag{5.8}$$

or

$$\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) = -\lambda \delta_i$$

Summing over m yields

$$\sum_{i=1}^{m} \sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) = -\lambda \sum_{i=1}^{m} \delta_i = -\lambda$$

As $\sum_{i=1}^{m} P(s_t = i \mid x_t, \Theta^{(k)}) = 1$, we get $\lambda = -T$.
The maximizing value of δ_i from Equation (5.8) is

$$\hat{\delta}_i = \frac{1}{T} \sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})$$

5.2.3.1 M-step for Poisson mixture

Under the Poisson assumption, the maximization is computationally tractable. Since

$$p_i(x_t \mid \theta^{(k)}) = e^{-\lambda_i} \frac{\lambda_i^{x_t}}{x_t!}$$

differentiating the second term in (5.7), with respect to λ_i, and equating to zero yields

$$\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) \left(-1 + \frac{x_t}{\lambda_i}\right) = 0$$

It follows immediately that

$$\hat{\lambda}_i = \frac{\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) x_t}{\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})}$$

Note that the Poisson mixture model is able to accommodate overdispersion better than the Poisson model with one component. For a two-state Poisson mixture, it follows immediately from (5.3) and (5.4) that the variance exceeds the mean by $\delta_1\delta_2(\lambda_1 - \lambda_2)^2$.

5.2.3.2 M-step for NB mixture

Different parameterizations for the negative binomial distribution exist. We choose the distribution function given by

$$p_i(x_t \mid \Theta^{(k)}) = \frac{\Gamma(x_t + r_i)}{\Gamma(x_t + 1)\Gamma(r_i)} {\pi_i}^{r_i} (1 - \pi_i)^{x_t}$$

where Γ denotes the Gamma-function; $r_i > 0$ and $\pi_i \in [0; 1]$ are the parameters of the NB.

We rewrite the Gamma-functions as $\exp\log(\Gamma)$ in the second part of (5.7) as follows

$$\begin{aligned}\sum_{i=1}^{m}\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) \log(p_i(x_t \mid \theta_i)) \quad = \quad & \sum_{i=1}^{m}\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) \\ & (\log\Gamma(x_t + r_i) - \log\Gamma(r_i) - \log\Gamma(x_t) \\ & + r_i \log\pi_i + x_t \log(1 - \pi_i)) \qquad (5.9)\end{aligned}$$

Differentiating (5.9) with respect to π_i and equating the derivative to zero yields

$$\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) \left(\frac{r_i}{\pi_i} - \frac{x_t}{1 - \pi_i}\right) = 0$$

That is

$$\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})(r_i - \pi_i(r_i + x_t)) = 0$$

The solution is as follows

$$\pi_i = \frac{r_i \sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})}{\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})(r_i + x_t)} \tag{5.10}$$

Maximizing (5.9) with respect to r_i gives

$$\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) \left(\psi(r_i + x_t) - \psi(r_i) + \log \pi_i\right) = 0 \tag{5.11}$$

where

$$\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x} = \frac{\Gamma'(x)}{\Gamma(x)}$$

Substituting π_i from Equation (5.10) in Equation (5.11) yields

$$\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)}) \left(\psi(r_i + x_t) - \psi(r_i) + \log \left[\frac{r \sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})}{\sum_{t=1}^{T} P(s_t = i \mid x_t, \Theta^{(k)})(r_i + x_t)}\right]\right) = 0$$

Accurate solution of r_i is obtained with a direct numerical maximization using `optim` in `R` [Nelder 1965]. The maximizing value of r_i is then substituted in (5.10) to derive π_i.

5.2.4 Example

We consider a Poisson mixture model Θ with $\delta = (0.1, 0.4, 0.5)$ and $\lambda = (10, 20, 30)$. This model is composed of three sub-populations. We generate a sample of $1,000$ observations from Θ model. We apply the EM algorithm to fit the simulated data. We use the k-means algorithm to set the initial starting values for the EM algorithm [Hartigan 1979]. The algorithm is stopped when $\sum_i \mid \theta_i^{(k)} - \theta_i^{(k-1)} \mid < 0.001$

After 36 iterations, the EM algorithm provides fair estimate of Θ with $\hat{\delta} = (0.0949, 0.4244, 0.4808)$ and $\hat{\lambda} = (10.0410, 19.6433, 30.2550)$. The 36 iterations are plotted in Figure 5.1. In this example, the EM algorithm works efficiently and converges to the same estimates under different starting values. In the case of NB mixtures, however, the EM algorithm is more sensitive to the choice of initial values, and it should be run several times using different starting values to avoid convergence to a local minimum.

5.3 EM Algorithm for Hidden Markov Models

5.3.1 HMMs

For a sequence data, the assumption of independent samples is too restrictive. The statistical dependence between sets of data may hide critical information. Hidden Markov models (HMMs) are a kind of mixture models where the mixing distribution is a Markov chain, in which, given present, the future is independent of

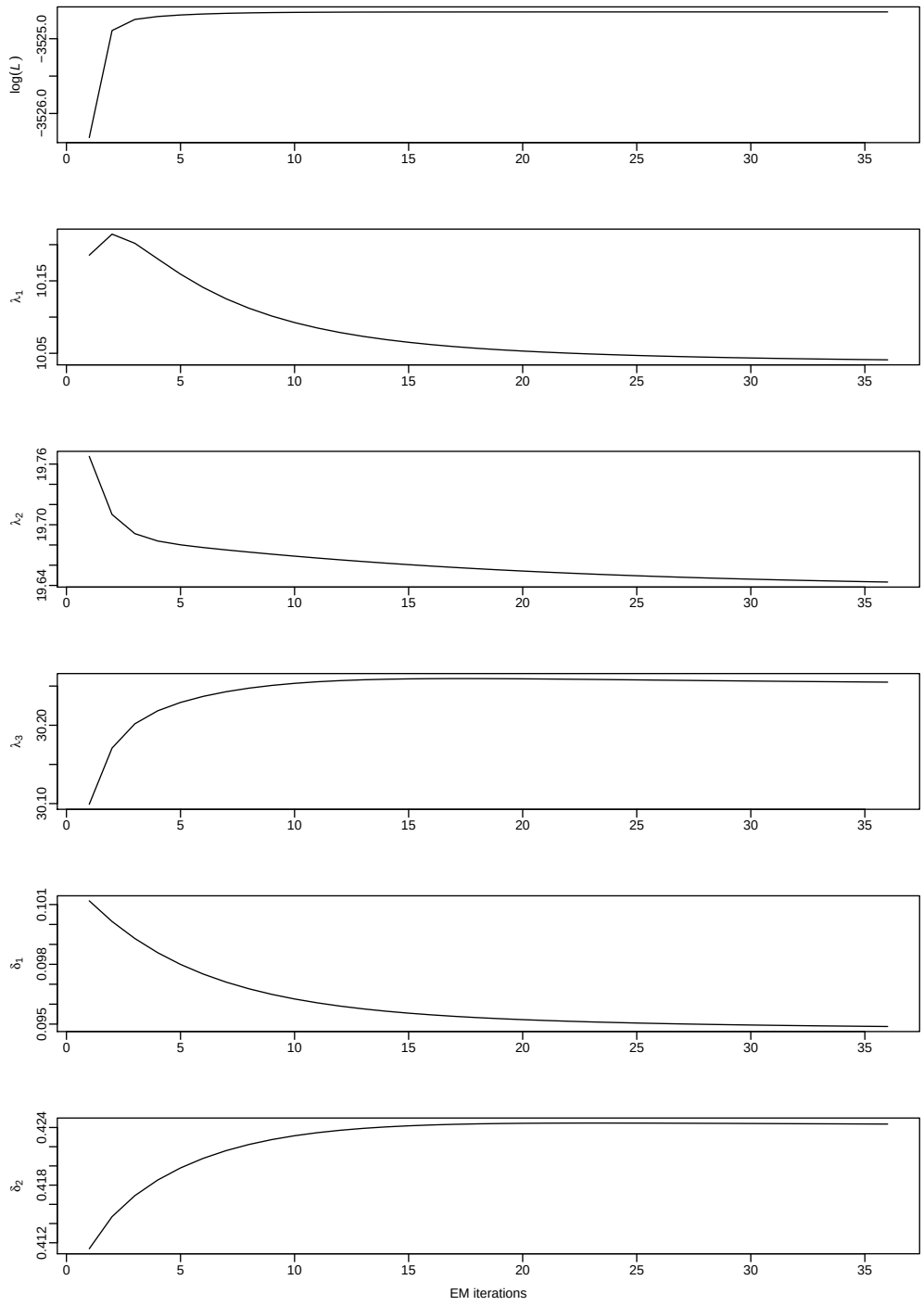

Figure 5.1: ***EM algorithm convergence.***
The EM algorithm is applied to a three states Poisson mixture model Θ in Example 5.2.4. After 36 iterations, the EM algorithm converges to $\hat{\delta} = (0.0949, 0.4244, 0.4808)$ and $\hat{\lambda} = (10.0410, 19.6433, 30.2550)$. The model log-likelihood $\log(L)$ is -3524.6410.

the past. Hidden states are treated as missing data in the estimation of HMM parameters. HMMs are an effective tool for modelling the dependence structure in data.

The model is composed of an observed sequence $\{X_t : t \geq 1\}$ and an unobserved (hidden) sequence $\{S_t : t \geq 1\}$. An observation X_t is generated by a hidden state S_t. Given the state S_t, the observation X_t is independent of other observations and states and only depends on the current state S_t. For a fixed state, the observation X_t is generated according to a fixed probability. If the markov chain $\{S_t : t \geq 1\}$ has m states, $\{X_t : t \geq 1\}$ is called an *m-state* HMM. The process can be drawn as a diagram of states (nodes) and transitions (edges) (see Figure 5.2).

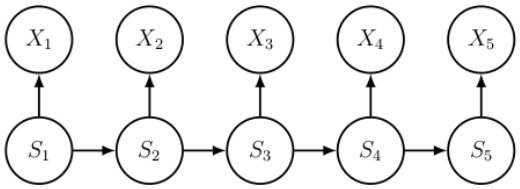

Figure 5.2: ***Basic HMM architecture.***
S_1, ..., S_5 are the hidden states and X_1,...,X_5 are the generated observations. Observations are independent of each other and only depend on the current state.

The HMM parameters are transition probabilities and emission probabilities. The transition probabilities $\gamma_{ij} = P(s_t = j \mid s_{t-1} = i)$ control the way the hidden state at time t is chosen given the hidden state at time $t-1$. The process can remain in the state i with probability γ_{ii}. Γ denotes the state transition matrix. The emission probabilities $P(X_t = x_t \mid S_t = i) = p_i(x_t)$ govern the distribution of the observed variable X at time t given the state of the hidden variable at that time.

$\delta_i = P(S_1 = i)$ denotes the initial state probability that specifies the starting state. δ_i may be fixed by specifying a particular state as starting state or under the stationary assumption of HMMs, $\delta = \delta\Gamma$.

We can summarize parameters features by

$$P(S_t \mid S_{1:(t-1)}) = P(S_t \mid S_{t-1}) \quad \forall t \in [\![2\,;T]\!]$$
$$P(X_t \mid X_{1:(t-1)}, S_{1:t}) = P(X_t \mid S_t) \quad \forall t \in [\![2\,;T]\!]$$

The complete-data likelihood is given by

$$P(X, S \mid \Theta) = P(S_1 \mid \Theta) \prod_{t=2}^{T} P(S_{t-1}, S_t \mid \Theta) \prod_{t=1}^{T} P(X_t \mid S_t, \Theta)$$

The CDLL may be straightforward to maximize even if the maximization of the observed data likelihood is tedious. This claim motivates the use of the EM algorithm to fit the HMMs.

5.3.2 Forward and backward probabilities

In order to apply the EM algorithm to HMMs, we need to compute the following probabilities

$$\begin{aligned}\sigma_i(t) &= P(S_t = i \mid X = x)\\ \phi_{ij}(t) &= P(S_{t-1} = i, S_t = j \mid X = x)\end{aligned}$$

To do so, we shall first define the forward probabilities, $\alpha_i(t)$, and the backward probabilities, $\beta_i(t)$.

Definition 1. The forward probability $\alpha_t(i)$ is the probability of the HMM emitting the output symbols $X_{1:t}$, and then ending up in state i at time t.

$$\alpha_t(i) = P(X_{1:t} = x_{1:t}, S_t = i) \quad \forall i \in [\![1\,;m]\!] \quad \forall t \in [\![1\,;T]\!]$$

Definition 2. The backward probability $\beta_t(i)$ is the probability of emitting symbols $X_{(t+1):T}$, then ending up in the final state, given the state at time t is i.

$$\begin{aligned}\beta_i(t) &= P(X_{t+1:T} = x_{t+1:T} \mid S_t = i) && \forall i \in [\![1\,;m]\!] \quad \forall t \in [\![1\,;T]\!]\\ \beta_i(T) &= 1 && \forall i \in [\![1\,;m]\!]\end{aligned}$$

Theorem 1. *Given a state sequence $\{S_t : t \in [\![1\,;T]\!]\}$ and an observed sequence $\{X_t : t \in [\![1\,;T]\!]\}$, the probability that X visits the state i at the time t is given by*

$$P(X = x, S_t = i) = \alpha_t(i)\beta_t(i) \tag{5.12}$$

proof 1.

$$\begin{aligned}P(X_{1:T}, S_t = i) &= P(X_{1:t}, X_{(t+1):T}, S_t = i)\\ &= P(X_{1:t}, X_{(t+1):T} \mid S_t = i)P(S_t = i)\\ &= P(X_{1:t} \mid S_t = i)P(X_{(t+1):T} \mid S_t = i)P(S_t = i)\\ &= P(X_{1:t}, S_t = i)P(X_{(t+1):T} \mid S_t = i)\\ &= \alpha_t(i)\beta_t(i)\end{aligned}$$

Proposition 1. *Summing Equation (5.12) over m yields*

$$\sum_{i=1}^{m} \alpha_t(i)\beta_t(i) = P(X = x)$$

Theorem 2. *The probability that X visited the state j at the time $t-1$ and enters the state i at time t is given by*

$$P(X = x, S_{t-1} = i, S_t = j) = \alpha_i(t-1)p_j(x_t)\gamma_{ij}\beta_j(t) \tag{5.13}$$

proof 2.

$$\begin{aligned}
P(X_{1:T} = x_{1:T}, S_{t-1} = i, S_t = j) &= P(X_{1:t-1}, X_{t:T}, S_{t-1} = i, S_t = j) \\
&= P(X_{1:t-1}, S_{t-1} = i) P(X_{t:T}, S_t = j \mid X_{1:t-1}, S_{t-1} = i) \\
&= P(X_{1:t-1}, S_{t-1} = i) P(X_{t:T}, S_t = j \mid S_{t-1} = i) \\
&= P(X_{1:t-1}, S_{t-1} = i) P(X_t, X_{t+1:T}, S_t = j \mid S_{t-1} = i) \\
&= P(X_{1:t-1}, S_{t-1} = i) P(X_{t+1:T} \mid X_t, S_t = j, S_{t-1} = i) \\
&\quad P(X_t, S_t = j \mid S_{t-1} = i) \\
&= P(X_{1:t-1}, S_{t-1} = i) P(X_{t+1:T} \mid S_t = j) \\
&\quad P(X_t, S_t = j \mid S_{t-1} = i) \\
&= P(X_{1:t-1}, S_{t-1} = i) P(X_{t+1:T} \mid S_t = j) \\
&\quad P(S_t = j \mid S_{t-1} = i) P(X_t = x_t \mid S_t = j) \\
&= \alpha_{t-1}(i) \beta_t(j) \gamma_{ij} p_j(x_t)
\end{aligned}$$

Proposition 2. *$\alpha_t(i)$ is computed by a recursion forward in time.*

$$\alpha_t(i) = \sum_{j=1}^{m} \alpha_{t-1}(j) \gamma_{ji} p_i(x_t)$$

$\beta_t(i)$ is computed by a recursion backward in time.

$$\beta_i(t) = \sum_{j=1}^{m} \beta_i(t+1) \gamma_{ij} p_j(x_{t+1})$$

proof 3. *The proof can be easily deduced from Equation (5.12) and (5.13).*

$$\begin{aligned}
P(X_{1:T}, S_t = j) &= \sum_{i=1}^{m} P(X_{1:T}, S_{t-1} = i, S_t = j) \\
\alpha_t(j) \beta_t(j) &= \sum_{i=1}^{m} \alpha_{t-1}(i) \beta_t(j) \gamma_{ij} p_j(x_t) \\
\alpha_t(j) &= \sum_{i=1}^{m} \alpha_{t-1}(i) \gamma_{ij} p_j(x_t)
\end{aligned}$$

The same holds for the second recursion.

$$\begin{aligned}
P(X_{1:T}, S_t = i) &= \sum_{j=1}^{m} P(X_{1:T}, S_t = i, S_{t+1} = j) \\
\alpha_t(i) \beta_t(i) &= \sum_{j=1}^{m} \alpha_t(i) \beta_{t+1}(j) \gamma_{ij} p_j(x_{t+1}) \\
\beta_t(i) &= \sum_{j=1}^{m} \beta_{t+1}(j) \gamma_{ij} p_j(x_{t+1})
\end{aligned}$$

Proposition 3. *The probability that the process visits the state i at time t given the observed sequence is*

$$\begin{aligned}\sigma_i(t) &= P(S_t = i \mid X = x) \\ &= \frac{P(X = x, S_t = i)}{P(X = x)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{P(X = x)} \end{aligned} \tag{5.14}$$

Proposition 4. *The probability that the process left state i at time $t-1$ and enters state j at t given the observed sequence is*

$$\begin{aligned}\phi_{ij}(t) &= P(S_{t-1} = i, S_t = j \mid X = x) \\ &= \frac{P(S_{t-1} = i, S_t = j, X = x)}{P(X = x)} \\ &= \frac{\alpha_{t-1}(i)\gamma_{ij}p_j(x_t)\beta_t(j)}{P(X = x)} \end{aligned} \tag{5.15}$$

5.3.3 The E-step

The CDLL is given by

$$\begin{aligned}\log L(\Theta \mid X, S) &= \log\left(\delta_{s_1}\prod_{t=2}^{T}\gamma_{s_{t-1},s_t}\prod_{t=1}^{T}p_{s_t}(x_t)\right) \\ &= \log(\delta_{s_1}) + \sum_{t=2}^{T}\log\gamma_{s_{t-1},s_t} + \sum_{t=1}^{T}\log p_{s_t}(x_t) \\ &= \sum_{i=1}^{m}\mathbf{1}_{\{s_1=i\}}\log\delta_i + \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=2}^{T}\mathbf{1}_{\{s_t=i,s_{t-1}=j\}}\right)\log\gamma_{ij} \\ &+ \sum_{i=1}^{m}\sum_{t=1}^{T}\mathbf{1}_{\{s_t=i\}}\log p_i(x_t) \end{aligned} \tag{5.16}$$

Hence

$$\begin{aligned}Q(\Theta \mid \Theta^{(k)}) &= \mathbb{E}_{S|X,\Theta^{(k)}}[\log L(\Theta \mid X, S)] \\ &= \sum_{i=1}^{m}\sigma_i(1)\log\delta_i + \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=2}^{T}\phi_{ij}(t)\right)\log\gamma_{ij} \\ &+ \sum_{i=1}^{m}\sum_{t=1}^{T}\sigma_i(t)\log p_i(x_t) \end{aligned} \tag{5.17}$$

where $\sigma_i(t)$ and $\phi_{ij}(t)$ are given in (5.14) and (5.15).

5.3.4 The M-step

In the M-step, we maximize the CDLL in (5.16) with respect to the parameter Θ. We maximize the first part with respect to the initial distribution δ_i. We use a Lagrange multiplier to maximize δ_i subject to the constraint $\sum_{i=1}^{m} \delta_i = 1$ as follows

$$\frac{\partial}{\partial \delta_i}\left[\sum_{i=1}^{m} \log(\delta_i)\sigma_i(1) + \lambda\left(\sum_{i=1}^{m} \delta_i - 1\right)\right] = 0$$

Then

$$\frac{1}{\delta_i}\sigma_i(1) + \lambda = 0 \tag{5.18}$$

or

$$\sigma_i(1) = -\lambda\delta_i$$

Summing over m yields

$$\sum_{i=1}^{m} \sigma_i(1) = -\lambda \sum_{i=1}^{m} \delta_i = -\lambda$$

As $\sum_{i=1}^{m} \sigma_i(1) = 1$, we get $\lambda = -1$.
The maximizing value of δ_i from Equation (5.18) is

$$\hat{\delta}_i = \sigma_i(1)$$

We maximize the second part of the CDLL in (5.16) with regard to γ_{ij} as follows

$$\frac{\partial}{\partial \gamma_{ij}}\left[\sum_{j=1}^{m}\left(\sum_{t=2}^{T} \phi_{ij}(t)\right)\log \gamma_{ij}\right] = 0$$

Equating the derivative to zero yields

$$\frac{\sum_{t=2}^{T} \phi_{ij}(t)}{\gamma_{ij}} - \frac{\sum_{t=2}^{T} \phi_{ii}(t)}{1 - \sum_{k \neq i} \gamma_{ik}} = 0$$

or

$$\frac{\sum_{t=2}^{T} \phi_{ij}(t)}{\gamma_{ij}} - \frac{\sum_{t=2}^{T} \phi_{ii}(t)}{\gamma_{ii}} = 0$$

This implies that

$$\gamma_{ij} \sum_{t=2}^{T} \phi_{ii}(t) = \gamma_{ii} \sum_{t=2}^{T} \phi_{ij}(t)$$

Summing over m gives

$$\gamma_{ii} \sum_{j=1}^{m} \sum_{t=2}^{T} \phi_{ij}(t) = \left(\sum_{j=1}^{m} \gamma_{ij} \right) \sum_{t=2}^{T} \phi_{ii}(t)$$

As $\sum_{j=1}^{m} \gamma_{ij} = 1$, then

$$\hat{\gamma}_{ii} = \frac{\sum_{t=2}^{T} \phi_{ii}(t)}{\sum_{t=2}^{T} \sum_{j=1}^{m} \phi_{ij}(t)}$$

Likewise, the maximizing value of γ_{ij} is then given by

$$\hat{\gamma}_{ij} = \frac{\sum_{t=2}^{T} \phi_{ij}(t)}{\sum_{t=2}^{T} \sum_{k=1}^{m} \phi_{ik}(t)}$$

The maximization of the third part of the CDLL in (5.16) depends on the nature of the state-dependent distribution (λ_i for the Poisson distribution, r_i and π_i for the NB distribution). For the Poisson distribution, analytic solutions are given. However, the maximization with respect to the parameters is not straightforward under the NB assumption. Hence, numerical maximization is needed. Maximizing values for λ_i and (r_i,π_i) are given in section 5.2.3. Note that for HMMs, $P(s_t = i \mid x_t, \Theta^{(k)})$ should be substituted from (5.14).

5.3.5 Example

We consider a NB-HMM model $\Theta = (\eta, \lambda, r)$ with m states. The transition probability matrix $\Gamma = (\gamma_{ij})$ is defined as follow

$$\gamma_{ij} = \begin{cases} 1 - \eta_i & \text{if } i = j \\ \frac{\eta_i}{m-1} & \text{if } i \neq j \end{cases}$$

With this parametrization of Γ, the second part of the CDLL in (5.16) becomes

$$\sum_{j=1}^{m} \left(\sum_{t=2}^{T} \phi_{ij}(t) \right) \log \gamma_{ij} = \log(1 - \eta_i) \left(\sum_{t=2}^{T} \phi_{ii}(t) \right) + \log \left(\frac{\eta_i}{1 - m} \right) \sum_{\substack{j=1 \\ j \neq i}}^{m} \left(\sum_{t=2}^{T} \phi_{ij}(t) \right)$$

Maximizing with regard to η_i and equating the derivative to zero yields

$$\eta_i \left(\sum_{t=2}^{T} \phi_{ii}(t) \right) = (1 - \eta_i) \sum_{\substack{j=1 \\ j \neq i}}^{m} \left(\sum_{t=2}^{T} \phi_{ij}(t) \right)$$

It follows immediately that

$$\hat{\eta}_i = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{t=2}^{T} \phi_{ij}(t)}{\sum_{j=1}^{m} \sum_{t=2}^{T} \phi_{ij}(t)}$$

We finally substitute $\hat{\eta}_i$ in (5.15) to derive $\phi_{ij}(t)$.

We simulated $1,000$ observations from Θ model given $\eta = (0.1, 0.2, 0.5)$, $\lambda = (5, 10, 20)$ and $r = (10, 10, 10)$. We apply the EM algorithm to fit the simulated data. We run the algorithm several times using different starting values to avoid convergence to a local minimum. The stopping rule is $\sum_i \mid \theta^{(k)} - \theta^{(k-1)} \mid < 0.001$.

After 34 iterations, the EM algorithm provides fair estimate of Θ with $\hat{\eta} = (0.0902, 0.1568, 0.5109)$, $\hat{\lambda} = (4.9011, 10.9209, 22.4128)$ and $\hat{r} = (14.2621, 5.4561, 16.0113)$. The estimated transition probability matrix is

$$\hat{\Gamma} = \begin{pmatrix} 0.9098 & 0.0451 & 0.0451 \\ 0.0784 & 0.8432 & 0.0784 \\ 0.2555 & 0.2555 & 0.4891 \end{pmatrix}$$

The model log-likelihood is -2947.769.

5.4 Discussion

In the context of independent mixtures and HMMs, a task of major importance is the choice of the optimal state-dependent distribution and number of states m of the latent process, since the choice of the optimal model leads to the improvement of the goodness-of-fit. The model fit can be increased with increasing m due to the model likelihood. However, increasing m implies an increase in the number of parameters. Without making assumptions on the transition probability matrix, the problem is quadratic, since the number of parameters is $m^2 + 2m - 1$ in the case of Poisson-HMMs and m^2+3m-1 in the case of NB-HMMs. In Example 5.3.5, we made specific assumptions on Γ to reduce the complexity of the model. Under such assumptions, the problem is linear, since the number of parameters is $3m-1$ in the case of Poisson-HMM and $4m - 1$ in the case of NB-HMM. Hence, a compromise has to be found between the model fit and the model complexity. Model selection criteria are used to balance the two situations. They are either based on the full-model log-likelihood (AIC and BIC) [Rydén 1995, MacDonald 1997, Gassiat 2003, Dannemann 2008], or on reducing the number of parameters by making assumptions on the state-dependent distribution or on the transition probability matrix in the case of HMMs [Zucchini 2000, Poskitt 2005]. Hypothesis tests, as LRT, can also be used in this

context. They have the advantage to allow decisions with a significance level. To the best of our knowledge, there is no common acceptance of the best criteria for determining the number of states. This issue can best be summarized by a quote from famous Bayesian statistician George Box, who said: *"All Models are wrong, but some are useful"* [Box 1976a].

Chapter 6

Overdispersion in the Distribution of Malaria Parasites and Leukocytes in Thick Blood Smears

"Beware of the problem of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confessions obtained under duress may not be admissible in the court of scientific opinion."

Stephen M. Stigler, 1987

Contents

The aim of this chapter is to explore overdispersion in parasite and leukocyte counts collected from the field. In the context of overdispersion, emphasis is laid on fitting data to appropriate models. The fitting models are fully presented in Chapter 5.

6.1 Introduction

Most of PD estimation methods assume that the distribution of the thickness of the TBS, and hence the distribution of parasites and leukocytes within

the TBS, is homogeneous; and that parasites and leukocytes are evenly distributed in TBSs, and thus can be modelled through a Poisson-distribution [Student 1907, Petersen 1996a, Bejon 2006, Hammami 2013]. PD data-based inferences also rely on such assumptions [Becher 2005, Damien 2010, Chandler 2006, Färnert 2009a, Mwangi 2005, Liljander 2011, Enosse 2006].

Identifying the distribution of parasite and leukocyte data on TBSs is the key to an appropriate analysis. Raghavan [Raghavan 1966] recognized that parasites may be missed due to the random variation within a slide. He used the binomial distribution to estimate the probability of missing a positive slide, when only a fixed number of HPFs is read. He assumed that parasites were randomly distributed in the blood film, and that each parasite has the same chance of occupying any of the HPFs read. Dowling & Shute [Dowling 1966] showed that leukocytes are evenly distributed in thick films, and that their number varies directly according to the thickness of the smear. They indicated a normal distribution of leukocytes per HPFs. In addition, they claim that parasites are also distributed evenly throughout the thick blood smear. However, they noticed, in the case of scanty parasitaemia, a phenomenon of "grouping", in which parasites tend to aggregate together in a specific area of the smear. Petersen *et al.* [Petersen 1996a] claimed that estimating the PD from the proportion of parasite-positive HPFs, instead of counting parasites in each field, underestimates the PD in TBSs, since a parasite-positive field may contain more than one parasite. To get ride of this problem, they suggested a correction of the estimation method. Their model was built under the assumption that parasites are Poisson-distributed on the TBSs. Under this assumption, the estimate of the mean number of parasite per field (λ) is then $\widehat{\lambda} = -\log(1-p)$, where p is the percentage of parasite-positive HPFs. However, due to the clustering of parasites in TBSs, $\widehat{\lambda}$ was corrected by a factor of 2. This factor of two was empirically chosen without a clear analytical proof. Bejon *et al.* [Bejon 2006] used the Poisson distribution to calculate the likelihood of sampling a parasite within the blood volume examined in microscopy. Alexander *et al.* [Alexander 2010] described the variation across the sample by a homogeneous Poisson distribution of parasites on TBSs. They unpacked -under the Poisson assumption- similar results to Raghavan's -under the Binomial assumption- at low densities, but he argued for the evidence of discrepancy as density increases.

Two assumptions specific to the Poisson model have been identified as sources of misspecification. The first is the assumption that variance equals the mean. The second is the assumption that events occur evenly. That assumption preludes, for instance, that occurrences in a field influence the probability of occurrences in neighbouring fields. But this type of contagion is to be suspected in the distribution of parasites and leukocytes in TBS. Violations of both assumptions lead to the same symptom: a violation of the Poisson variance assumption. Overdispersion, or extra-Poisson variation, denotes a situation in which the variance exceeds the mean. Unobserved heterogeneity and positive contagion lead to overdispersion [Selby 1965, Darwin 1957, Cox 1983, McCullagh 1989]. Undetected heterogeneity may entail important misleading inferences, so its detection is essential.

Three lines of research exist to account for overdispersion. Firstly, an overdispersion test is helpful, since the lack of significance in testing overdispersion might indicate that a further investigation of latent heterogeneity might not be necessary. Various tests for detecting overdispersion have been developed [Dean 1989, Gurmu 1991, Dean 1992, Lee 1986, Lu 1997]. Secondly, the effect of overdispersion has been analysed and corrected within the maintained Poisson model [Gourieroux 1984, Petersen 1996a]. Thirdly, various models have been proposed that account for unobserved heterogeneity while nesting the Poisson model as a special case [Cameron 1986, Gschlößl 2008, J.F. 1987, Winkelmann 1991, Mullahy 1986, Joe 2005, Winkelmann 2003, Yau 2003]. Standard approaches employ mixture distributions, either parametrically by introducing models that accommodate overdispersion, for example the negative binomial models, or semiparametrically by leaving the mixing distribution unspecified [Gurmu 1998, Petersen 1996a]. These parametric and semiparametric models involve an extra-dispersion parameter, which requires numerical methods for its estimation [Clark 1989, Piegorsch 1990, Boes 2007].

In published studies, malariological data are presented as summary statistics (e.g. parasite density per microlitre, prevalence, absolute or assumed WBC count). Parasite and leukocyte counts per field, while of great importance, are not available in the open literature or in archived sources. A dataset of parasite and leukocyte counts per HPF was then constituted and published in this study. Three TBSs of 12-month-old children were entirely examined. All HPFs were read sequentially. The number of parasites and the number of leukocytes per HPF were recorded. The aim of this study is twofold: to examine the presence of overdispersion in the distribution of parasites and leukocytes in TBSs, and to fit the appropriate model that allows for overdispersion in these data. To do so, two sources of overdispersion are explored: the latent heterogeneity in parasite and leukocyte counts, i.e. the presence of homogeneous zones (where the data have a similar distribution) associated to an unobserved state, and the spatial dependence in data, i.e: the correlation between neighbouring occurrences.

The aim of this chapter is twofold: to examine the presence of overdispersion in the distribution of parasites and leukocytes in TBSs, and to fit the appropriate model that allows for overdispersion in the data.

6.2 Materials and methods

6.2.1 Epidemiological data

The data accompanying this study were gathered from a field study of *Plasmodium falciparum* malaria in the district of Tori Bossito located 40 km North-East of Cotonou, South Benin. Across this field study, 550 infants were followed weekly from birth to 12 months [Le Port 2011a, Le Port 2012]. Malaria is perennial in the study area, and according to a recent entomological survey *P. falciparum* is the commonest species (95%), *Plasmodium malariae* and *Plasmodium ovale* representing respectively 3% and 2% [Djenontin 2010]. From the *Tori-Bossito* study, three

thick films of 12-month-old children were randomly selected among positive slides and included in this study. TBSs were stained with Giemsa. All high power fields (HPFs), defined as oil immersion microscopic fields ($\times 1,000$), were re-examined by visually scanning the entire film horizontally from edge to edge. The number of parasites (p) per field and the number of leukocytes (ℓ) per field were derived. The letters "a", "b" "c" denote the three selected TBSs throughout this paper. A summary of the data is given in Table 6.1. Histograms of the data are plotted in Figure 6.1 in order to help for visualizing the shape of the data before the distributions are fitted.

TBS	***a***		***b***		***c***	
Number of HPFs	754		938		836	
Volume of blood* (μl)	1.51		1.88		1.67	
PD† (parasites /μl)	16,190.79		31,783.18		3,725.95	
Parasites and leukocytes	p_a	ℓ_a	p_b	ℓ_b	p_c	ℓ_c
Total number	20621	10189	38112	9593	5989	12859
Mean (per HPF)	27.35	13.51	40.63	10.23	7.16	15.38
Median	25	13	37	10	7	14
Range	0-111	0-43	0-131	0-35	0-22	2-47
IQR‡	12-40	8-17	20-60	6-14	4-10	11-19
Standard deviation	18.76	7.22	25.94	5.90	3.92	6.62
% negative§	1.06	1.06	0.75	1.39	1.08	0.00

Table 6.1: ***Descriptive statistics of parasite and leukocyte counts on TBSs.***
Three thick blood smears are studied "a", "b", "c".
Parasite and leukocyte counts for each TBS are denoted (p_a, ℓ_a), (p_b, ℓ_b) et (p_c, ℓ_c).
* Assuming that the volume of blood in one HPF is approximately 0.002 μl [Dowling 1966, Bruce-Chwatt 1985, Warrell 2002]
† $PD = \frac{p}{\ell} \times 8,000$, assuming that the number of leukocytes per microlitre of blood is 8,000 [Bruce-Chwatt 1958, Greenwood 1987, WHO 2010a].
‡ Inter-Quartile Range.
§ Percentage of negative high-power fields (HPFs) where no parasites and/or no leukocytes are seen.

6.2.2 Statistical models for parasite and leukocyte data

Some laboratory counting techniques consist in reading a certain volume of blood (say u μl) before the film is declared negative. If parasites are seen in u μl, then an additional volume (say v μl) is read. The volume of blood contained in one HPF is approximately 0.002 μl [Dowling 1966, Bruce-Chwatt 1985, Warrell 2002]. The assumed number of white blood cells per microlitre of blood is 8,000 [Greenwood 1987, WHO 2010a]. In practice, u μl may correspond to 100 HPFs (i.e. $u = 0.2$ μl), and v μl may correspond to 200 white blood cells (i.e. $v = 0.025$ μl) [Reyburn 2007, WHO 2010a, Allen 2011, Adu-Gyasi 2012]. In this example, parasites are assumed to be spread evenly throughout the TBS with density θ μl. Under the Poisson assumption, the probability of seeing no parasites in u volume

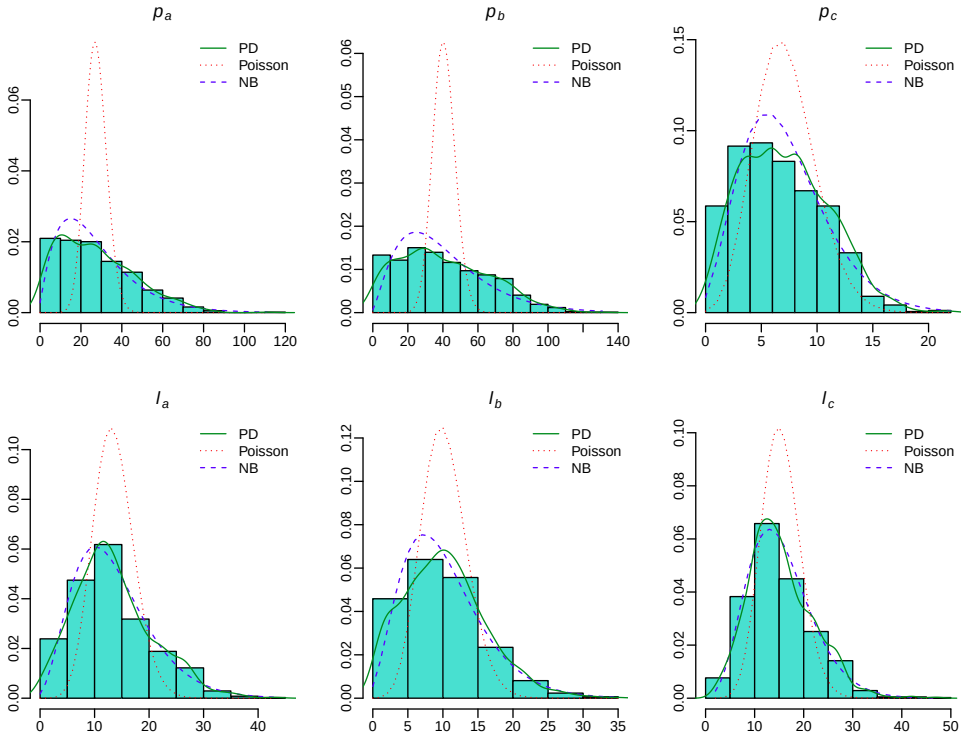

Figure 6.1: ***Histograms of parasite and leukocyte counts per HPF.*** The empirical density function and the fitted distributions (Poisson, NB) are displayed on the top of each histogram.

of blood is $e^{-\theta u}$, and the probability of seeing exactly x parasites ($x > 0$) is then $(1-e^{\theta u})e^{-\theta v}(\theta v)^{x-1}/(x-1)!$. The latter probability is the product of the probability of seeing at least one parasite in volume u, and the probability of seeing $(x-1)$ more parasites in volume v. Under this procedure, the estimation of the PD depends on volumes u and v, which are not the same for all slides.

The restrictive nature of the equidispersion assumption in the Poisson model led to the development of numerous techniques both for detecting and modelling overdispersion [Zorn 1996, Gurmu 1991, Lee 1986, Cameron 1986, Dean 1989, King 1989, Hausman 1984]. This section details alternative models used to fit the PD and leukocyte data.

6.2.2.1 Simple parametric models

The typical alternative to the Poisson model is the negative binomial (NB) model, which is an attractive model that allows overdispersion. The dispersion parameter ϕ in the NB controls the deviation from the Poisson. This makes the NB distribution suitable as a robust alternative to the Poisson. However, it is useful to obtain more general specifications through other modelling frameworks that handle overdispersion or zero-inflation (NB, geometric, logistic, Gaussian, exponential, zero-

inflated Poisson (ZIP), Poisson hurdle (HP), zero-inflated negative binomial (ZINB), negative binomial hurdle (HNB)). The main motivation behind using zero-inflated [Lambert 1992, W. 1994] and hurdle count models [Mullahy 1986, Heilbron 1994] is that PD data frequently display excess zeros at low parasitaemia levels. Zero-inflated and hurdle count models provide a way of modelling the excess zeros in addition to allowing for overdispersion. These models include two possible data generation processes (one generates only zero counts, whereas the other process generates counts from either a Poisson or a negative binomial model).

6.2.2.2 Finite mixture models

One method of dealing with overdispersed observations with a bimodal or more generally multimodal distribution is to use a finite mixture model. Mixture models are designed to account for unobserved heterogeneity in a set of data. The sample may consist of unobserved groups, each having a distinct distribution for the observed variable. Consider for example the distribution of parasites per HPF, X_t. The fields can be divided into groups according to its locations, e.g. edges and center of the film. Even if the number of parasites within each group was Poisson-distributed, the distribution of X_t would be overdispersed relative to the Poisson. In the case of a two-component mixture with weights (δ_1, δ_2), means (λ_1, λ_2) and variances (σ_1^2, σ_2^2), the total variance exceeds the mean by $\delta_1\delta_2(\lambda_1 - \lambda_2)^2$ (details of the proof are given in Additional file 2). Hence, the two-state Poisson mixture is able to accommodate overdispersion better than the Poisson model with one component. The mixture component identities are defined by some latent variables (also called the *parameter process*). If the latent variables are independent, the resulting distribution is called *independent* mixture. An independent mixture distribution consists of a finite number, say m, of component distributions and a mixing distribution which selects from these components. Note, however, that the above definition of mixture models ignores the possibility of spatial dependence in data, a point that shall be addressed by introducing Hidden Markov Models (HMMs), which connect the latent variables into a Markov chain instead of assuming that they are independent.

6.2.2.3 Hidden Markov models (HMMs)

Unlike the mixture models, where observations are assumed independent of each other and the spatial relationship between neighbouring data is not taken into account, HMMs incorporate this spatial relationship, and show promise as flexible general purpose models to account for such dependency [Baum 1966, Baum 1967, Rabiner 1989]. HMMs can be used to describe observable events that depend on underlying factors, which are not directly observable, namely the *hidden states*. A HMM consists of two stochastic processes: an invisible process of hidden states, namely the *hidden process* (also called the *parameter process*), and a visible process of observable events, namely the *observed process* (or the *state-dependent process*). The hidden states follow a Markov chain, in which, given the present state, the

future is independent of the past. Modelling observations in these two layers, one visible and the other invisible, is very useful to classify observations into a number of classes, or clusters, and to incorporate the spatial-dependent information among neighbouring observations. In the context of parasite and leukocyte counts per HPF, emphasis is put on predicting the sequence of regions on the TBS (i.e. the states) that gave rise to the actual parasite and leukocyte counts (i.e. the observations). Since a variation in the distribution of parasites and leukocytes in the TBS is suspected, these regions cannot be directly observed, and need to be predicted. Inference in HMMs is often carried out using the expectation-maximization (EM) algorithm [Baum 1970, Dempster 1977, Cappé 2005], but examples of Bayesian estimation implemented through Markov chain Monte Carlo (MCMC) sampling are also frequent in the literature [Robert 2000, Rydén 2008]. In most practical cases, the number of hidden states is unknown and has to be estimated. We shall return to the latter point later in the discussion.

6.2.3 Methodology

Firstly, the problem of testing whether the data come from a single Poisson distribution is considered. The basic null hypothesis of interest is that "variance = mean" (equidispersion). In a context such as this, the focus is put on alternatives that are overdispersed, in the sense that "variance > mean". The hypothesis being tested is commonly referred to as the homogeneity hypothesis. A commonly used statistic for testing the Poisson assumption is Pearson's test, which in spatial statistics is known as the index of dispersion test [Fisher 1950, Rao 1956]. The statistic is the ratio of the sample variance to the sample mean, multiplied by $(n-1)$, where n is the sample size.

In the case of the Poisson distribution, the variance is equal to the mean, i.e. the index of dispersion is equal to one. In the case of the binomial distribution, the index of dispersion is less than 1; this situation is called *underdispersion*. For all mixed Poisson distributions, that show overdispersion in data, the index of dispersion is greater than 1. Fisher [Fisher 1950] showed that under the assumption that data are generated by a Poisson distribution with some parameter λ, then the test statistic approximately has a Chi-squared distribution (χ_2) with $(n-1)$ degrees of freedom.

If the Poisson assumption is violated, the goodness of fit of alternative simple parametric models should be assessed. In order to estimate model parameters, a direct optimization of the log-likelihood is performed using `optim` [Nelder 1965]. The Kolmogorov-Smirnov (k.s) goodness-of-fit test is used [Chakravarti 1967] to test the validity of the assumed distribution for the data. The test evaluates the null hypotheses (that the data are governed by the assumed distribution) against the alternative (that the data are not drawn from the assumed distribution). Model selection criteria are used to determine which of the simple parametric models best fits the data. The selection criteria used in this paper are presented in the next section.

Secondly, the first source of overdispersion in count data is investigated, which

is unobserved heterogeneity. The unobserved heterogeneity among parasite and leukocyte data is explored using mixture models. The motivation behind the use of mixture models is that they can handle situations where a single parametric family is unable to provide a satisfactory model for local variations in data. The objective here is to describe the data as a finite collection of homogeneous populations on TBSs. The form of these sub-populations is modelled using Poisson and NB distributions.

Thirdly, the second source of overdispersion is explored, which is positive contagion [King 1989]. When contagion is present, the value of X_t positively influences the value of $X_{t'}$ $(t \neq t')$. For example, a high number of parasites in one HPF leads to correspondingly high numbers of parasites in neighbouring HPFs; likewise, a low number of parasites in one HPF drive down counts for other neighbouring HPFs. Since this data-generating process directly influences the occurrence of parasites in HPFs, it has important implications for the observed level of dispersion in data.

The autocorrelation plots [Box 1976b] are a commonly-used tool for checking randomness and spatial dependence in data. The autocorrelation function (ACF) will first test whether adjacent observations are autocorrelated; that is, whether there is correlation between observations x_1 and x_2, x_2 and x_3, x_3 and x_4, etc. This is known as lag one autocorrelation, since one of the pair of tested observations lags the other by one period (ie. one HPF). Similarly, it will test at other lags. For instance, the autocorrelation at lag five tests whether observations x_1 and x_6, x_2 and x_7, ..., x_{27} and x_{32}, etc, are correlated. If random, such autocorrelations should be "near zero" for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero. HMMs are used to account for autocorrelations in data. The state-dependent distribution is modelled using Poisson and NB. Note that HMMs are an extension of mixture models with spatial dependence taken into consideration, and the two types of models are nested.

The proposed mixture models and HMMs are fitted by maximum likelihood using the EM algorithm, and validated by direct numerical maximization using `nlm` in `R` [Dennis 1983, Schnabel 1985]. Initialization of the EM algorithm is based on incremental k-means [Hartigan 1979]. Details on the maximization of the complete-data log-likelihood with regard to parameters of the unobserved state distribution (Poisson, NB) for mixture models and HMMs are given in chapter 5.

6.2.4 Model selection and checking

Models comparison was based on three measures. One is the deviance statistic, also called the likelihood-ratio test statistic or likelihood-ratio chi-squared test statistic, which is a measure of the difference in log-likelihood between two models. If data have been generated by Model A (a simpler model) and are analysed with Model B (a more complex model within which model A is nested), the expected distribution of the test statistic, which is twice the difference in log-likelihoods $2(\mathcal{L}_B - \mathcal{L}_A)$ computed using the data, follows a χ_2-distribution with degrees of freedom equal to the difference in the number of parameters. Hence, LRT permits a probabilistic decision as to whether one model is adequate or whether an alter-

native model is superior. This statistic is appropriate when one model is nested within another model. Negative binomial and Poisson models are nested because as ϕ converges to 0, the negative binomial distribution converges to Poisson. But the situation is non-standard, because under the null hypothesis the extra parameter ϕ lies on the boundary of its parameter space. The standard asymptotic result of a χ^2-distribution is not applicable. For this purpose, Akaike's Information Criterion (AIC) [Akaike 1973] and the Bayesian Information Criterion (BIC) [Schwarz 1978] are used. These two measures penalize for model complexity and permit comparison of nonnested models. Models are nonnested if there is no parametric restriction on one model that produces the second model specification. The AIC (resp. BIC) can be thought of as the amount of information lost when a specific model to approximate the real distribution of data is being used. Thus, the model with the smallest AIC (resp. BIC) is favored.

In the area of statistical modelling (e.g: regression, generalised linear models), residuals are broadly used to check the validity of the fitted model. In this context, residuals are calculated from the model predictions and the observed data. In the context of HMMs, no strict analog to a residual exists since the value of a residual depends on the unobservable state. Pseudo-residuals offer a convenient way for model checking in HMMs [MacDonald 1997, Patterson 2009]. The HMM version of residuals is used to check the validity of the model as well as to identify outliers, since their absolute value indicate the deviation from the median of the distribution. While information criteria for model selection compare the relative goodness-of-fit, the analysis of pseudo-residuals provides a measure of the absolute goodness-of-fit. Zucchini and MacDonald [MacDonald 1997] provide details for calculating and assessing two types of pseudo-residuals (ordinary and forecast), for both continuous and discrete state distributions. Model pseudo-residuals can also be extracted using the function "`Residuals`" in the `R` package `HiddenMarkov`. Here, the ordinary pseudo-residuals are used to evaluate the suitability of selected HMMs. The ordinary pseudo-residual for the observation x_t is based on its conditional distribution given all other data. In the case of discrete observations, pseudo-residuals are defined as intervals $[r_t^-, r_t^+]$ as

$$r_t^- = \Phi^{-1}\left(P(X_t < x_t \mid x_{t-1}, x_{t-2}, ..., x_1)\right) \quad \forall t \in [\![1\,;T]\!]$$
$$r_t^+ = \Phi^{-1}\left(P(X_t \leq x_t \mid x_{t-1}, x_{t-2}, ..., x_1)\right) \quad \forall t \in [\![1\,;T]\!]$$

where Φ is the c.d.f. of a standard normal-distributed random variable. If the fitted model is correct, the pseudo-residuals are standard normal-distributed. Graphically, QQ-plots and pseudo-residual ACFs were used to assess the goodness-of-fit of selected HMMs.

6.3 Results

6.3.1 Overdispersion in parasite and leukocyte distributions

Histograms in Figure 6.1 show that parasite and leukocyte counts are clearly skewed to the right. The fitted "candidate" distributions, Poisson and NB, are displayed on the top of each histogram and compared to the empirical density function in order to visualize how well they match the data. The Poisson distribution clearly does not fit the data. On the other hand, the NB distribution fits the data much more closely than the Poisson distribution. This result was expected because of the implicit restriction of the Poisson model on the distribution of the observed counts. It is true that the negative binomial distribution converges to the Poisson distribution, but the former will be always more skewed to the right than the latter with similar parameters.

The initial visualization of the histograms motivates the use of Pearson's test to check for overdispersion. In all TBSs, the Poisson model was highly significantly rejected in favor of a model with heterogeneity ($p \ll .0001$ using Pearson's test). The authors considered fitting data to alternative models allowing for overdispersion: NB, geometric, logistic, Gaussian, exponential. The k.s test was significant ($p \ll .0001$), then it indicated that the distribution of the parasite and leukocyte data was significantly different from the distribution against which it was being compared. However, this test is frequently found to be too sensitive. Given a large enough sample size, it can detect differences that are meaningless to the present purpose, in the sense that even very small divergences of the model from the data would be flagged up and cause significance of the test. It is certainly worth judging the results of the test in light of other statistical measures. The AIC is used to assess the goodness-of-fit of alternative models to data. The difference in fit between the Poisson model (resp. NB model) and its corresponding ZIP and HP models (resp. ZINB and HNB models) is trivial. This result might be expected due to the non-excess of zeros in data (see Table 6.1). The AIC selects the NB model, which is estimated to be "closest" to the unknown distribution that generated the data ($\triangle\text{AIC} \gg 10$) (see Table 6.2).

The maximum likelihood estimators (MLE) for the dispersion parameter of the negative binomial models (ϕ) are: $\hat{\phi}_{\text{MLE}}(p_a) = 0.53$, $\hat{\phi}_{\text{MLE}}(p_b) = 0.53$, $\hat{\phi}_{\text{MLE}}(p_c) = 0.18$, $\hat{\phi}_{\text{MLE}}(\ell_a) = 0.23$, $\hat{\phi}_{\text{MLE}}(\ell_b) = 0.28$, $\hat{\phi}_{\text{MLE}}(\ell_c) = 0.12$ (the maximum likelihood equations are solved iteratively). The positivity of the dispersion parameter of the negative binomial models indicates that parasites (resp. leukocytes) tend to be aggregated together, leaving some areas with high parasite (resp. leukocyte) densities, and other areas with very few parasites (resp. leukocytes) [Bliss 1953]. These findings indicate that there is significant overdispersion in the distribution of parasites and leukocytes across all TBSs used in the analysis.

	Poisson			Negative Binomial		
	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	6801.59	13605.17	13609.80	3200.63	**6405.25**	**6414.50**
p_b	10838.95	21679.91	21684.75	4344.27	**8692.54**	**8702.23**
p_c	2472.18	4946.36	4951.08	2302.96	**4609.92**	**4619.38**
ℓ_a	3108.25	6218.51	6223.13	2532.77	**5069.53**	**5078.79**
ℓ_b	3547.53	7097.06	7101.90	2965.34	**5934.69**	**5944.38**
ℓ_c	3051.08	6104.15	6108.88	2728.46	**5460.91**	**5470.37**
	Geometric			**Logistic**		
	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3249.22	6500.44	6505.06	3287.80	6579.60	6588.86
p_b	4413.13	8828.26	8833.10	4407.19	8818.38	8828.06
p_c	2488.96	4979.93	4984.65	2344.83	4693.66	4703.12
ℓ_a	2719.04	5440.09	5444.72	2560.46	5124.92	5134.17
ℓ_b	3122.84	6247.69	6252.53	2998.50	6001.01	6010.69
ℓ_c	3122.55	6247.11	6251.84	2762.37	5528.74	5538.20
	Gaussian			**Exponential**		
	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3279.99	6563.99	6573.24	3248.74	6499.48	6504.10
p_b	4384.43	8772.85	8782.54	4412.85	8827.71	8832.55
p_c	2327.71	4659.41	4668.87	2482.13	4966.25	4970.98
ℓ_a	2560.19	5124.39	5133.64	2717.17	5436.34	5440.96
ℓ_b	2995.11	5994.21	6003.90	3118.89	6239.77	6244.62
ℓ_c	2765.26	5534.51	5543.97	3120.93	6243.86	6248.59

Table 6.2: ***Comparison of simple parametric models fitted to parasite and leukocyte counts per field.***
Parasite (p_a, p_b, p_c) and leukocyte (ℓ_a, ℓ_b, ℓ_c) counts are fitted to Poisson, Negative Binomial, Geometric, Logistic, Gaussian and Exponential models. Minus log-likelihood ($-\mathcal{L}$) and information measures (AIC and BIC) are given. Direct optimization of the log-likelihood was performed using `optim` in `R`. The best AIC and BIC values are highlighted in bold.

6.3.2 Modeling heterogeneity in parasite and leukocyte data

Mixture models fitted to parasite and leukocyte counts are presented in Table 6.3. Using a two-state Poisson mixture instead of a one-state Poisson model dramatically improved the fit to data as judged by the AIC and BIC contrary to NB case. The simple parametric NB model was preferred to NB mixtures. The goodness-of-fit of Poisson mixtures increased with m values. Poisson mixtures (slightly) outperformed the one-state NB model according to AIC for TBSs "a" and "b". However, the one-state NB model was preferred to the Poisson mixtures according to BIC for all TBSs.

Spatial dependence between data is explored through autocorrelation plots (see Figure 6.2). Autocorrelations should be near-zero for randomness, which was not

	Poisson mixture			Negative binomial mixture		
$m = 1$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	6801.59	13605.17	13609.80	3200.63	6405.25	6414.50
p_b	10838.95	21679.91	21684.75	4344.27	8692.54	8702.23
p_c	2472.18	4946.36	4951.08	2302.96	4609.92	4619.38
ℓ_a	3108.25	6218.51	6223.13	2532.77	5069.53	5078.79
ℓ_b	3547.53	7097.06	7101.90	2965.34	5934.69	5944.38
ℓ_c	3051.08	6104.15	6108.88	2728.46	5460.91	5470.37
$m = 2$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3962.18	7930.35	7944.23	3200.63	6409.25	6430.53
p_b	5882.41	11770.81	11785.34	4344.27	8696.54	8718.69
p_c	2289.73	4585.47	4599.65	2302.96	4613.93	4635.61
ℓ_a	2633.87	5273.75	5287.62	2532.77	5073.54	5094.81
ℓ_b	3029.67	6065.33	6079.86	2965.35	5938.69	5960.84
ℓ_c	2756.98	5519.97	5534.15	2728.45	5464.91	5486.59
$m = 3$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3397.75	6805.50	6828.63	3200.63	6413.25	6447.60
p_b	4761.19	9532.38	9556.60	4344.27	8700.54	8736.20
p_c	2288.39	4586.77	4610.41	2302.96	4617.93	4652.89
ℓ_a	2527.85	5065.70	5088.83	2532.77	5077.54	5111.88
ℓ_b	2945.87	5901.74	5925.95	2965.35	5942.69	5978.35
ℓ_c	2729.21	5468.42	5492.06	2728.45	5468.90	5503.87
$m = 4$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3267.46	6548.92	6581.29	3189.16	6394.32	6442.42
p_b	4470.16	8954.33	8988.24	4344.27	8704.54	8754.38
p_c	2288.21	4590.42	4623.52	2302.96	4621.93	4670.85
ℓ_a	2519.22	5052.44	5084.81	2532.77	5081.54	5129.63
ℓ_b	2938.52	5891.05	5924.95	2965.35	5946.69	5996.53
ℓ_c	2721.23	5456.47	5489.57	2728.45	5472.90	5521.82

Table 6.3: ***Comparison of independent mixture models fitted to parasite and leukocyte counts by AIC and BIC.***
Parasite (p_a, p_b, p_c) and leukocyte (ℓ_a, ℓ_b, ℓ_c) counts are fitted to Poisson mixtures and negative binomial mixtures. The number of components is m. Minus log-likelihood ($-\mathcal{L}$) and information measures (AIC and BIC) are given. Models were fitted by maximum likelihood using the expectation-maximization (EM) algorithm, and validated by direct numerical maximization using `nlm` in `R`.

the case for parasite and leukocyte data. Thus, the randomness assumption failed as expected. The confidence limits are provided to show when ACF appears to be significantly different from zero. Lags having values outside these limits (shown as blue dotted bars) should be considered to have significant correlations. For "p_a", "p_b" and "ℓ_a", the autocorrelation plots start with a moderate autocorrelation at lag 1 (between 0.5 and 0.6) that gradually decreases. The decreasing autocorrelation

is generally linear, but with significant noise. Such a pattern is the autocorrelation plot signature of a "moderate autocorrelation", which in turn provides moderate predictability if modelled properly. For parasite data "p_c", a very few lags > 4 slightly lie outside the 95% confidence limits. For leukocyte data "ℓ_b" and "ℓ_c", with the exception of lags < 5, almost all of the autocorrelations fall within the 95% confidence limits. For all TBSs, the ACF suggests the existence of a spatial dependence between data. HMMs are therefore used to account for this dependence.

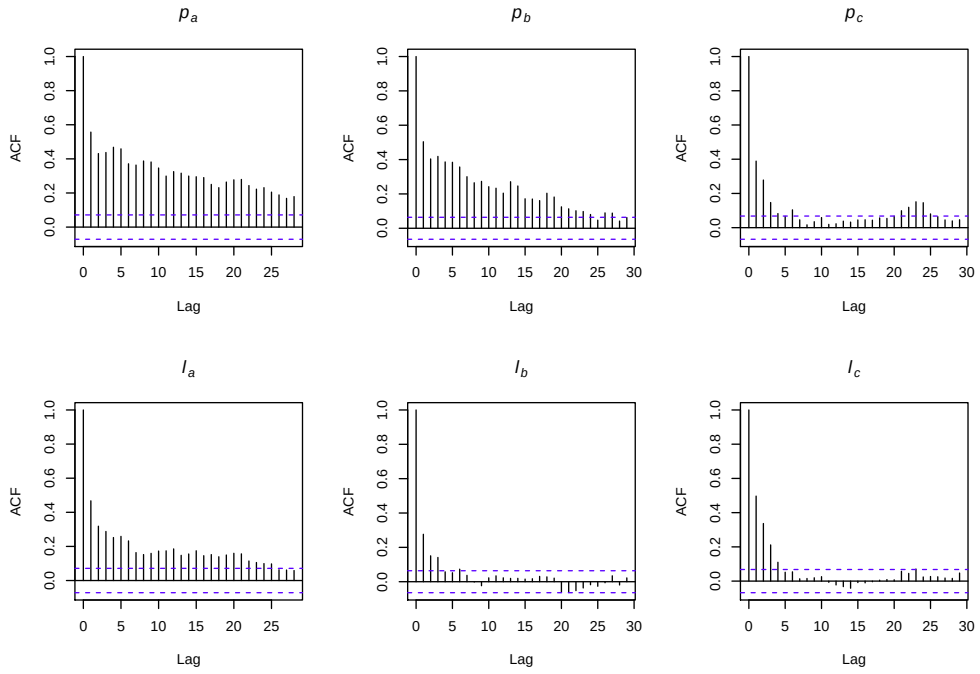

Figure 6.2: ***Sample autocorrelation function (ACF).***
Autocorrelation plots for parasite (p_a, p_b, p_c) and leukocyte (ℓ_a, ℓ_b, ℓ_c) counts show correlations between values x_i and lagged values of the counts for lags from 0 to 30. The lagged values can be written as x_{i-1}, x_{i-2}, x_{i-3}, and so on. ACF gives correlations between x_i and x_{i-1}, x_i and x_{i-2}, and so on. The lag is shown along the x-axis, and the autocorrelation is on the y-axis. The blue dotted lines indicate bounds for statistical significance.

The comparison of independent mixture models in Table 6.3 and HMMs in Table 6.4 shows that, on the basis of AIC and BIC, HMMs are superior to mixture models. Although more parameters need to be evaluated for HMMs than for comparable independent mixtures, the corresponding AIC and BIC were lower than those obtained for the independent mixtures. Given the spatial depedence shown in Figure 6.2, one would expect that independent mixture models will not perform well relative to HMMs.

Due to its higher complexity, an m-state model will always have a higher likelihood than an (m-1)-state model. Model selection criteria are used to see if the improvement in the likelihood was great enough to indicate that the m-state model

	Poisson HMM			Negative binomial HMM		
$m = 1$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	6801.59	13605.17	13609.80	3200.63	6405.25	6414.50
p_b	10838.95	21679.91	21684.75	4344.27	8692.54	8702.23
p_c	2472.18	4946.36	4951.08	2302.96	4609.92	4619.38
ℓ_a	3108.25	6218.51	6223.13	2532.77	5069.53	5078.79
ℓ_b	3547.53	7097.06	7101.90	2965.34	5934.69	5944.38
ℓ_c	3051.08	6104.15	6108.88	2728.46	5460.91	5470.37
$m = 2$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3877.14	7764.27	7787.40	3043.31	6098.62	6126.37
p_b	5794.89	11599.77	11623.99	4166.23	8344.45	8373.51
p_c	2228.73	4467.47	4491.11	2224.71	4461.42	4489.79
ℓ_a	2578.83	5167.66	5190.79	2433.86	4879.72	4907.47
ℓ_b	2993.67	5997.35	6021.57	2889.88	5791.76	5820.82
ℓ_c	2667.70	5345.41	5369.05	2640.61	5293.22	5321.59
$m = 3$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	6447.60	3265.54	6553.09	6603.97	3008.87	6035.74
p_b	4634.75	9291.50	9344.78	4126.32	8270.64	8314.23
p_c	2210.74	4443.48	4495.49	2215.95	4449.90	4492.46
ℓ_a	2414.70	4851.41	4902.28	2394.82	4807.64	4849.27
ℓ_b	2898.08	5818.17	5871.45	2884.03	5786.06	5829.65
ℓ_c	2609.50	5241.00	5293.01	2619.57	5257.14	5299.69
$m = 4$	$-\mathcal{L}$	AIC	BIC	$-\mathcal{L}$	AIC	BIC
p_a	3096.91	6231.82	6319.70	2985.36	5994.73	6050.23
p_b	4322.77	8683.53	8775.57	4117.57	8259.14	8317.27
p_c	2206.93	4451.87	4541.71	2214.22	4452.45	4509.19
ℓ_a	2380.19	4798.38	4886.26	2390.87	4805.74	4861.24
ℓ_b	2880.72	5799.44	5891.48	2881.97	5787.95	5846.07
ℓ_c	2599.52	5237.05	5326.89	2615.98	5255.96	5312.71

Table 6.4: ***Comparison of hidden Markov models fitted to parasite and leukocyte counts by AIC and BIC.***
Parasite (p_a, p_b, p_c) and leukocyte (ℓ_a, ℓ_b, ℓ_c) counts are fitted to Poisson HMMs and negative binomial HMMs. The number of components is m. Minus log-likelihood ($-\mathcal{L}$) and information measures (AIC and BIC) are given. Models were fitted by maximum likelihood using the expectation-maximization (EM) algorithm, and validated by direct numerical maximization using `nlm` in `R`.

captures more heterogeneity in data than the (m-1)-state model. Both AIC and BIC, try to identify a model that optimally balances model fit and model complexity. These two criteria are plotted against the number of states m of the negative binomial HMM in Figure 6.3. Several comments arise from Figure 6.3. Unlike the NB mixtures, using two-state NB-HMM instead of one-state NB-HMM dramatically improves the fit to data. Little to no improvement in AIC is gained for $m \geq 3$. Ac-

cording to both AIC and BIC, the model with four states is the most appropriate for p_a. For the other counts, AIC and BIC selected different models. The Optimal numbers of states selected by LRT ($p \ll .0001$), AIC and BIC are given in Table 6.5. AIC and LRT selected the same models. Models selected by AIC and LRT are more complex than those selected by BIC since BIC penalizes larger models more. As it turns out, there is no clear "best" final model. One can narrow down his decision to the two selected NB-HMMs or investigate whether BIC, which selected a smaller "best" model, is more appropriate than AIC in this situation. This would be hard to pin down without extra-statistical information (scientific or practical). It should be noted, however, that the BIC increases consistently after a minimum is attained, while the AIC is more flat around the minimum. This evidence weighs in favour of the BIC.

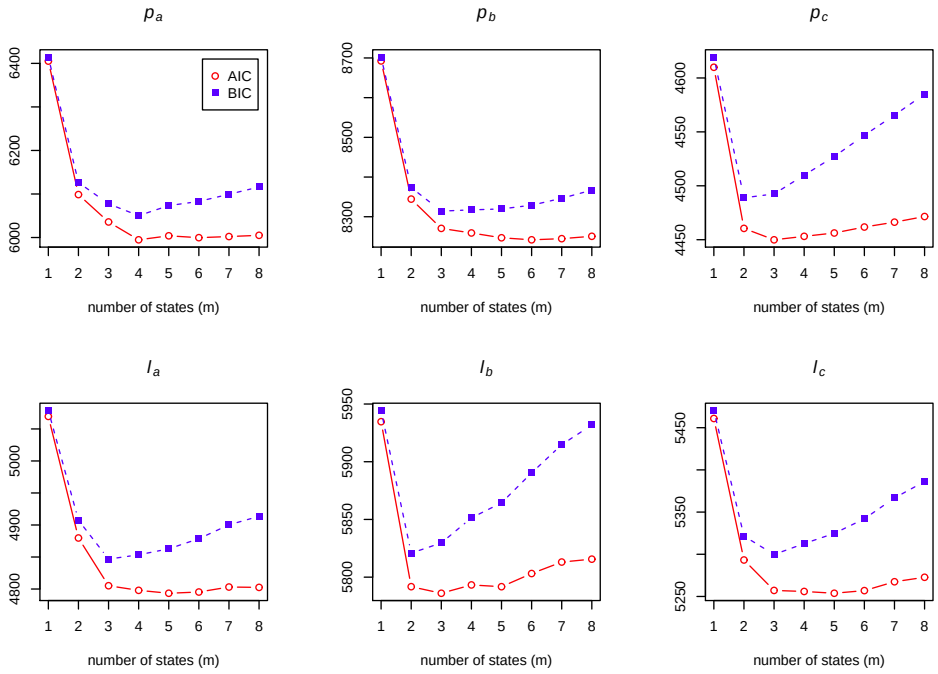

Figure 6.3: ***Model selection criteria of the fitted NB-HMMs.***
AIC and BIC are plotted against the number of states m of the negative binomial HMMs fitted to parasite (p_a, p_b, p_c) and leukocyte (ℓ_a, ℓ_b, ℓ_c) counts.

Even though the AIC and BIC selected two or three-state NB-HMMs for the parasite data p_c, one may consider the Poisson-HMMs as an acceptable alternative, since its AIC and BIC scores were only marginally higher than the competing models ($\triangle$AIC < 10 and $\triangle$BIC < 10). The latter has the advantage of being computationally tractable, while the NB-HMM is more complex as shown in Additional file 2 (higher number of parameters, no analytical solution for the MLE). Hence, one may check whether the Poisson-HMMs provides an adequate fit for the parasite

	p_a	p_b	p_c	ℓ_a	ℓ_b	ℓ_c
LRT	4	6	3	5	3	5
AIC	4	6	3	5	3	5
BIC	4	3	2	3	2	3

Table 6.5: ***Selection of the number of states of the fitted NB-HMMs.*** Three selection criteria (LRT, AIC and BIC) were used to select the optimal number of states of the negative binomial HMMs fitted to parasite (p_a, p_b, p_c) and leukocyte (ℓ_a, ℓ_b, ℓ_c) counts.

data p_c using pseudo-residuals. Figure 6.4 shows that the single Poisson distribution is definitely not appropriate since the pseudo-residuals deviate substantially from the standard normal distribution. In addition, many pseudo-residuals segments lie outside the bands of 0.5% and 99.5%. For the other models, very few observations stand out as extreme, histograms of pseudo-residuals are approximately normal-shaped and autocorrelations are "near zero" indicating low correlation in the residuals. However, the QQ-plots show that the upper quantiles are badly represented for the three and four-state Poisson-HMMs. Considering only the diagnostic plots, and not the model selection criteria, one can accept the two-state Poisson-HMM as the final fitting model for p_c.

6.4 Discussion

The Poisson formulation is seductive in its simplicity. It captures the discrete and nonnegative nature of count data, and naturally accounts for heteroscedastic and skewed distributions through its equidispersion property [Winkelmann 1995]. However, in most real data situations, equidispersion rarely occurs. The primary objective of the analysis reported in this paper was to test overdispersion in the distribution of parasites and leukocytes per HPF. Pearson's test was used to test for overdispersion in data. The data are shown to have too much variability to be represented by the Poisson distribution. The primary focus is on fitting the appropriate alternative model to parasite and leukocyte data. The goodness-of-fit of alternative models, designed to address the problem of overdispersion, is illustrated and discussed. The results show that the negative binomial (NB) model is the most appropriate (among simple parametric models), which suggests that parasites and leukocytes tend to aggregate together. The negative binomial has been widely used to inflate the Poisson dispersion as needed [Anderson 1993], and to analyse extra-dispersed count data [Shaw 1995, Alexander 2000, Saha 2009]. In addition, typical justifications for using the negative binomial formulation for count data go far beyond the existing critiques of overdispersion. Using the negative binomial distribution instead of the Poisson, allow to fix important errors in model specification [Berk 2008]. However, both the Poisson and the negative binomial distributions impose some special requirements the credibility of which also needs to be seriously assessed when statistical models for count data are constructed.

To explicitly account for the heterogeneity factor, an alternative model with additional free parameters may provide a better fit. In the case of the parasite and leukocytes counts, the Poisson mixture model and the negative binomial mixture model are proposed. The four-state Poisson model is prefered for two of the three TBSs. In order to further the analysis in the light of the authors' first intuition (that data tend to aggregate together), autocorrelation plots are examined. ACF suggests the existence of spatial dependence between neighbouring parasite and leukocyte counts. Moreover, investigating sources of overdispersion in data is enhanced by contrasting mixture models to HMMs. On the basis of AIC and BIC, HMMs are prefered. Information from neighbouring regions on TBSs is needed to better estimate this spatial dependence.

In this study, LRT and AIC select the same NB-HMMs, which seem to be the best fit for parasite and leukocyte distributions per field on selected TBSs. However, BIC selects less complex NB-HMMs. While it is true that, when fitted to the parasite and leukocyte data, the NB-HMM performed slightly better than the Poisson-HMM on the basis of AIC and BIC, both are reasonable models capable of describing the principal features of the data without using an excessive number of parameters. The NB-HMM perhaps has the advantage to incorporate an extra parameter to allow for overdispersion in parasite and leukocyte counts. However, with small differences in AIC (or BIC) score, i.e: $\triangle\text{AIC} < 10$ (or $\triangle\text{BIC} < 10$), a statistician may be tempted to choose the Poisson-HMM, which is computationally tractable, rather than its NB counterpart. Either more observations from TBSs or a convincing biological interpretation for one model rather than the other would be needed to take the discussion further. Contrary to the assumptions implicit within widely used simple parametric models, the fit to mixtures and HMMs viewed together are a reflection of the need for an heterogeneous modelling approach that explores the overdispersion in parasite and leukocyte counts.

While at first glance intuitively appealing for a statistician, detecting overdispersion in data is of highly questionable utility for malariologists. From a statistical standpoint, failure to take overdispersion into account leads to serious underestimation of the standard errors, biased parameter estimates and misleading inferences [Wang 1996]. In addition, changes in deviance (likelihood ratio statistic) will be very large and overly complex models will be selected accordingly. When overdispersion is present and ignored, using the Poisson model may overstate the significance of some covariates[Lee 2012] or give inconclusive evidence of interactions among them [McCullagh 1989]. From an epidemiological point of view, the importance of checking for overdispersion in parasite and leukocyte data stems from the need for epidemiological interpretations to be based on solid evidence. However, most existing PD estimation methods assume homogeneity in the distribution of parasites and leukocytes in TBSs. This assumption clearly does not hold. Likewise, the distribution of blood thickness within the smear will never be completely homogeneous [Dowling 1966], even under optimal conditions. Hence, the validity of the results of many statistical analyses, where PD is related to other explanatory variables, becomes suspect. For example, Enosse *et al.* [Enosse 2006] used a Poisson regression

to estimate the RTS,S/AS02A malaria vaccine effect, adjusted for parasite density, age, and time to infection. However, the comparison of the analysis outcomes with the primary outcomes of a non-parametric analysis using Mann-Whitney U test appears to show discrepancies. The authors concluded that the Poisson distribution did not adequately describe the data. Another example is the use of logistic regression to model the risk of fever as a continuous function of parasite density in order to estimate the fraction of fever attributable to malaria and to establish a case definition for the diagnosis of clinical malaria [Smith 1994, Mwangi 2005, Chandler 2006]. Case definition for symptomatic malaria is widely used in endemic areas. It requires fever together with a parasite density above a specific threshold. Even under declining levels of malaria endemicity, this method remains the reference method for discriminating malaria from other causes of fever and assessing malaria burden and trends [Roucher 2012]. Such estimates of the attributable fraction may be imprecise if the PD is not being estimated correctly. Furthermore, PD estimation methods potentially induce variability [Hammami 2013]. A proportion of this variability may be explained by the heterogeneity factor. An alternative PD estimation method that accounts for heterogeneity and spatial dependence between parasites and leukocytes in TBSs should be seriously considered in future epidemiological studies with field-collected PD data.

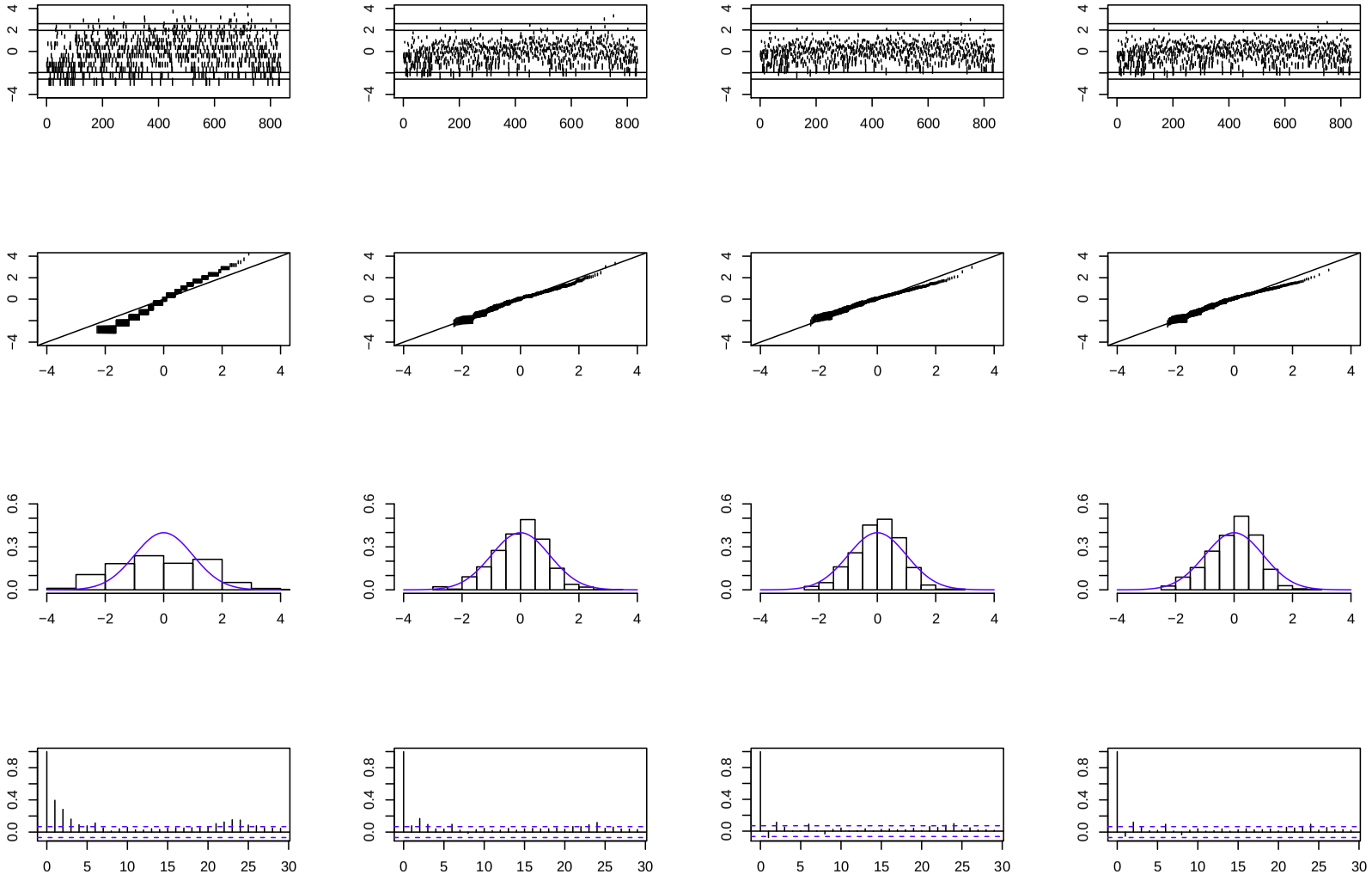

Figure 6.4: ***Diagnostic plots based on normal ordinary pseudo-residuals.***
Rows correspond to (1) index plots of the normal pseudo-residuals with horizontal lines at ±1.96 (2.5% and 97.5%) and ±2.58 (0.5% and 99.5%), (2) histograms of the normal pseudo-residuals with normal distribution curves in blue, (3) QQ-plots of the normal pseudo-residuals with theoretical quantiles on the horizontal axis, and (4) autocorrelation functions of the normal pseudo-residuals. Columns correspond to the Poisson-HMMs fitted to p_c data with 1, 2, 3 and 4 states respectively.

Chapter 7

Conclusion & Perspectives

"Do not follow where the path may lead. Go instead where there is no path and leave a trail."

Ralph Waldo Emerson

The work presented within these pages started from a field reality; namely, the problem of parasite density estimation in malaria. Accurate estimation of the parasite density is an essential clinical and epidemiological endpoint. Operational issues, as observed in the field, hided a more complex dilemma, which is the need to accurately assess the parasite density with a reasonable cost. Against this backdrop, we used a toolkit of statistical treatments to address the problem. Finally, we returned back to the field with an innovative parasite density estimation method.

Firstly, we studied the statistical properties (mean error, coefficient of variation, false negative rates) of parasite density estimators of commonly used threshold-based counting techniques depending on variable threshold values. We also assessed the influence of the thresholds on the cost-effectiveness of parasite density estimation methods. We were interested in the variability generated by these methods. We showed that this variability is a function of the parasite density and the threshold value. In addition, we gave more insights on the behavior of measurement errors according to varying threshold values, and on what should be the optimal threshold values that minimize this variability. Another important aspect of this study is that we observed how estimators perform at different parasitemia levels, and how much the choice of threshold values may influence the performance of estimators relative to each parasitemia level. We showed that estimators perform quite differently according to threshold values and the level of parasitemia, and that an overall performance measure probably hides a lot of complexity in the behavior of each estimator. While all estimators that have been considered had some deficiencies, adaptative methods, which take into account the level of parasitemia (and/or the abundance of leukocytes), outperformed non-adaptative methods in terms of accuracy and cost-effectiveness, and should therefore be seriously considered in future epidemiological surveys. In further support of the arguments cited here, empirical validation of the theoretical results is needed through a re-reading experience conducted in the field. Towards a better understanding of threshold effects, we are also interested in the study of the consequences of the quality of the parasite density estimators in models classically used in epidemiology and starting from these measures

(mixed effects linear and logistic regression, generalized linear models, etc). These aspects of the problem are now under consideration, and will be the subject of a publication at a later stage. Furthermore, variability in parasite density measurements may be explained by other sources including the reader skills, the inter-rater reliability, the quality of the slide, the amount of blood examined, the loss of parasite during the staining and the dehaemoglobinization of the thick blood smear and the distribution of parasites and leukocytes in thick blood smears. These sources of error have also to be addressed and assessed. The training of microscopists and the standardization of estimation methods could potentially reduce this variability, and then increase the accuracy and the efficiency of parasite density estimates.

Secondly, in order to investigate overdispersion in the distribution of parasite and leukocyte in thick blood smears, we constituted and published the first dataset on parasite and leukocyte counts per high power field. Based on these data, we found evidence that the Poisson assumption is inconsistent with the parasite and leukocyte distributions. Among simple parametric models considered, the negative binomial (NB) model is the closest to the unknown distribution that generates the data. On the basis of model selection criteria AIC and BIC, the NB-HMMs provide a better fit to data than Poisson mixtures. While it is true that, when fitted to the parasite and leukocyte counts, the NB-HMM is a slightly better choice than the Poisson-HMM model on the basis of AIC and BIC, both are reasonable models capable of describing the principal features of the data without using an excessive number of parameters. The NB-HMM model perhaps has the advantage to incorporate an extra parameter to allow for overdispersion in the parasite and leukocytes data. Either more observations from thick blood smears or a convincing biological interpretation for one model rather than the other would be needed to take the discussion further. Contrary to the assumptions implicit within widely used simple parametric models, the fit to mixtures and HMMs viewed together are a reflection of the need for an heterogeneous modeling approach that explores the overdispersion in parasite and leukocyte counts.

Finally, we devised a reduced reading procedure for the examination of the thick blood smears that aims to a better operational optimization and to a practical assessing of the heterogeneity in the distribution of parasites in the thick film. The motivations behind the design of this alternative protocol are the need to optimize the cost of epidemiological surveys and to reduce the inescapable loss of information. This new counting device is an appropriate protocol for field experience. It allows for heterogeneity detection. It is at least as accurate and precise as threshold-based counting methods. This technique is potentially useful for laboratories that routinely perform malaria parasite enumeration, since it requires neither special equipments, nor operator decisions that might bias the outcome. A patent application process has been launched in October, 2012. A prototype development of the counter is in process. The outcomes and impact of the new counting method are under consideration and can be viewed as a starting point for further developments.

Looking back, I got to notice that until now different steps have been tackled in the project, but much still needs to be done. The problem of parasite density

estimation is as appealing as critical. The overall aim of this work was to make a small contribution towards better understanding of the malaria issue in an attempt to improve malaria diagnostic tools. We deeply believe that the overall contributions and results generated by the scientific community, in all fields of science, from both a public health and a socio-economic perspective, will make a difference. Perhaps in some years from now, the reader of these pages will find the subject completely outdated.

Bibliography

[Abeku 2007] Tarekegn A. Abeku. *Response to Malaria Epidemics in Africa.* Emerg Infect Dis, vol. 13, no. 5, pages 681–686, 2007. (Cited on page 13.)

[Achan 2011] Jane Achan, Ambrose Talisuna, Annette Erhart, Adoke Yeka, James Tibenderana, Frederick Baliraine, Philip Rosenthal and Umberto D'Alessandro. *Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria.* Malaria Journal, vol. 10, no. 1, page 144, 2011. (Cited on page 25.)

[Acker 1967] P. Acker, L. Maydat, P. Trapet, C. Fourcade and H. Sagnet. *Quelques constantes biologiques actuelles de l'Africain congolais normal.* Bulletin de la Société de Pathologie exotique, vol. 60, pages 460–467, 1967. (Cited on page 38.)

[Adu-Gyasi 2012] Dennis Adu-Gyasi, Mohammed Adams, Sabastina Amoako, Emmanuel Mahama, Maxwell Nsoh, Seeba Amenga-Etego, Frank Baiden, Kwaku Asante, Sam Newton and Seth Owusu-Agyei. *Estimating malaria parasite density: assumed white blood cell count of 10,000/µl of blood is appropriate measure in Central Ghana.* Malaria Journal, vol. 11, no. 1, page 238, 2012. (Cited on pages 38 and 84.)

[Akaike 1973] H Akaike. Information theory and an extension of the maximum likelihood principle, volume 1, pages 267–281. Akademiai Kiado, 1973. (Cited on page 89.)

[Alexander 2000] Neal Alexander, Rana Moyeed and Julian Stander. *Spatial modelling of individual-level parasite counts using the negative binomial distribution.* Biostatistics, vol. 1, no. 4, pages 453–463, 2000. (Cited on page 96.)

[Alexander 2010] Neal Alexander, David Schellenberg, Billy Ngasala, Max Petzold, Chris Drakeley and Colin Sutherland. *Assessing agreement between malaria slide density readings.* Malaria Journal, vol. 9, no. 1, page 4, 2010. (Cited on pages 3, 4, 39, 43, 45, 61, 62, 63 and 82.)

[Allen 2011] Lisa Allen, Jennifer Hatfield, Giselle DeVetten, Jeremy Ho and Mange Manyama. *Reducing malaria misdiagnosis: the importance of correctly interpreting Paracheck Pf(R) "faint test bands" in a low transmission area of Tanzania.* BMC Infectious Diseases, vol. 11, page 308, 2011. (Cited on page 84.)

[Alonso 1994] P.L. Alonso, T. Smith and J.R.M. Armstrong Schellenberg. *Randomised trial of efficacy of Spf66 vaccine against Plasmodium falciparum malaria in children in southern Tanzania.* Lancet, vol. 344, pages 1175–1181, 1994. (Cited on page 36.)

[Amexo 2004] M. Amexo, R. Tolhurst, G. Barnish and I. Bates. *Malaria misdiagnosis: effects on the poor and vulnerable.* Lancet, vol. 364, pages 1896–1898, 2004. (Cited on pages 2 and 36.)

[Anderson 1993] R. M. Anderson. *Epidemiology.* In F. E. G. Cox, editeur, Modern Parasitology, chapitre 4, pages 75–116. Blackwell Publishing Ltd., 2nd édition, 1993. (Cited on page 96.)

[Aribodor 2009] Dennis Aribodor, Obioma Nwaorgu, Christine Eneanya, Ikechukwu Okoli, Reed Pukkila-Worley and Harrison Etaga. *Association of low birth weight and placental malarial infection in Nigeria.* The Journal of Infection in Developing Countries, vol. 3, no. 8, pages 620–623, 2009. (Cited on page 19.)

[Bartlett 2003] John M.S. Bartlett and David Stirling. *A Short History of the Polymerase Chain Reaction.* In John M.S. Bartlett and David Stirling, editeurs, PCR Protocols, volume 226 of *Methods in Molecular Biology*, pages 3–6. Humana Press, 2003. (Cited on page 24.)

[Bates 2004] I. Bates, V. Bekoe and A. Asamoa-Adu. *Improving the accuracy of malaria-related laboratory tests in Ghana.* Malar J, vol. 3, page 38, 2004. (Cited on pages 36 and 39.)

[Baum 1966] Leonard E. Baum and Ted Petrie. *Statistical Inference for Probabilistic Functions of Finite State Markov Chains.* The Annals of Mathematical Statistics, vol. 37, no. 6, pages 1554–1563, 1966. (Cited on page 86.)

[Baum 1967] Leonard E Baum and J A Eagon. *An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology.* Bulletin of the American Mathematical Society, vol. 73, no. 3, pages 360–363, 1967. (Cited on page 86.)

[Baum 1970] Leonard E. Baum, Ted Petrie, George Soules and Norman Weiss. *A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains.* The Annals of Mathematical Statistics, vol. 41, no. 1, pages 164–171, 1970. (Cited on page 87.)

[Becher 2005] H. Becher and B.B. Kouyaté. Health research in developing countries: A collaboration between burkina faso and germany. European Consortium for Mathematics in Industry. Springer London, Limited, 2005. (Cited on pages 3 and 82.)

[Bejon 2006] Philip Bejon, Laura Andrews, Angela Hunt-Cooke, Frances Sanderson, Sarah Gilbert and Adrian Hill. *Thick blood film examination for Plasmodium falciparum malaria has reduced sensitivity and underestimates parasite density.* Malaria Journal, vol. 5, no. 1, page 104, 2006. (Cited on pages 3, 21, 33, 39 and 82.)

[Bejon 2008] Philip Bejon, John Lusingu, Ally Olotu, Amanda Leach, Marc Lievens, Johan Vekemans, Salum Mshamu, Trudie Lang, Jayne Gould, Marie-Claude Dubois, Marie-Ange Demoitié, Jean-Francois Stallaert, Preeti Vansadia, Terrell Carter, Patricia Njuguna, Ken O. Awuondo, Anangisye Malabeja, Omar Abdul, Samwel Gesase, Neema Mturi, Chris J. Drakeley, Barbara Savarese, Tonya Villafana, W. Ripley Ballou, Joe Cohen, Eleanor M. Riley, Martha M. Lemnge, Kevin Marsh and Lorenz von Seidlein. *Efficacy of RTS,S/AS01E Vaccine against Malaria in Children 5 to 17 Months of Age.* New England

Journal of Medicine, vol. 359, no. 24, pages 2521–2532, 2008. (Cited on page 27.)

[Berk 2008] Richard Berk and John M MacDonald. *Overdispersion and Poisson Regression*. Journal of Quantitative Criminology, vol. 24, no. 3, pages 269–284, 2008. (Cited on page 96.)

[Bland 1986] MJ Bland and DG Altman. *Statistical methods for assessing agreement between two methods of clinical measurement*. Lancet, vol. 1, pages 307–310, 1986. (Cited on pages 39 and 62.)

[Bliss 1953] C. I. Bliss and R. A. Fisher. *Fitting the Negative Binomial Distribution to Biological Data*. Biometrics, vol. 9, no. 2, pages 176–200, 1953. (Cited on page 90.)

[Blistein 1950] I. Blistein. *Hématologie normale des Noirs du Congo. Sang et moelle osseuse des adultes*. Annales de la Société belge de Médecine tropicale, vol. 12, pages 273–294, 1950. (Cited on page 38.)

[Boes 2007] Stefan Boes. *Count Data Models with Unobserved Heterogeneity: An Empirical Likelihood Approach*. Rapport technique 0704, University of Zurich, Socioeconomic Institute, March 2007. (Cited on page 83.)

[Box 1976a] G. E. P. Box. *Science and Statistics*. Journal of the American Statistical Association, vol. 71, no. 356, pages 791–799, 1976. (Cited on page 80.)

[Box 1976b] G.E.P. Box and G.M. Jenkins. Time series analysis: forecasting and control. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976. (Cited on pages 5 and 88.)

[Boyd 1949] M. F. Boyd, R. Christophers and L. T. Coggeshall. *The mechanism of immunity against malaria in communities living under hyperendemic conditions*. In M. F. Boyd, editeur, Malariology. Philadelphia and London:Saunders Company, 1949. (Cited on page 3.)

[Braun-Munzinger 1992] RA Braun-Munzinger and BA Southgate. *Repeatability and reproducibility of egg counts of Schistosoma haematobium in urine*. Trop Med Parasitol, vol. 43, pages 149–154, 1992. (Cited on page 62.)

[Briand 2009] Valérie Briand, Julie Bottero, Harold Noël, Virginie Masse, Hugues Cordel, José Guerra, Hortense Kossou, Benjamin Fayomi, Paul Ayemonna, Nadine Fievet, Achille Massougbodji and Michel Cot. *Intermittent Treatment for the Prevention of Malaria during Pregnancy in Benin: A Randomized, Open-Label Equivalence Trial Comparing Sulfadoxine-Pyrimethamine with Mefloquine*. Journal of Infectious Diseases, vol. 200, no. 6, pages 991–1001, 2009. (Cited on page 26.)

[Bruce-Chwatt 1958] L.J. Bruce-Chwatt. *Parasite Density Index in Malaria*. Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 52, no. Issue 4, page 389, 1958. (Cited on pages 38 and 84.)

[Bruce-Chwatt 1985] L.J. Bruce-Chwatt. Essential malariology. A Wiley medical publication. Wiley, 1985. (Cited on pages 37 and 84.)

[Cameron 1986] A. Colin Cameron and Pravin K. Trivedi. *Econometric models based on count data. Comparisons and applications of some estimators and tests.* Journal of Applied Econometrics, vol. 1, no. 1, pages 29–53, 1986. (Cited on pages 83 and 85.)

[Cappé 2005] Olivier Cappé, Eric Moulines and Tobias Rydén. Inference in hidden markov models. Springer series in statistics. Springer, New York, 2005. (Cited on page 87.)

[Carter 2002] Richard Carter and Kamini N. Mendis. *Evolutionary and Historical Aspects of the Burden of Malaria.* Clinical Microbiology Reviews, vol. 15, no. 4, pages 564–594, 2002. (Cited on page 20.)

[Cartwright 1968] G.E. Cartwright. Diagnostic laboratory hematology. Grune & Stratton, 1968. (Cited on page 38.)

[CDC 2010] CDC. *Eradication of Malaria in the United States (1947-1951)*, February 2010. [Retrieved 2012-05-02.]. (Cited on page 26.)

[Chadli 1986] A. Chadli, M.F. Kennou and J. Kooli. *Campaigns for the eradication of malaria in Tunisia: history and current situation.* Arch Inst Pasteur Tunis, vol. 63, no. 1, pages 35–50, 1986. (Cited on page 30.)

[Chakravarti 1967] I.M. Chakravarti, R.G. Laha and J. Roy. Handbook of methods of applied statistics. Numeéro vol. 1 de Wiley series in probability and mathematical statistics. Wiley, 1967. (Cited on pages 4 and 87.)

[Chandler 2006] Clare I. R. Chandler, Chris J. Drakeley, Hugh Reyburn and Ilona Carneiro. *The effect of altitude on parasite density case definitions for malaria in northeastern Tanzania.* Tropical Medicine & International Health, vol. 11, no. 8, pages 1178–1184, 2006. (Cited on pages 3, 82 and 98.)

[Chippaux 1991] Jean-Philippe Chippaux, M. Akogbeto, A. Massougbodji and J. Adjagba. *Mesure de la parasitémie palustre et évaluation du seuil pathogène en région de forte transmission permanente.* In Le paludisme en Afrique de l'Ouest : études entomologiques et épidémiologiques en zone rizicole et en milieu urbain, Etudes et Thèses, pages 55–65. ORSTOM, 1991. (Cited on pages 38 and 42.)

[Christophers 1924] S.R. Christophers. *The mechanism of immunity against malaria in communities living under hyperendemic conditions.* Indian Journal of Medical Research, vol. 12, pages 273–294, 1924. (Cited on page 2.)

[Christophers 1951] Rickard Christophers. *Microscopic Diagnosis of Malaria.* British medical journal, vol. 1, no. 4697, pages 75–76, 1951. (Cited on page 21.)

[Cibulskis 2011] Richard E. Cibulskis, Maru Aregawi, Ryan Williams, Mac Otten and Christopher Dye. *Worldwide Incidence of Malaria in 2009: Estimates, Time Trends, and a Critique of Methods.* PLoS Med, vol. 8, no. 12, page e1001142, 12 2011. (Cited on page 15.)

[Clark 1989] Suzanne J. Clark and Joe N. Perry. *Estimation of the Negative Binomial Parameter κ by Maximum Quasi -Likelihood.* Biometrics, vol. 45, no. 1, pages pp. 309–316, 1989. (Cited on page 83.)

[Clendennen 1995] T.E. Clendennen, G.W. Long and K. Baird. *QBC©and Giemsa-stained thick blood films: diagnostic performance of laboratory technologists.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 89, pages 183–184, 1995. (Cited on pages 39 and 62.)

[Cohen 2005] Cheryl Cohen, Alan Karstaedt, John Frean, Juno Thomas, Nelesh Govender, Elizabeth Prentice, Leigh Dini, Jacky Galpin and Heather Crewe-Brown. *Increased Prevalence of Severe Malaria in HIV-Infected Adults in South Africa.* Clinical Infectious Diseases, vol. 41, no. 11, pages 1631–1637, 2005. (Cited on page 20.)

[Colbourne 1971] M J Colbourne. *The laboratory diagnosis of malaria.* Tropical Doctor, vol. 1, no. 4, pages 161–163, 1971. (Cited on page 21.)

[Coleman 2006] R.E. Coleman, J. Sattabongkot, S. Promstaporm, N. Maneechai, B. Tippayachai, A. Kengluecha, N. Rachapaew, G. Zollner, R.S. Miller, J.A. Vaughan, K. Thimasarn and B. Khuntirat. *Comparison of PCR and microscopy for the detection of asymptomatic malaria in a Plasmodium falciparum/vivax endemic area in Thailand.* Malaria Journal, vol. 5, page 121, 2006. (Cited on pages 3 and 24.)

[Collier 1983] J A Collier and J M Longmore. *The reliability of the microscopic diagnosis of malaria in the field and in the laboratory.* Annals of Tropical Medicine and Parasitology, vol. 77, no. 2, pages 113–117, 1983. (Cited on page 21.)

[Cox 1983] D.R. Cox. *Some Remarks on Overdispersion.* Biometrika, vol. 70, page 269, 1983. (Cited on page 82.)

[Craig 1999] M.H. Craig, R.W. Snow and D. le Sueur. *A Climate-based Distribution Model of Malaria Transmission in Sub-Saharan Africa.* Parasitology Today, vol. 15, no. 3, pages 105–111, 1999. (Cited on page 14.)

[Curtis 1986] C.F. Curtis and L.N. Otoo. *A simple model of the build-up of resistance to mixtures of antimalarial drugs.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 80, pages 889–892, 1986. (Cited on page 25.)

[Damien 2010] Georgia Damien, Armel Djenontin, Christophe Rogier, Vincent Corbel, Sahabi Bangana, Fabrice Chandre, Martin Akogbeto, Dorothee Kinde-Gazard, Achille Massougbodji and Marie-Claire Henry. *Malaria infection and disease in an area with pyrethroid-resistant vectors in southern Benin.* Malaria Journal, vol. 9, no. 1, page 380, 2010. (Cited on pages 3 and 82.)

[Dannemann 2008] Jörn Dannemann and Hajo Holzmann. *Testing for two states in a hidden Markov model.* Canadian Journal of Statistics, vol. 36, no. 4, pages 505–520, 2008. (Cited on page 79.)

[Darwin 1957] J.H. Darwin. *The Power of the Poisson Index of Dispersion.* Biometrika, vol. 44, page 286, 1957. (Cited on page 82.)

[Dean 1989] C. Dean and J. F. Lawless. *Tests for Detecting Overdispersion in Poisson Regression Models.* Journal of the American Statistical Association, vol. 84, no. 406, pages pp. 467–472, 1989. (Cited on pages 83 and 85.)

[Dean 1992] C. Dean. *Testing for Overdispersion in Poisson and Binomial Regression Models.* Journal of the American Statistical Association, vol. 87, page 451, 1992. (Cited on page 83.)

[Dempster 1977] A. Dempster, N. Laird and D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, vol. 39 (Series B), pages 1–38, 1977. (Cited on pages 66, 67 and 87.)

[Dennis 1983] J.E. Dennis and R.B. Schnabel. Numerical methods for unconstrained optimization and nonlinear equations. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1983. (Cited on page 88.)

[Dini 2003] Leigh Dini and John Frean. *Quality assessment of malaria laboratory diagnosis in South Africa.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 97, no. 6, pages 675–677, 2003. (Cited on pages 36 and 39.)

[Djenontin 2010] Armel Djenontin, Sahabi Bio-Bangana, Nicolas Moiroux, Marie-Claire Henry, Olayide Bousari, Joseph Chabi, Razaki Osse, Sebastien Koudenoukpo, Vincent Corbel, Martin Akogbeto and Fabrice Chandre. *Culicidae diversity, malaria transmission and insecticide resistance alleles in malaria vectors in Ouidah-Kpomasse-Tori district from Benin (West Africa): A pre-intervention study.* Parasites & Vectors, vol. 3, no. 1, page 83, 2010. (Cited on page 83.)

[Dondorp 2010] Arjen M Dondorp, Caterina I Fanello, Ilse CE Hendriksen, Ermelinda Gomes, Amir Seni, Kajal D Chhaganlal, Kalifa Bojang, Rasaq Olaosebikan, Nkechinyere Anunobi, Kathryn Maitland, Esther Kivaya, Tsiri Agbenyega, Samuel Blay Nguah, Jennifer Evans, Samwel Gesase, Catherine Kahabuka, George Mtove, Behzad Nadjm, Jacqueline Deen, Juliet Mwanga-Amumpaire, Margaret Nansumba, Corine Karema, Noella Umulisa, Aline Uwimana, Olugbenga A Mokuolu, Olanrewaju T Adedoyin, Wahab BR Johnson, Antoinette K Tshefu, Marie A Onyamboko, Tharisara Sakulthaew, Wirichada Pan Ngum, Kamolrat Silamut, Kasia Stepniewska, Charles J Woodrow, Delia Bethell, Bridget Wills, Martina Oneko, Tim E Peto, Lorenz von Seidlein, Nicholas PJ Day and Nicholas J White. *Artesunate versus quinine in the treatment of severe falciparum malaria in African children (AQUAMAT): an open-label, randomised trial.* The Lancet, vol. 376, no. 9753, pages 1647 – 1657, 2010. (Cited on page 25.)

[Doolan 2009] Denise L. Doolan, Carlota Dobaño and J. Kevin Baird. *Acquired Immunity to Malaria.* Clinical Microbiology Reviews, vol. 22, no. 1, pages 13–36, January 2009. (Cited on page 20.)

[Dowling 1966] M. A. C. Dowling and G. T. Shute. *A comparative study of thick and thin blood films in the diagnosis of scanty malaria parasitaemia.* Bulletin of the World Health Organization, vol. 34, pages 249–267, 1966. (Cited on pages 3, 34, 36, 37, 38, 39, 46, 49, 62, 82, 84 and 97.)

[Draper 1971] C C Draper. *Malaria. Laboratory diagnosis.* British medical journal, vol. 2, no. 5753, pages 93–95, 1971. (Cited on page 21.)

[Dubey 1999] M.L. Dubey, C. Weingken, N.K. Ganguly and R.C. Mahajan. *Comparative evaluation of methods of malaria parasite density determination in blood samples from patients and experimental animals.* Indian J Med Res, vol. 109, pages 20–27, 1999. (Cited on pages 3, 39 and 62.)

[Earle 1932] W. C. Earle and M. Perez. *Enumeration of parasites in the blood of malaria patients.* Journal of Laboratory and Clinical Medicine, vol. 17, pages 1124–1130, 1932. (Cited on pages 3 and 37.)

[Earle 1939] W. C. Earle, M. Perez, J. Delrio and C. Arzola. *Observations on the course of naturally acquired malaria in Puerto Rico.* Puerto Rico Journal of Public Health and Tropical Medicine, vol. 14, pages 391–406, 1939. (Cited on page 2.)

[Enosse 2006] Sonia Enosse, Carlota Dobaño, Diana Quelhas, John J Aponte, Marc Lievens, Amanda Leach, Jahit Sacarlal, Brian Greenwood, Jessica Milman, Filip Dubovsky, Joe Cohen, Ricardo Thompson, W. Ripley Ballou, Pedro L Alonso, David J Conway and Colin J Sutherland. *RTS,S/AS02A Malaria Vaccine Does Not Induce Parasite CSP T Cell Epitope Selection and Reduces Multiplicity of Infection.* PLOS Clin Trial, vol. 1, no. 1, page e5, 05 2006. (Cited on pages 3, 82 and 97.)

[Färnert 2009a] Anna Färnert, Thomas N. Williams, Tabitha W. Mwangi, Anna Ehlin, Greg Fegan, Alex Macharia, Brett S. Lowe, Scott M. Montgomery and Kevin Marsh. *Transmission-Dependent Tolerance to Multiclonal Plasmodium falciparum Infection.* Journal of Infectious Diseases, vol. 200, no. 7, pages 1166–1175, 2009. (Cited on pages 3 and 82.)

[Färnert 2009b] Anna Färnert, Thomas N. Williams, Tabitha W. Mwangi, Anna Ehlin, Greg Fegan, Alex Macharia, Brett S. Lowe, Scott M. Montgomery and Kevin Marsh. *Transmission-Dependent Tolerance to Multiclonal Plasmodium falciparum Infection.* Journal of Infectious Diseases, vol. 200, no. 7, pages 1166–1175, 2009. (Cited on page 20.)

[Faulde 2007] Michael K. Faulde, Ralf Hoffmann, Khair M. Fazilat and Achim Hoerauf. *Malaria Reemergence in Northern Afghanistan.* Emerg Infect Dis, vol. 13, no. 9, pages 1402–1404, 2007. (Cited on page 14.)

[Field 1963] J. W. Field, A. A. Sandosham and Y. I. Fong. The microscopical diagnosis of human malaria. I. A morphological study of the erythrocytic parasites in thick blood films, volume 30. Studies from the Institute for Medical Research. Federation of Malaya, second édition, 1963. (Cited on page 3.)

[Filmer 2005] Deon Filmer. *Fever and its treatment among the more and less poor in sub-Saharan Africa.* Health Policy and Planning, vol. 20, no. 6, pages 337–346, 2005. (Cited on page 29.)

[Finkel 2007] Michael Finkel. *Malaria- Stopping a Global Killer.* National Geographic, 2007. (Cited on page 13.)

[Fisher 1950] R.A. Fisher. *The significance of deviations from expectation in a Poisson series.* Biometrics, vol. 6, pages 17–24, 1950. (Cited on page 87.)

[Fleiss 1973] J. L. Fleiss and J. Cohen. *The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.* Educational and Psychological Measurement, vol. 33, pages 613–619, 1973. (Cited on page 62.)

[Garcia 1998] A. Garcia, M. Cot, J.P. Chippaux, S. Ranques, J. Feingold, F. Demenais and L. Abel. *Genetic control of blood infection levels in human malaria: evidence for a complex genetic model.* Am J Trop Med Hyg, vol. 58, pages 480–488, 1998. (Cited on page 36.)

[Garcia 2004] A. Garcia, A. Dieng, F. Rouget, F. Migot-Nabias, J.Y. Le Hesran and O. Gaye. *Role of environment and behaviour in familial resemblances of Plasmodium falciparum infection in a population of Senegalese children.* Microbes and Infections, vol. 6, pages 68–75, 2004. (Cited on page 36.)

[Garner 2005] Paul Garner and Patricia M Graves. *The Benefits of Artemisinin Combination Therapy for Malaria Extend Beyond the Individual Patient.* PLoS Med, vol. 2, no. 4, page e105, 04 2005. (Cited on page 25.)

[Gassiat 2003] E. Gassiat and S. Boucheron. *Optimal error exponents in hidden Markov models order estimation.* Information Theory, IEEE Transactions on, vol. 49, no. 4, pages 964 – 980, april 2003. (Cited on page 79.)

[Geels 2011] Mark J Geels, Egeruan B Imoukhuede, Nathalie Imbault, Harry van Schooten, Terry McWade, Marita Troye-Blomberg, Roland Dobbelaer, Alister G Craig and Odile Leroy. *European Vaccine Initiative: lessons from developing malaria vaccines.* Expert Review of Vaccines, vol. 10, no. 12, pages 1697–1708, 2011. (Cited on page 27.)

[Gething 2011] Peter Gething, Anand Patil, David Smith, Carlos Guerra, Iqbal Elyazar, Geoffrey Johnston, Andrew Tatem and Simon Hay. *A new world malaria map: Plasmodium falciparum endemicity in 2010.* Malaria Journal, vol. 10, no. 1, page 378, 2011. (Cited on page 17.)

[Gourieroux 1984] C. Gourieroux, A. Monfort and A. Trognon. *Pseudo Maximum Likelihood Methods: Theory.* Econometrica, vol. 52, no. 3, pages 681–700, 1984. (Cited on page 83.)

[Graves 2006a] Patricia M. Graves and Hellen Gelband. *Vaccines for preventing malaria (blood-stage).* Cochrane Database of Systematic Reviews, vol. 18, no. 4, 2006. (Cited on page 27.)

[Graves 2006b] Patricia M. Graves and Hellen Gelband. *Vaccines for preventing malaria (pre-erythrocytic).* Cochrane Database of Systematic Reviews, vol. 18, no. 4, 2006. (Cited on page 27.)

[Graves 2006c] Patricia M. Graves and Hellen Gelband. *Vaccines for preventing malaria (SPf66).* Cochrane Database of Systematic Reviews, vol. 19, no. 2, 2006. (Cited on page 27.)

[Greenwood 1987] BM. Greenwood, AK. Bradley, AM. Greenwood and et al. *Mortality and morbidity from malaria among children in rural area of Gambia, West Africa.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 81, pages 478–486, 1987. (Cited on pages 38 and 84.)

[Greenwood 1991] B.M. Greenwood and J.R. Armstrong. *Comparison of two simple methods for determining malaria parasite density.* Trans R Soc Trop Med Hyg, vol. 85, pages 186–188, 1991. (Cited on pages 3, 38, 39, 42 and 62.)

[Gschlößl 2008] Susanne Gschlößl and Claudia Czado. *Modelling count data with overdispersion and spatial effects.* Statistical Papers, vol. 49, pages 531–552, 2008. (Cited on page 83.)

[Guerra 2007] Carlos Guerra, Simon Hay, Lorena Lucioparedes, Priscilla Gikandi, Andrew Tatem, Abdisalan Noor and Robert Snow. *Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project.* Malaria Journal, vol. 6, no. 1, page 17, 2007. (Cited on page 14.)

[Guerrant 2006] R.L. Guerrant, D.H. Walker and P.F. Weller. Tropical infectious diseases: principles, pathogens & practice. Numeéro vol. 2 de Tropical Infectious Diseases: Principles, Pathogens & Practice. Elsevier Churchill Livingstone, 2006. (Cited on page 22.)

[Gurmu 1991] Shiferaw Gurmu. *Tests for Detecting Overdispersion in the Positive Poisson Regression Model.* Journal of Business & Economic Statistics, vol. 9, no. 2, pages pp. 215–222, 1991. (Cited on pages 83 and 85.)

[Gurmu 1998] S. Gurmu, P. Rilstone and S. Stern. *Semiparametric Estimation of Count Regression Models.* Journal of Econometrics, vol. 88, pages 123–150, 1998. (Cited on page 83.)

[Guyatt 2004] Helen L. Guyatt and Robert W. Snow. *Impact of Malaria during Pregnancy on Low Birth Weight in Sub-Saharan Africa.* Clinical Microbiology Reviews, vol. 17, no. 4, pages 760–769, 2004. (Cited on page 19.)

[Hammami 2013] Imen Hammami, Grégory Nuel and André Garcia. *Statistical Properties of Parasite Density Estimators in Malaria.* PLoS ONE, vol. 8, no. 3, 2013. (Cited on pages 3, 82 and 98.)

[Hartigan 1979] J. A. Hartigan and M. A. Wong. *Algorithm AS 136: A k-means clustering algorithm.* Applied Statistics, vol. 28, no. 1, pages 100–108, 1979. (Cited on pages 71 and 88.)

[Hausman 1984] Jerry A. Hausman, Bronwyn H. Hall and Zvi Griliches. *Econometric Models for Count Data with an Application to the Patents-R&D Relationship.* Working Paper 17, National Bureau of Economic Research, October 1984. (Cited on page 85.)

[Heilbron 1994] David C. Heilbron. *Zero-Altered and other Regression Models for Count Data with Added Zeros.* Biometrical Journal, vol. 36, no. 5, pages 531–547, 1994. (Cited on page 86.)

[Hermsen 2001] C.C. Hermsen, Telgt D.S., E.H. Linders, van de Locht L.A., W.M. Eling, E.J. Mensink and R.W. Sauerwein. *Detection of Plasmodium falciparum malaria parasites in vivo by real-time quantitative PCR.* Mol Biochem Parasitol, vol. 118, no. 8, pages 247–251, 2001. (Cited on page 24.)

[Huho 2012] Bernadette Huho, Gerard Killeen, Heather Ferguson, Adriana Tami, Christian Lengeler, J Derek Charlwood, Aniset Kihonda, Japhet Kihonda, S Patrick Kachur, Thomas Smith and Salim Abdulla. *Artemisinin-based combination therapy does not measurably reduce human infectiousness to vectors in a setting of intense malaria transmission.* Malaria Journal, vol. 11, no. 1, page 118, 2012. (Cited on page 25.)

[Jacquerioz 2009] F.A. Jacquerioz and M. Croft Ashley. *Drugs for preventing malaria in travellers.* Cochrane Database of Systematic Reviews, no. 4, 2009. (Cited on page 28.)

[Jamieson 2006] A. Jamieson and S. Toovey. Malaria: A traveller's guide. Struik, 2006. (Cited on page 13.)

[Jelliffen 1968] E. F. Patricia Jelliffen. *Low birth-weight and malarial infection of the placenta.* Bull World Health Organ, vol. 38, no. 1, pages 69–78, 1968. (Cited on page 19.)

[Jeremiah 2007] Zaccheaus Awortu Jeremiah and Emmanuel Kufre Uko. *Comparative analysis of malaria parasite density using actual and assumed white blood cell counts.* Annals of Tropical Paediatrics: International Child Health, vol. 27, no. 1, pages 75–79, 2007. (Cited on page 38.)

[J.F. 1987] Lawless J.F. *Negative Binomial and Mixed Poisson Regression.* Canadian Journal of Statistics, vol. 15, page 209, 1987. (Cited on page 83.)

[Joe 2005] Harry Joe and Rong Zhu. *Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution.* Biometrical Journal, vol. 47, no. 2, pages 219–229, 2005. (Cited on page 83.)

[Kamya 2006] Moses R. Kamya, Anne F. Gasasira, Adoke Yeka, Nathan Bakyaita, Samuel L. Nsobya, Damon Francis, Philip J. Rosenthal, Grant Dorsey and Diane Havlir. *Effect of HIV-1 Infection on Antimalarial Treatment Outcomes in Uganda: A Population-Based Study.* Journal of Infectious Diseases, vol. 193, no. 1, pages 9–15, 2006. (Cited on page 20.)

[Kilian 2000] A.H. Kilian, W.G. Metzger, E.J. Mutschelknauss, G. Kabagambe, P. Langi, R. Korte and F. Von Sonnenburg. *Reliability of malaria microscopy in epidemiological studies: results of quality control.* Trop Med Int Health, vol. 5, pages 3–8, 2000. (Cited on pages 21, 34 and 36.)

[King 1989] Gary. King. *Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator.* American Journal of Political Science, vol. 33, pages 762–784, 1989. (Cited on pages 4, 85 and 88.)

[Kirkwood 2001] Betty Kirkwood and Jonathan Sterne. Essentials of medical statistics. Wiley-Blackwell, 2 édition, April 2001. (Cited on pages 4 and 45.)

[Kokwaro 2009] Gilbert Kokwaro. *Ongoing challenges in the management of malaria.* Malaria Journal, vol. 8, no. Suppl 1, page S2, 2009. (Cited on page 25.)

[Krefis 2010] Anne Krefis, Norbert Schwarz, Bernard Nkrumah, Samuel Acquah, Wibke Loag, Nimako Sarpong, Yaw Adu-Sarkodie, Ulrich Ranft and Jurgen May. *Principal component analysis of socioeconomic factors and their association with malaria in children from the Ashanti Region, Ghana.* Malaria Journal, vol. 9, no. 1, page 201, 2010. (Cited on page 29.)

[Kremsner 1988] P.G. Kremsner, G.M. Zotter, H. Feldmeier, W. Graninger, R.M. Rocha and G. Wiedermann. *A comparative trial of three regimens for treating uncomplicated falciparum malaria in Acre, Brazil.* J Infect Dis, vol. 158, pages 1368–1371, 1988. (Cited on pages 36 and 37.)

[Krishna 2003] BV Krishna and AR Deshpande. *Comparison between conventional and QBC methods for diagnosis of malaria.* Indian J Pathol Microbiol, vol. 46, no. 3, pages 517–520, 2003. (Cited on page 23.)

[Kublin 2005] James G Kublin, Padmaja Patnaik, Charles S Jere, William C Miller, Irving F Hoffman, Nelson Chimbiya, Richard Pendame, Terrie E Taylor and Malcolm E Molyneux. *Effect of Plasmodium falciparum malaria on concentration of HIV-1-RNA in the blood of adults in rural Malawi: a prospective cohort study.* The Lancet, vol. 365, no. 9455, pages 233 – 240, 2005. (Cited on page 20.)

[Lambert 1992] Diane Lambert. *Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing.* Technometrics, vol. 34, no. 1, pages 1–14, 1992. (Cited on page 86.)

[Le Hesran 1997] Jean Yves Le Hesran, Michel Cot, Philippe Personne, Nadine Fievet, Béatrice Dubois, Mathilde Beyeme, Christian Boudin and Philippe Deloron. *Maternal Placental Infection with Plasmodium falciparum and Malaria Morbidity during the First 2 Years of Life.* American Journal of Epidemiology, vol. 146, no. 10, pages 826–831, 1997. (Cited on page 19.)

[Le Port 2011a] A. Le Port, L. Watier, G. Cottrell, S. Ouédraogo, C. Dechavanne and et al. *Infections in Infants during the First 12 Months of Life: Role of Placental Malaria and Environmental Factors.* PLoS ONE, vol. 6, no. 11, 2011. (Cited on pages 19 and 83.)

[Le Port 2011b] Agnès Le Port, Gilles Cottrell, Célia Dechavanne, Aziz Bouraima, Valérie Briand, José Guerra, Isabelle Choudat, Achille Massougbodji, Benjamin Fayomi, Florence Migot-Nabias, André Garcia and Michel Cot. *Prevention of malaria during pregnancy: assessing the effect of the distribution of IPTp through the national policy in Benin.* Am J Trop Med, vol. 82, no. 2, pages 270–275, 2011. (Cited on pages 36, 38 and 43.)

[Le Port 2012] Agnès Le Port, Gilles Cottrell, Yves Martin-Prevel, Florence Migot-Nabias, Michel Cot and André Garcia. *First malaria infections in a cohort of infants in Benin: biological, environmental and genetic determinants. Description of the study site, population methods and preliminary results.* BMJ Open, vol. 2, no. 2, 2012. (Cited on page 83.)

[Lee 1986] L.F. Lee. *Specification Test for Poisson Regression Models.* International Economic Review, vol. 27, page 689, 1986. (Cited on pages 83 and 85.)

[Lee 2012] J.-H. Lee, G. Han, W. J. Fulp and A. R. Giuliano. *Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) Study.* Epidemiology & Infection, vol. 140, pages 1087–1094, 6 2012. (Cited on page 97.)

[Liljander 2011] Anne Liljander, Philip Bejon, Jedidah Mwacharo, Oscar Kai, Edna Ogada, Norbert Peshu, Kevin Marsh and Anna Färnert. *Clearance of Asymptomatic P. falciparum Infections Interacts with the Number of Clones to Predict the Risk of Subsequent Malaria in Kenyan Children.* PLoS ONE, vol. 6, no. 2, page e16940, 02 2011. (Cited on pages 3 and 82.)

[Lu 1997] Wang-Shu. Lu. *Score tests for overdispersion in poisson regression models.* Journal of Statistical Computation and Simulation, vol. 56, no. 3, pages 213–228, 1997. (Cited on page 83.)

[MacDonald 1997] MacDonald and W. Zucchini. Hidden markov and other models for discrete-valued time series. Chapman & Hall, London, 1997. (Cited on pages 79 and 89.)

[Mathew 2010] Joseph L. Mathew. *Artemisinin derivatives Versus quinine for severe malaria in children: A systematic review and meta-analysis.* Indian Pediatrics, vol. 47, no. 5, pages 423–428, 2010. (Cited on page 25.)

[McCullagh 1989] P. McCullagh and J.A. Nelder. Generalized linear models, second edition. Taylor and Francis, 1989. (Cited on pages 82 and 97.)

[McCutchan 2008] TF McCutchan, RC Piper and MT Makler. *Use of malaria rapid diagnostic test to identify Plasmodium knowlesi infection.* Emerging Infectious Diseases, vol. 14, no. 5, pages 1750–1752, 2008. (Cited on page 22.)

[Melville 1910] C.H. Melville. *The prevention of malaria in war.* In Ronald Ross, editeur, The Prevention of Malaria by Ronald Ross, pages 577–599. New York: E.P. Dutton., 2d edition. édition, 1910. (Cited on pages 12 and 29.)

[Menéndez 2007] Clara Menéndez, Umberto D'Alessandro and Feiko O ter Kuile. *Reducing the burden of malaria in pregnancy by preventive strategies.* The Lancet Infectious Diseases, vol. 7, no. 2, pages 126–135, 2007. (Cited on page 26.)

[Mermin 2006] Jonathan Mermin, John R. Lule and John P. Ekwaru. *Association Between Malaria and CD4 Cell Count Decline Among Persons With HIV.* JAIDS Journal of Acquired Immune Deficiency Syndromes, vol. 41, no. 1 suppl, pages 129–130, 2006. (Cited on page 20.)

[Milet 2010] J. Milet, G. Nuel, L. Watier, D. Courtin, Y. Slaoui, P. Senghor, F. Migot-Nabias, O. Gaye and A. Garcia. *Genome wide linkage study, using a 250K SNP map, of Plasmodium falciparum infection and mild malaria attack in a Senegalese population.* PLoS One, vol. 5, no. 7, 2010. (Cited on pages 36 and 39.)

[Miller 1958] M. J. Miller. *Observations on the natura1 history of malaria in the semi-resistant West African.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 52, pages 152–168, 1958. (Cited on page 2.)

[Milne 1994] L.M. Milne, M.S. Kyi, P.L. Chiodini and D.C. Warhurst. *Accuracy of routine laboratory diagnosis of malaria in the United Kingdom.* Journal of clinical pathology, vol. 47, pages 740–742, 1994. (Cited on page 24.)

[Mitiku 2003] K. Mitiku, G. Mengistu and B. Gelaw. *The reliability of blood film examination for malaria at the peripheral health unit.* Ethiop.J.Health Dev, vol. 17, no. 2, pages 197–204, 2003. (Cited on page 31.)

[Molineaux 1980] L. Molineaux and C. Gramiccia. The Garki Project. Geneva: World Health Organization, 1980. (Cited on pages 37, 42 and 62.)

[Mulder 1998] B. Mulder, W. van der Ligt, R. Sauerwein and P. Verhave. *Detection of Plasmodium falciparum gametocytes with the QBC© test and Giemsa-stained thick blood films for malaria transmission studies in Cameroon.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 92, pages 395–396, 1998. (Cited on pages 39 and 62.)

[Mullahy 1986] John Mullahy. *Specification and testing of some modified count data models.* Journal of Econometrics, vol. 33, no. 3, pages 341–365, 1986. (Cited on pages 83 and 86.)

[Murray 2008] Clinton K. Murray, Robert A. Gasser, Alan J. Magill and R. Scott Miller. *Update on Rapid Diagnostic Testing for Malaria.* Clinical Microbiology Reviews, vol. 21, no. 1, pages 97–110, January 2008. (Cited on page 22.)

[Murray 2012] Christopher JL Murray, Lisa C Rosenfeld, Stephen S Lim, Kathryn G Andrews, Kyle J Foreman, Diana Haring, Nancy Fullman, Mohsen Naghavi, Rafael Lozano and Alan D Lopez. *Global malaria mortality between 1980 and 2010: a systematic analysis.* The Lancet, vol. 379, no. 9814, pages 413 – 431, 2012. (Cited on pages 2 and 15.)

[Mutabingwa 2005] T.K. Mutabingwa. *Artemisinin-based combination therapies (ACTs): Best hope for malaria treatment but inaccessible to the needy!* Acta Tropica, vol. 95, no. 3, pages 305 – 315, 2005. (Cited on page 25.)

[Mwangi 2005] Tabitha W Mwangi, Amanda Ross, Robert W Snow and Kevin Marsh. *Case Definitions of Clinical Malaria under Different Transmission Conditions in Kilifi District, Kenya.* Journal of Infectious Diseases, vol. 191, no. 11, pages 1932–1939, 2005. (Cited on pages 3, 82 and 98.)

[Nelder 1965] J. A. Nelder and R. Mead. *A simplex algorithm for function minimization.* Computer Journal, vol. 7, pages 308–313, 1965. (Cited on pages 71 and 87.)

[Olliaro 2011] Piero Olliaro, Abdoulaye Djimdé, Corine Karema, Andreas Märtensson, Jean-Louis Ndiaye, Sodiomon B. Sirima, Grant Dorsey and Julien Zwang. *Standardised versus actual white cell counts in estimating thick film parasitaemia in African children under five.* Tropical Medicine & International Health, vol. 16, no. 5, pages 551–554, 2011. (Cited on page 38.)

[Olotu 2013] Ally Olotu, Gregory Fegan, Juliana Wambua, George Nyangweso, Ken O. Awuondo, Amanda Leach, Marc Lievens, Didier Leboulleux, Patricia Njuguna, Norbert Peshu, Kevin Marsh and Philip Bejon. *Four-Year Efficacy of RTS,S/AS01E and Its Interaction with Malaria Exposure.* New England Journal of Medicine, vol. 368, no. 12, pages 1111–1120, 2013. (Cited on page 28.)

[Parrot 1950] L. Parrot and A. Catanei. *Les Cléments de mesure du réservoir de virus paludéen.* Archives de l'Institut Pasteur d'Algérie, vol. 28, pages 71–92, 1950. (Cited on page 2.)

[Patnaik 2005] Padmaja Patnaik, Charles S. Jere, William C. Miller, Irving F. Hoffman, Jack Wirima, Richard Pendame, Steven R. Meshnick, Terrie E. Taylor, Malcolm E. Molyneux and James G. Kublin. *Effects of HIV-1 Serostatus, HIV-1 RNA Concentration, and CD4 Cell Count on the Incidence of Malaria Infection in a Cohort of Adults in Rural Malawi.* Journal of Infectious Diseases, vol. 192, no. 6, pages 984–991, 2005. (Cited on page 20.)

[Patterson 2009] Toby A. Patterson, Marinelle Basson, Mark V. Bravington and John S. Gunn. *Classifying movement behaviour in relation to environmental conditions using hidden Markov models.* Journal of Animal Ecology, vol. 78, pages 1113–1123, 2009. (Cited on page 89.)

[Payne 1988] D Payne. *Use and limitations of light microscopy for diagnosing malaria at the primary health care level.* Bull World Health Organ, vol. 66, pages 621–626, 1988. (Cited on pages 39 and 62.)

[Petersen 1996a] E Petersen, N T Marbiah, L New and A Gottschau. *Comparison of two methods for enumerating malaria parasites in thick blood films.* Am J Trop Med Hyg, vol. 55, no. 5, pages 485–489, 1996. (Cited on pages 3, 4, 36, 45, 82 and 83.)

[Petersen 1996b] E. Petersen, N.T. Marbiah, L. New and A. Gottschau. *Comparison of Two Methods for Enumerating Malaria Parasites in Thick Blood Films.* Am J Trop Med Hyg, vol. 55, pages 485–489, 1996. (Cited on page 3.)

[Piegorsch 1990] Walter W. Piegorsch. *Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter.* Biometrics, vol. 46, no. 3, pages pp. 863–867, 1990. (Cited on page 83.)

[Pinto 2001] M. Pinto, M. Verenkar, R. Desouza and S. Rodrigues. *Usefulness of quantitative buffy coat blood parasite detection system in diagnosis of malaria.* Indian Journal of Medical Microbiology, vol. 19, no. 4, pages 219–221, 2001. (Cited on page 23.)

[Planche 2001] T. Planche, S. Krishna, M. Kombila, K. Engel, J. F. Faucher, E. Ngou-Milama and P. G. Kremsner. *Comparison of methods for the rapid laboratory assessment of children with malaria.* Am. J. Trop. Med. Hyg., vol. 65, no. 5, pages 599–602, 2001. (Cited on pages 3, 36 and 37.)

[Poskitt 2005] D.S. Poskitt and Jing Zhang. *Estimating components in finite mixtures and hidden Markov models.* Australian & New Zealand Journal of Statistics, vol. 47, no. 3, pages 269–286, 2005. (Cited on page 79.)

[PrayGod 2008] George PrayGod, Albie de Frey and Michael Eisenhut. *Artemisinin derivatives versus quinine in treating severe malaria in children: a systematic review.* Malaria Journal, vol. 7, no. 1, page 210, 2008. (Cited on page 25.)

[Prudhomme O'Meara 2005] W Prudhomme O'Meara, FE McKenzie, AJ Magill, JR Forney, B Permpanich, C Lucas, RA Gasser and C Wongsrichanalai. *Sources of variability in determining malaria parasite density by microscopy.* Am J Trop Med Hyg, vol. 73, pages 593–598, 2005. (Cited on pages 39 and 62.)

[Prudhomme O'Meara 2006a] Wendy Prudhomme O'Meara, Mazie Barcus, Chansuda Wongsrichanalai, Sinuon Muth, Jason Maguire, Robert Jordan, William Prescott and F Ellis McKenzie. *Reader technique as a source of variability in determining malaria parasite density by microscopy.* Malaria Journal, vol. 5, no. 1, page 118, 2006. (Cited on pages 36, 38, 39 and 42.)

[Prudhomme O'Meara 2006b] Wendy Prudhomme O'Meara, Shon Remich, Bernhards Ogutu, Martin Lucas, Ramadan Mtalib, Peter Obare, Frederick Oloo, Caroline Onoka, Joseph Osoga, Colin Ohrt and F. McKenzie. *Systematic comparison of two methods to measure parasite density from malaria blood smears.* Parasitology Research, vol. 99, pages 500–504, 2006. (Cited on pages 3 and 62.)

[Rab 2001] M. A. Rab, T. W. Freeman, N. Durrani, D. De Poerck and M. W. Rowland. *Resistance of Plasmodium falciparum malaria to chloroquine is widespread in eastern Afghanistan.* Annals of Tropical Medicine and Parasitology, vol. 95, no. 1, pages 41–46, 2001. (Cited on page 14.)

[Rabiner 1989] Lawrence R. Rabiner. *A tutorial on hidden markov models and selected applications in speech recognition.* In Proceedings of the IEEE, pages 257–286, 1989. (Cited on page 86.)

[Raghavan 1966] K. Raghavan. *Statistical considerations in the microscopical diagnosis of Malaria, with special reference to the role of cross-checking.* Bulletin of the World Health Organization, vol. 34, pages 788–791, 1966. (Cited on pages 3 and 82.)

[Raghavendra 2011] Kamaraju Raghavendra, TapanK. Barik, B.P.Niranjan Reddy, Poonam Sharma and AdityaP. Dash. *Malaria vector control: from past to future.* Parasitology Research, vol. 108, no. 4, pages 757–779, 2011. (Cited on page 27.)

[Rao 1956] C.R. Rao and I.M. Chakravarti. *Some small sample tests of significance for a Poisson distribution.* Biometrics, vol. 12, pages 264–282, 1956. (Cited on page 87.)

[Reyburn 2004] H. Reyburn, R. Mbatia, C. Drakeley, I. Carneiro, E. Mwakasungula, O. Mwerinde, K. Saganda, J. Shao, A. Kitua, R. Olomi, B.M. Greenwood and C.J. Whitty. *Overdiagnosis of malaria in patients with severe febrile illness in Tanzania: a prospective study.* BMJ, vol. 329, pages 1212–1217, 2004. (Cited on pages 2 and 36.)

[Reyburn 2007] Hugh Reyburn, Hilda Mbakilwa, Rose Mwangi, Ombeni Mwerinde, Raimos Olomi, Chris Drakeley and Christopher J M Whitty. *Rapid diagnostic tests compared with malaria microscopy for guiding outpatient treatment of febrile illness in Tanzania: randomised trial.* BMJ, vol. 334, page 403, 2 2007. (Cited on page 84.)

[Ricci 2012] Francesco Ricci. *Social Implications of Malaria and Their Relationships with Poverty.* Mediterr J Hematol Infect Dis, vol. 4, no. 1, pages 337–346, 2012. (Cited on page 29.)

[Robert 2000] C. P. Robert, T. Rydén and D. M. Titterington. *Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method.* Journal of the Royal Statistical Society Series B, vol. 62, no. 1, pages 57–75, 2000. (Cited on page 87.)

[Rogier 1993] C. Rogier and J. F. Trape. *Malaria attacks in children exposed to high transmission: who is protected?* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 87, pages 245–246, 1993. (Cited on page 36.)

[Rogier 1996] C. Rogier, D. Commenges and J.F. Trape. *Evidence for an age-dependent pyrogenic threshold of Plasmodium falciparum parasitemia in highly endemic populations.* Am J Trop Med Hyg, vol. 54, pages 613–619, 1996. (Cited on page 36.)

[Roshanravan 2003] B. Roshanravan, E. Kari, R.H. Gilman, L. Cabrera, E. Lee, J. Metcalfe, M. Calderon, A.G. Lescano, S.H. Montenegro, C. Calampa and J.M. Vinetz. *Endemic malaria in the Peruvian Amazon region of Iquitos.* Am J Trop Med Hyg, vol. 69, pages 45–52, 2003. (Cited on page 24.)

[Ross 1910] R. Ross and D. Thompson. *Some enumerative studies on malaria fever.* Ann Trop Med Parasitol, vol. 4, pages 267–313, 1910. (Cited on page 3.)

[Roucher 2012] Clémentine Roucher, Christophe Rogier, Fambaye Dieye-Ba, Cheikh Sokhna, Adama Tall and Jean-François Trape. *Changing Malaria Epidemiology and Diagnostic Criteria for Plasmodium falciparum Clinical Malaria.* PLoS ONE, vol. 7, page e46188, 09 2012. (Cited on page 98.)

[Rougemont 1991] A. Rougemont, N. Breslow, E. Brenner and et al. *Epidemiological basis for clinical diagnosis of childhood malaria in endemic zone in West Africa.* Lancet, vol. 338, pages 1292–1295, 1991. (Cited on page 38.)

[Russel 1963] P. F. Russel, S. L. West, R. D. Manwell and G. Macdonald. Practical Malariology. London : Oxford University Press, 1963. (Cited on page 3.)

[Rydén 1995] Tobias Rydén. *Estimating the Order of Hidden Markov Models.* Statistics, vol. 26, no. 4, pages 345–354, 1995. (Cited on page 79.)

[Rydén 2008] Tobias Rydén. *EM versus Markov chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective.* Bayesian Analysis, 2008. (Cited on page 87.)

[Saha 2009] Krishna K. Saha and Roger Bilisoly. *Testing the homogeneity of the means of several groups of count data in the presence of unequal dispersions.* Computational Statistics & Data Analysis, vol. 53, no. 9, pages 3305–3313, July 2009. (Cited on page 96.)

[Sakuntabhai 2008] A. Sakuntabhai, R. Ndiaye, I. Casademont, C. Peerapittayamongkol and C. Rogier. *Genetic determination and linkage mapping of Plasmodium falciparum malaria related traits in Senegal.* Microbes and Infections, vol. 3, page 2000, 2008. (Cited on page 36.)

[Sarkar 2010] Pralay K. Sarkar, Gautam Ahluwalia, Vannan K. Vijayan and Arunabh Talwar. *Critical Care Aspects of Malaria.* Journal of Intensive Care Medicine, vol. 25, no. 2, pages 93–103, 2010. (Cited on page 25.)

[Schnabel 1985] Robert B. Schnabel, John E. Koonatz and Barry E. Weiss. *A modular system of algorithms for unconstrained minimization.* ACM Trans. Math. Softw., vol. 11, no. 4, pages 419–440, December 1985. (Cited on page 88.)

[Schwarz 1978] Gideon Schwarz. *Estimating the Dimension of a Model.* The Annals of Statistics, vol. 6, no. 2, pages 461–464, 1978. (Cited on page 89.)

[Schwetz 1941] J. Schwetz. *Recherches sur le paludisme dans les villages et les camps de la division de Mongbwalu des mines d'or de Kilo (Congo Belge).* Mémoires de l'Institut Royal Colonial Belge, vol. 11, page 2, 1941. (Cited on page 2.)

[Scott 1955] W. Scott. *Reliability of content analysis: The case of nominal scale coding.* Public Opinion Quarterly, vol. 17, pages 321–325, 1955. (Cited on page 62.)

[Selby 1965] B. Selby. *The index of Dispersion as a Test Statistic.* Biometrika, vol. 52, page 627, 1965. (Cited on page 82.)

[Sergent 1935] E. Sergent and L. Parrot. *L'immunité, la prémunition et la résistance innée.* Archives Institut Pasteur Algérie, vol. 13, pages 279–319, 1935. (Cited on page 20.)

[Sharma 1996] VP. Sharma. *Re-emergence of malaria in India.* Indian J Med Res, vol. 103, pages 26–45, 1996. (Cited on page 14.)

[Shaw 1995] D. J. Shaw and A. P. Dobson. *Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review.* Parasitology, vol. 111, pages S111–S133, 1995. (Cited on page 96.)

[Sidhu 2002] A.B. Sidhu, D. Verdier-Pinard and D.A. Fidock. *Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations.* Science, vol. 298, no. 5591, pages 210–213, 2002. (Cited on page 28.)

[Sinton 1924] J. A. Sinton. *Methods for the enumeration of parasites and leucocytes in the blood of malarial patients.* Indian Journal of Medical Research, vol. 12, pages 341–346, 1924. (Cited on page 3.)

[Small 2010] Dylan S. Small, Jing Cheng and Thomas R. Ten Have. *valuating the Efficacy of a Malaria Vaccine.* The International Journal of Biostatistics, vol. 6, no. 2, 2010. (Cited on page 36.)

[Smith 1994] T. Smith, JA. Schellenberg and R. Hayes. *Attributable fraction estimates and case definitions for malaria in endemic areas.* Statistics in Medicine, vol. 13, pages 2345–2358, 1994. (Cited on page 98.)

[Smith 2007] David Smith, Carlos Guerra, Robert Snow and Simon Hay. *Standardizing estimates of the Plasmodium falciparum parasite rate.* Malaria Journal, vol. 6, no. 1, page 131, 2007. (Cited on page 17.)

[Student 1907] Student. *On the error of counting with a haemacytometer.* Biometrika, vol. 5, pages 351–360, 1907. (Cited on pages 3 and 82.)

[The RTS S Clinical Trials Partnership 2011] . The RTS S Clinical Trials Partnership. *First Results of Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Children.* New England Journal of Medicine, vol. 365, no. 20, 2011. (Cited on page 27.)

[Thomson 1911] D. Thomson. *A new blood-counting pipette for estimating the number of leucocytes and blood parasites per cubic millimeter.* Annals of Tropical Medicine and Parasitology, vol. 5, pages 471–478, 1911. (Cited on page 3.)

[Timmann 2007] Christian Timmann, Jennifer A Evans, Inke R König, André Kleensang, Franz Rüschendorf, Julia Lenzen, Jürgen Sievertsen, Christian Becker, Yeetey Enuameh, Kingsley Osei Kwakye, Ernest Opoku, Edmund N. L Browne, Andreas Ziegler, Peter Nürnberg and Rolf D Horstmann. *Genome-Wide Linkage Analysis of Malaria Infection Intensity and Mild Disease.* PLoS Genet, vol. 3, page e48, 03 2007. (Cited on page 36.)

[Tiono 2009] Alfred Tiono, Alphonse Ouedraogo, Edith Bougouma, Amidou Diarra, Amadou Konate, Issa Nebie and Sodiomon Sirima. *Placental malaria and low birth weight in pregnant women living in a rural area of Burkina Faso following the use of three preventive treatment regimens.* Malaria Journal, vol. 8, no. 1, page 224, 2009. (Cited on page 19.)

[Trape 1985] J. F. Trape. *Rapid evaluation of malaria parasite density and standardization of thick smear examination for epidemiological investigations.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 79, pages 181–184, 1985. (Cited on pages 3, 21, 36, 37, 38, 39, 42 and 62.)

[van den Berg 2009] Henk van den Berg. *Global Status of DDT and Its Alternatives for Use in Vector Control to Prevent Disease.* Environmental Health Perspectives, vol. 117, no. 11, pages 1656–1663, 2009. (Cited on page 26.)

[W. 1994] Greene W. *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.* In W. Working Paper EC- 94-10. Department of Economics, New York University, 1994. (Cited on page 86.)

[Wang 1996] Peiming Wang, Martin L. Puterman, Iain Cockburn and Nhu Le. *Mixed Poisson Regression Models with Covariate Dependent Rates.* Biometrics, vol. 52, pages 381–400, 1996. (Cited on page 97.)

[Wangai 2011] L.N. Wangai, F.T. Kimani, S.A. Omar, S.M. Karanja, D.W. Nderu, G. Magoma and D. Mutua. *Chloroquine resistance status a decade after : Re-emergence of sensitive Plasmodium falciparum strains in malaria endemic and epidemic areas in Kenya.* The Journal of protozoology research, vol. 21, no. 1, pages 20–29, 2011. (Cited on page 14.)

[Warhurst 1996] DC Warhurst and JE Williams. *ACP Broadsheet no 148. July 1996. Laboratory diagnosis of malaria.* J Clin Pathol, vol. 49, pages 533–538, 1996. (Cited on pages 3, 22, 38 and 42.)

[Warrell 2002] DM. Warrell and HM. Gilles. Essential malariology. (eds). Arnold, London, 2002. (Cited on pages 37 and 84.)

[Waters 2012] N.C. Waters and M.D. Edstein. *8-Aminoquinolines: primaquine and tafenoquine.* In H.M. Staines and S. Krishna, editeurs, Treatment and Prevention of Malaria: Antimalarial Drug Chemistry, Action and Use, Milestones in Drug Therapy, pages 69–93. Springer Basel, 2012. (Cited on page 25.)

[White 1999a] N White. *Antimalarial drug resistance and combination chemotherapy.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, vol. 354, pages 739–749, 1999. (Cited on pages 12 and 25.)

[White 1999b] N J White, F Nosten, S Looareesuwan, W M Watkins, K Marsh, R W Snow, G Kokwaro, J Ouma, T T Hien, M E Molyneux, T E Taylor, C I Newbold, T K Ruebush, M Danis, B M Greenwood, R M Anderson and P Olliaro. *Averting a malaria disaster.* Lancet, vol. 353, no. 9168, pages 1965–1967, 1999. (Cited on page 25.)

[WHO 1959] WHO. *The world health organization program to rid the world of malaria.* The American Journal of Nursing, vol. 59, pages 1402–4, 1959. (Cited on pages 1 and 13.)

[WHO 1961] WHO. Manual for processing and examination of blood slides in malaria eradication programmes. WHORA1262-61, 1961. (Cited on page 37.)

[WHO 1963] WHO. *Terminology of malaria and of malaria eradication.* World Health Organization, 1963. (Cited on page 38.)

[WHO 1999] WHO. *New perspectives. Malaria Diagnosis. Report of a Joint WHO/USAID Informal Consultation 25-27 October 1999.* Geneva: World Health Organization, 1999. (Cited on page 33.)

[WHO 2000] WHO. *Spatial targeting of interventions against malaria.* Bull World Hlth Organ, 2000. (Cited on page 36.)

[WHO 2007a] WHO. *Policy and Procedures of the WHO/NICD Microbiology External Quality Assessment Programme in Africa.* World Health Organization. Geneva, 2007. (Cited on page 39.)

[WHO 2007b] WHO. *Report of the technical expert group meeting on intermittent preventive treatment in pregnancy (IPTp).* World Health Organization. Geneva, 2007. (Cited on page 26.)

[WHO 2010a] WHO. *Basic Malaria Microscopy: Part I. Learner's Guide, Second Edition.* World Health Organization, April 2010. (Cited on pages 25, 27, 35, 36, 38, 42, 49 and 84.)

[WHO 2010b] WHO. Guidelines for the treatment of malaria. Non-serial Publication. World Health Organization, 2010. (Cited on page 25.)

[WHO 2011a] WHO. *Malaria Rapid Diagnostic Test Performance: Summary results of WHO Malaria RDT Product Testing: Rounds 1-3 (2008-2011). Geneva.* World Health Organization, 2011. (Cited on page 23.)

[WHO 2011b] WHO. *World Malaria Report 2011.* World Health Organization, 2011. (Cited on pages 1, 2 and 15.)

[WHO 2012] WHO. *World Malaria Report 2012.* World Health Organization, 2012. (Cited on pages 2 and 15.)

[Wilson 1936] D. B. Wilson. *Rural hyper-endemic malaria in Tanganyika territory.* Transactions of the Royal Society of Tropical Medicine and Hygiene, vol. 29, pages 583–618, 1936. (Cited on page 2.)

[Wilson 1950] D. B. Wilson, P. C. C. Garnham and N. H. Swellengrebel. *A review of hyperendemic malaria.* Tropical Diseases Bulletin, vol. 47, pages 677–698, 1950. (Cited on page 2.)

[Winkelmann 1991] Rainer Winkelmann and Klaus F. Zimmermann. *A new approach for modeling economic count data.* Economics Letters, vol. 37, no. 2, pages 139–143, 1991. (Cited on page 83.)

[Winkelmann 1995] Rainer Winkelmann and Klaus F Zimmermann. *Recent Developments in Count Data Modelling: Theory and Application.* Journal of Economic Surveys, vol. 9, no. 1, pages 1–24, March 1995. (Cited on page 96.)

[Winkelmann 2003] Rainer Winkelmann. Econometric analysis of count data. 4th rev. ed. Springer, 2003. (Cited on page 83.)

[Wintrobe 1967] M. Wintrobe. Clinical Hematology, volume 30. Philadelphia: Lea and Febiger, 6th édition, 1967. (Cited on pages 3 and 38.)

[Yau 2003] Kelvin K. W. Yau, Kui Wang and Andy H. Lee. *Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros.* Biometrical Journal, vol. 45, no. 4, pages 437–452, 2003. (Cited on page 83.)

[Zorn 1996] Christopher Zorn. *Evaluating Zero-inflated and Hurdle Poisson Specifications.* In JSAI Workshops, 1996. (Cited on page 85.)

[Zucchini 2000] W. Zucchini. *An Introduction to Model Selection.* J Math Psychol, vol. 44, no. 1, pages 41–61, 2000. (Cited on page 79.)

[Zucchini 2009] W. Zucchini and I.L. MacDonald. Hidden markov models for time series: An introduction using r. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2009. (Cited on page 65.)

[Zurovac 2006] D. Zurovac, B. Midia, S.A. Ochola, M. English and R.W. Snow. *Microscopy and outpatient malaria case management among older children and adults in Kenya*. Trop Med Int Health, vol. 11, pages 1–9, 2006. (Cited on pages 2 and 36.)

Statistical properties of parasite density estimators in malaria and field applications

Abstract: Malaria is a devastating global health problem that affected 219 million people and caused 660,000 deaths in 2010. Inaccurate estimation of the level of infection may have adverse clinical and therapeutic implications for patients, and for epidemiological endpoint measurements. The level of infection, expressed as the parasite density (PD), is classically defined as the number of asexual parasites relative to a microliter of blood. Microscopy of Giemsa-stained thick blood smears (TBSs) is the gold standard for parasite enumeration. Parasites are counted in a predetermined number of high-power fields (HPFs) or against a fixed number of leukocytes. PD estimation methods usually involve threshold values; either the number of leukocytes counted or the number of HPFs read. Most of these methods assume that (1) the distribution of the thickness of the TBS, and hence the distribution of parasites and leukocytes within the TBS, is homogeneous; and that (2) parasites and leukocytes are evenly distributed in TBSs, and thus can be modeled through a Poisson-distribution. The violation of these assumptions commonly results in overdispersion. Firstly, we studied the statistical properties (mean error, coefficient of variation, false negative rates) of PD estimators of commonly used threshold-based counting techniques and assessed the influence of the thresholds on the cost-effectiveness of these methods. Secondly, we constituted and published the first dataset on parasite and leukocyte counts per HPF. Two sources of overdispersion in data were investigated: latent heterogeneity and spatial dependence. We accounted for unobserved heterogeneity in data by considering more flexible models that allow for overdispersion. Of particular interest were the negative binomial model (NB) and mixture models. The dependent structure in data was modeled with hidden Markov models (HMMs). We found evidence that assumptions (1) and (2) are inconsistent with parasite and leukocyte distributions. The NB-HMM is the closest model to the unknown distribution that generates the data. Finally, we devised a reduced reading procedure of the PD that aims to a better operational optimization and a practical assessing of the heterogeneity in the distribution of parasites and leukocytes in TBSs. A patent application process has been launched and a prototype development of the counter is in process.

Keywords: Malaria epidemiology, threshold-based counting techniques, parasite density estimators, mean error, coefficient of variation, false-negative rates, cost-effectiveness, parasite and leukocyte counts per high-power field, Poisson distribution, overdispersion, heterogeneity, negative binomial distribution, mixture models, HMMs, patent.