

**Analytical frameworks for studies prone to misclassification bias to identify
correlates of disease protection**

Case study: Immunity to malaria



By

KENNEDY MWAI WAMBUI

A Thesis Submitted to the Faculty of Health Sciences,

The University of the Witwatersrand

Johannesburg, South Africa

in fulfilment of the requirements for the

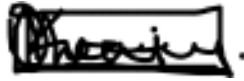
Degree of

DOCTOR OF PHILOSOPHY

7 December 2021

DECLARATION

I, **Kennedy Mwai Wambui** declare that this PhD thesis is my own my own, unaided work, except where specific references cited to the work of others. It is being submitted for the Degree of Doctor of Philosophy degree at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

Date: 06th day of April 2021

DEDICATION

This research is dedicated to my loving mum Wambui Mwai for the prayers and support during my academic and research career.

Kũgũrũ nĩ ĩrata thĩ (It is perseverance that prevails).

To *Jehova Ngai*, all is unto you.

Thesis Abstract

Background:

The process of identifying statistical associations between important biological features is challenging especially for outcomes prone to misclassification and significant biological complexity. Additionally, experimental conditions may introduce systematic and non-systematic bias when measuring biological features.

These misclassification and systematics errors if not handled can lead to bias in describing, framing, selecting, and comparing cases and controls. Consequently, researchers across the globe, are facing challenges in making data-driven decisions and recommendations. This is particularly relevant for studies on natural immunity to malaria given the approaches used to categorize the non-case and clinical malaria cases in studies aimed at identifying correlates of immunity against malaria.

The primary aim of this study was to develop analytical approaches that are sensitive to the challenge that the outcome and biological features are prone to differential misclassification and systematic bias respectively, then identify the correlates of protection against disease.

Methodology

This project had three major steps to achieve the overall objective. First, we used methodologies to reduce bias in malaria case definition, then developed approaches to handle bias in high throughput protein micro array data. Finally, we utilised machine learning approaches to select correlates of malaria protection.

To reduce bias in clinical data, we utilised logistic regression and Bayesian approaches adjusting for the repeated measures. To do this, we used longitudinal data from Kilifi, Kenya to investigate whether the varying intensity of malaria transmission had a significant impact on parasite density thresholds and case definition. Further we used the same dataset to explore an alternative statistical approach of using the probability of developing fevers in defining cases rather than discrete threshold cut-offs as the former has been reported to increase predictive power.

To handle the bias associated with biological measurements, we developed and compared methods of correcting systematic and non-systematic biases in high throughput laboratory data. Additionally, we created an R package suite for dealing with handling the systematic and non-systematic biases in protein micro-array data. Lastly, we compared the performance of Random Forest, GPBoosting and Bayesian with spike and slab prior machine learning approaches to in identifying merozoite targets of protective immunity against *Plasmodium falciparum* malaria using multi-centre cohort studies data. Random forest approach has been reported to handle overfitting bias through bagging. GPBoosting utilizes tree-based approach with Gaussian Process and Mixed Effects Models this adjusting for the site random effect. The Bayesian with spike and slab prior performs well in data prone to multicollinearity and produces readily available uncertainty estimates.

This final part covers two main themes, the first theme involves using dichotomous outcome and secondly, utilizing probability of having fever given a certain parasite density as the outcome in feature selection using Kilifi data only. A network model was used to

help visualize complex associations and linkages among different sites and the identified features.

Results.

To perform case definition, malaria attributable fractions were estimated using logistic power and Bayesian latent class approaches. Both approaches estimated similar patterns of fevers attributable to malaria with changing transmission intensities. The former performed well in estimating the probabilities of having fever, while the latter was efficient in determining the parasite density threshold. However, compared to the logistic power model, the Bayesian algorithm estimates were less than the former for both malaria attributable fractions and probabilities of fever. The Bayesian latent class gave a lower MAF estimate, Bland-Altman bias = 0.20 (0.16-0.24), compared to the logistic model.

After correcting (or accounting for) the bias associated with the outcome, we develop a generic one-stop-shop pre-processing suite for protein microarrays that is compatible with data from the major protein microarray scanners. The suite incorporated graphical and tabular interfaces to facilitate a detailed inspection of data and is coupled with supporting guidelines that enable users to select the most appropriate algorithms to systematically address bias arising in customized experiments. From the various machine learning approaches, proteins PF113, PF3D7_0525800(IMC1g), AMA1, MSP11, PF3D7_1136200 were identified as the features most highly associated with immunity to malaria. In addition, the utilization of probability outcome had a better statistical fit compared to utilizing a dichotomous outcome in the subset analysis that was done for data from a high transmission area of Kilifi.

Discussion.

Logistic regression and Bayesian approaches to handle misclassification bias gave similar estimates of malaria attributable fractions used for case definition. In addition, we observe that utilizing probabilities as a way of handling misclassification bias in malaria outcome had a better fit compared to using dichotomous variable. However, a training sample is required to utilize the Bayesian latent class models. Finally, we developed an R-package to handle systematic and non-systematic biases associated with high throughput assays such as protein micro-arrays. Specifically, we develop functions to correct for background noise, within-sample variation, mean-variance dependence, and batch correction. We also included a user-friendly interactive Shiny web-based platform thus eliminating the need for prowess in programming.

Combining the data from case definition and protein micro-array and utilizing machine learning approaches we identified previously described and novel antigens (IMC1g and PF3D7_0629500) that are potential targets for naturally acquired immunity against malaria. However, these and other antigens with a protective effect would be recommended for further functional assays to inform vaccine development.

The probability outcome-based model had a better statistical fit than the binary model helping in reducing the misclassification bias. Nevertheless, in our study, there was a limitation of lack of training samples for most of the multicentre cohorts thus we only tested the model on data from only one site in Kilifi (Junju, Kilifi cohort).

To the best of our knowledge this is the largest malaria multicentre study to compare different methodologies for identifying important correlates of immunity to malaria and to demonstrate that probability-based model is better than a binary outcome model in identifying important features. We propose a workflow of ensuring disease is defined without bias, handle systematic and non-systematic laboratory data bias and perform feature selection.

Preface

This thesis is submitted to the University of the Witwatersrand, Faculty of Health Sciences for the fulfilment of the requirements for the degree of Doctor of Philosophy by publication in the major area-subject of Biostatistics and Epidemiology. The doctoral research has been undertaken at the School of Public Health in the Department of Biostatistics and Epidemiology. My supervisors were Professor Faith Osier, Professor Samson Kinyanjui and Professor Eustasius Musenge.

This doctoral work was supported by the National Institute for Health Research, using Official Development Assistance (ODA) funding 16/136/33, Tackling Infections to Benefit Africa (TIBA). Further support from the DELTAS Africa Initiative [DEL-15-003], The Initiative to Develop African Research Leaders (IDeAL), based at the KEMRI – Wellcome Trust Research Programme in Kenya.

Presentations arising from this study

- 1) Analytical approaches for selecting correlates of malaria protection, TIBA virtual Postgraduate Seminar Day - Tuesday 12 October 2021
- 2) A comparison of data-normalisation approaches for one-color protein microarray data: *Plasmodium falciparum* Merozoite protein microarray. *2020 DELTAS SSACAB Virtual Conference* oral presentation.
- 3) protGear: A tool for protein micro-array data pre-processing. *Symposium on Data Science and Statistics Bellevue, Washington, May 29 – June 1, 2019: Beyond Big Data: Building Data Tools*, oral presentation.
- 4) Integrated web tool for protein micro array data management and normalization. Kennedy Mwai, Nelson Kibinge PhD, James Tuju PhD, Samson Kinyanjui PhD, Eustasius Musenge, Faith Osier. *Second TIBA Annual General Meeting, Accra Ghana, September 16-18, 2019*, poster presentation. (Awarded the second-best Poster prize).
- 5) protGear: A tool for protein micro-array data pre-processing. Kennedy Mwai, Nelson Kibinge PhD, Samson Kinyanjui PhD, David Moinina Sengeh PhD, Eustasius Musenge, Faith Osier. *IDeAL Midterm Conference and KEMRI – Wellcome's Decade*

Celebration of Research Capacity Building 20th to 22nd June 2018, poster presentation. (Awarded the Best Poster prize)

- 6) protGear: A tool for protein micro-array data pre-processing. Kennedy Mwai, Nelson Kibinge PhD, Samson Kinyanjui PhD, David Moinina Sengeh PhD, Ass. Prof Eustasius Musenge, Prof. Faith Osier. *First Tiba Annual General Meeting, Durban, South Africa, 29-31 May 2018*, poster presentation. (Awarded the Best Poster prize)

Publications arising from this study

My involvement in the PhD research work includes being the first and co-author in the following publications and tool

1. Estimating Malaria Attributable Fractions With Changing Transmission Intensity: Bayesian Latent Class Vs Logistic Models. (First Author)
doi: <https://doi.org/10.21203/rs.3.rs-916409/v1>
2. protGear: A protein microarray data pre-processing suite. (First Author)
doi: <https://doi.org/10.1016/j.csbj.2021.04.044>
3. protGear R package: Bioconductor submission in process. (Package Maintainer)
Package repository: <https://github.com/Keniajin/protGear/> and <https://keniajin.github.io/protGear/>
4. Defining the merozoite targets of protective immunity against Plasmodium falciparum malaria through multi-center cohort studies. (Shared first Author)
Paper under interactive review with co-authors.
5. Distinct kinetics in antibody responses to 111 Plasmodium falciparum antigens identifies novel serological markers of recent malaria exposure. (Co-Author)
doi: <https://doi.org/10.1101/2020.12.14.20242768>

Acknowledgements

Ūgĩ wa mūdũ ūmwe ndūrĩmaga (A single man's ability cannot till all fields.)

The success of this work is owed to different people who have contributed either directly or indirectly. I greatly appreciate my supervisors, Professor Faith Osier, Professor Eustasius Musenge and Professor Sam K. Muchina for their professional guidance in shaping the research objectives and mentorship during my PhD journey. I am grateful to Dr Nelson Kibinge and Dr James Tuju for their invaluable advice and feedback towards addressing the statistical, immunological, and epidemiological implications of the research findings. I also thank my third-party monitor Professor George Warimwe for his encouragement and advise. A special thank you to The Initiative to Develop African Research Leaders (IDeAL) program, KEMRI-Wellcome Trust Kilifi, Centre for Infectious Diseases, Parasitology Unit, Heidelberg and School of Public Health, University of the Witwatersrand for hosting me and providing the resources needed to undertake this PhD training.

I extend special thanks to the South-South Malaria Antigen Research Partnership (SMART) led by Prof. Faith Osier for permitting me to use their data that helped me achieve my set objectives in line with my doctoral work. I would also like to acknowledge the KEMRI-Wellcome Trust, IT department for providing server platform to process and run my models. Special thanks to Dr Gathoni Kamuyu, Rinter Kimathi, Timothy Chege, Lydia Nyamako, Emily Chepsat, Patience Kiyuka and the entire FO group laboratory team that was involved in the protein microarray experiments and data production. Thanks to Barrack Angujo and Lucian Mshote for creating the data warehouse. I extend my

appreciation to Dr Irene Nailain, David Odera, Dr Roland Frank for their support and input especially while I was based at the Heidelberg University Hospital. I would also like to thank the IDeAL department, specifically, Liz Murabu, David Ngui, Florence Kirimi and Rita Baya for their administrative support during my PhD. I would also like to acknowledge Dr Innocent Maposa and Dr John Bradley for their inputs during my interim seminar. I also appreciate Dr Alice Kamau and James Mburu for their statistical advice and inputs. Thanks to Jacqueline Waeni for the support to traceback and debug errors for the protGear package. I extend my gratitude to the Witwatersrand Biostatistics Unit, especially Professor Jonathan Levin and Professor Tobias Chirwa for their encouraging words. I would like to thank TIBA and IDeAL for providing me with a fellowship that supported me during the entire period of study.

Many friends and colleagues at the University of the Witwatersrand and KEMRI-Wellcome trust were equally encouraging and urged me on. They include George Kinyanjui, Maria Njambi, Rodney Ogwang, Charles Kolombo, Diana Nyabundi and Koko Mutai. Special thanks to my host department, KWTRP-IT especially Denis Kimurgor, Lucas Wambua, Denis Kiboi and Anne Wairimu.

Immense gratitude, Kinya Kirima for always cheering me on! Thank you for always offering the much-needed support and love. Finally, I would also like to say a heartfelt thank you to my Mum, Wambui Mwai, for always believing in me and encouraging me to follow my dreams.

Table of Contents

DECLARATION	ii
DEDICATION	iii
Thesis Abstract.....	iv
Preface	ix
Presentations arising from this study	x
Publications arising from this study.....	xii
Acknowledgements.....	xiii
List of Figures	xx
List of Tables	xxi
Nomenclature/List of Abbreviations and Symbols	ii
CHAPTER 1 : INTRODUCTION Equation Chapter 2 Section 1	1
1.1 Overview	1
1.1.1 Malaria background	1
1.1.2 Malaria case definition.....	5
1.1.3 Identifying correlates of disease protection	7
1.1.4 Research problem and justification: What is novel?	13
1.1.5 Thesis research questions	15
1.1.6 Thesis broad and specific objectives.....	16
1.2 Literature Review.....	17
1.2.1 Evolution of case definition approaches in malaria.....	17
1.2.2 Protein microarray data pre-processing tools	20
1.2.3 Machine learning approaches and correlates of malaria protection.	21
CHAPTER 2 : METHODOLOGY	25
2.1 Study population and design	25
2.2 Epidemiological data.....	25
2.2.1 Kombewa, Kenya.....	27
2.2.2 Nyamisati, Tanzania	27
2.2.3 Muheza, Tanzania	28
2.2.4 Kilifi, Kenya.....	28
2.2.5 Bandiagara, Mali	29
2.2.6 Sotuba, Mali	29
2.2.7 Samako, Mali.....	29

2.2.8	Dielmo, Senegal	30
2.2.9	Ndiop, Senegal	30
2.2.10	Kintampo, Ghana	30
2.2.11	Balonghin, Burkina Faso	31
2.2.12	Jinja, Kanungu and Tororo, Uganda	31
2.3	Laboratory data.....	32
2.4	Case Definition methods.....	33
2.4.1	Classical case definition	33
2.4.2	Logistic power model	34
2.4.3	Non-parametric monotone regression	35
2.4.4	Latent class model.	35
2.4.5	Bayesian latent class models	36
2.4.6	Sensitivity and specificity of case definitions.....	38
2.4.7	Estimation of individual MAF probabilities.....	38
2.5	Protein microarray data processing.....	39
2.5.1	Gain scanning for protein microarray	40
2.5.1	Background correction.....	42
2.5.2	Coefficient of variation	44
2.5.3	Normalisation and Batch correction	45
2.6	Feature selection	49
2.6.1	Modified Poisson.....	50
2.6.2	Penalized regression	51
2.6.3	Bayesian regularization methods.....	54
2.6.4	Random Forest.....	56
2.6.5	Boosting trees	60
2.6.6	sparse Partial Least Squares discriminant analysis (sPLSDA)	62
2.6.7	Block Forest and Networks	63
2.7	The inter-linkages of the different methodologies.....	65
2.8	Ethical considerations	67
CHAPTER 3 : Equation Chapter (Next) Section 1 CASE DEFINITION Paper 1. Estimating malaria		
attributable fractions with changing transmission intensity: Bayesian latent class vs logistic models.....		68
3.1	Abstract:.....	68
3.2	Background	70

3.3	Methods.....	73
3.3.1	Study area and population.....	73
3.3.2	Malaria parasite prevalence cross-sectional survey.....	74
3.3.3	Statistical analysis	75
3.4	Results.....	78
3.4.1	Study population.....	78
3.4.2	Temporal distribution of fever and parasite density	78
3.4.3	Relationship of fever to parasitaemia over time	79
3.4.4	Impact of age on MAF.....	81
3.5	Discussion:	83
3.6	Conclusion:.....	85
3.7	Ethics approval and consent to participate	86
3.8	Consent for publication.....	86
3.9	Availability of data and materials	86
3.10	Competing interests.....	86
3.11	Funding	86
3.12	Authors' contributions	87
3.13	Acknowledgements.....	87
3.14	Supplementary.....	88
CHAPTER 4	: PROTEIN MICRO-ARRAY DATA PROCESSING Equation Chapter 5 Section 1	93
4.1	Abstract.....	94
4.2	Introduction.	95
4.3	Methods.....	97
4.3.1	Data extraction.....	98
4.3.2	Background correction.....	98
4.3.3	Coefficient of variation.	101
4.3.4	Normalisation.....	102
4.3.5	Batch Correction.	107
4.4	Results.....	108
4.4.1	Implementation.	108
4.4.2	Background Correction	108
4.4.3	Within sample variation and purification tag subtraction.....	109
4.4.4	Normalisation and Batch Correction.....	111

4.5	Discussion:	112
4.6	Availability and implementation.....	115
4.7	Funding	115
4.8	Declaration of Competing Interest	115
4.9	Supplementary.....	116
4.9.1	Supplementary A: Definition of terms and quantitation results file contents	116
4.9.2	Supplementary B: ComBat Batch Correction.....	118
4.9.3	Supplementary C-1.....	120
4.9.4	Supplementary C-2.....	122
4.9.5	Supplementary D:	123
CHAPTER 5 : FEATURE SELECTION		124
5.1	Introduction	125
5.2	Methods.....	127
5.2.1	Epidemiological Data	127
5.2.2	Protein micro-array data.....	129
5.2.3	Feature selection using the Binary outcome	133
5.2.4	Network Models.....	139
5.2.5	Feature selection using probability outcome.	140
5.3	Results.....	141
5.3.1	Binary Outcome Results.....	141
5.3.2	Probability Outcome Results.....	150
5.4	Discussion:	154
5.4.1	Selected feature and biological importance:.....	160
5.4.2	Conclusion:.....	161
5.5	Supplementary.....	162
CHAPTER 6 : DISCUSSION.....		165
6.1	Identifying correlates of disease protection in malaria	166
6.2	Case definition	166
6.3	Protein microarray data processing.....	168
6.3.1	Background Correction	169
6.3.2	Within sample variation.....	169
6.3.3	Normalisation.....	170
6.3.4	Batch Correction	170

6.3.5	Protein micro-array data processing suites	171
6.4	Feature selection of Merozoite targets of protective immunity	172
6.4.1	Modified Poisson.....	172
6.4.2	Random Forest.....	173
6.4.3	Boosting trees	173
6.4.4	Bayesian Lasso	174
6.4.5	Network Models.....	175
6.4.6	Misclassification bias and feature selection	176
6.4.7	Selected feature and biological importance	177
6.4.8	Strengths and limitations	177
6.4.9	Recommendations.	179
6.4.10	Summary and Conclusions	180

List of Figures

Figure 1-1: Malaria life cycle: Source CDC.org.....	5
Figure 1-2: Forward phase and reverse phase protein micro-array.....	10
Figure 2-1: Epidemiological design of the cohort studies	26
Figure 2-2: Background correction: artefacts add noise to the signal intensity	40
Figure 2-3: General schematic flow of the random forest algorithm (136)	57
Figure 2-4: Schematic flow of Bagging.....	59
Figure 2-5: Sequential boosting approach workflow, where the h^* are the residuals	61
Figure 2-6: Interdependencies of omics data from different blocks. Source: Zierer et al. (146). 64	
Figure 2-7: The schematic presentation presenting the linking to of the thesis objectives.	66
Figure 3-1: Junju and Chonyi study sites in the Kilifi Health and Demographic Surveillance System (KHDSS)	74
Figure 3-2: A) Temporal estimates of attributable fraction (AF) from 2005 to 2017 using Bayesian Latent class models and logistic power models.....	80
Figure 3-3: Malaria attributable fractions and probabilities over age group for all the study participants.....	82
Figure 4-1: protGear data processing scheme.....	97
Figure 4-2 Background correction: artefacts add noise to the signal intensity	100
Figure 4-3: Example of background diagnostic plots produced by protGear.....	109
Figure 4-4: The visualization of the CV	110
Figure 4-5: Standard deviation vs mean plots (meanSdPlot) of A) Non normalised data B) log2 normalisation C) Cyclic loess normalisation D) Robust Linear Model normalisation and E) VSN normalisation	111
Figure 4-6: Comparison of Batch m1 and Batch m2 before and after batch correction using Combat() approach.....	119
Figure 5-1: The multicentre cohorts from West and East African countries.....	128
Figure 5-2: Parasite positivity and malaria positivity for the multicentre cohorts.....	129
Figure 5-3: Antibody data processing workflow for the laboratory data.....	131
Figure 5-4: Protein micro array data.....	132

Figure 5-5: Modified Poisson regression RR estimates.	143
Figure 5-6: Estimated risk ratios and the Confidence intervals from the MP model	144
Figure 5-7: The OOB error before SMOTE and after SMOTE with “mtry” of 57 and 10000 trees	145
Figure 5-8: Rank importance using the Random Forest approach	146
Figure 5-9 Rank importance using the GPBoost approach	147
Figure 5-10: Bayesian Lasso estimates.	148
Figure 5-11: A network diagram showing the relationship of the top antigens per site. T	150
Figure 5-12: Sample size and probability distribution of the Junju cohort	151
Figure 5-13: Top antigens picked by zero inflated and binomial mixed effect models for Junju, Kilifi dataset.	153

List of Tables

Table 1-1: Logistic models of case definition.....	18
Table 2-1: A summary of the study sites	26
Table 3-1: Prevalence’s of Plasmodium falciparum positivity, fever, and presumptive malaria (fever + parasitemia) in the pre-transmission season cross-sectional survey and the number of active follow-up events in the Junju Cohort..	79
Table 3-2: A comparison of a binary and probability outcome using high vs low antibody levels in Junju 2008 and Chonyi 2000 cohort.	83

Nomenclature/List of Abbreviations and Symbols

WHO - World Health Organization

SSA - sub-Saharan Africa

EIR - Entomological Inoculation Rates

ITNs - Insecticide-Treated Nets

IRS - Indoor Residual Spraying

ELISA - Indirect enzyme-linked immunosorbent assay

PMT - Photo-Multiplier Tube

GLM - generalized linear models

GPL-3 - General Public License version 3 (GPL-3)

EM - Expectation-Maximization

DNA - Deoxyribonucleic Acid

RNA - ribonucleic acid

CRAN - Comprehensive R Archive Network

PAA - Protein Array Analyser

PMA - Protein Microarray Analyser

PMD - Protein Microarray Database

PAWER - Protein Array Web Explorer

HIV - Human Immunodeficiency Virus

LASSO - Least absolute shrinkage and selection operator

PLS-DA - partial least squares discriminant analysis

OOB - out-of-bag

ROC -Receiver operating characteristic

AUC - area under the curve

SMOTE - synthetic minority oversampling technique

SMART - South-South Malaria Antigen Research Partnership

MAF - malaria attributable fraction

MFI - Mean Fluorescent Intensity

SNR - Signal to Noise Ratio

FG - Foreground

CV - Coefficient of Variation

VSN - variance stabilization normalization

RLM - robust linear model

CI - Confidence Interval

LOESS - local regression

PCA - principal component analysis

SVA - single value decomposition

DWD - Distance-weighted discrimination (DWD)

SVM Support Vector Machines (SVM)

ANOVA - way analysis of variance (ANOVA),

RR - Relative Risk

OR - ODDS Ratio

RSS - residual sum of squares

RF - Random Forest

RF-RI - Random Forest Random-Input

RMSE- Root Mean Squared Error

PLS - Partial Least Squares

sPLS - sparse Partial Least Squares regression

sPLS-DA - sparse Partial Least Squares discriminant analysis

KHDSS - Kilifi Health and Demographic Surveillance System

BIC - Bayesian Information Criterion

CD4 - CD4 hexa-histidine

MBP - Maltose binding protein
GST -Glutathione S-transferase
MVD - mean-variance dependence
EB - Empirical Bayes
CART - Classification and Regression Trees (CART)
MDA - Mean Decrease Accuracy -MDA
VI - Variable importance
MP - Modified Poisson
MCMC - Markov chain Monte Carlo
JAGS – Just Another Gibbs Sampler
MERF - Mixed-effects random forest for clustered data
LD - linkage disequilibrium
SNP - Single nucleotide polymorphisms
CPS - chloroquine chemoprophylaxis
WAIC – Watanabe Information Criterion
SADC - Southern African Development Community
E8 - Elimination Eight
MOSASWA – Mozambique , South Africa and Swaziland region

CHAPTER 1 : INTRODUCTION

This chapter introduces the research thesis by giving an overview of malaria immunity, misclassification bias, background of disease immunity, and identifying correlates of protection against disease using high throughput experiments. The chapter also lists the thesis research question and justification together with the thesis objectives and highlights the statistical methodologies applied to deriving case definitions and for the identification of correlates of protection from malaria. Finally, the chapter links the thesis framework with the research articles generated from this work.

1.1 Overview

The process of identifying important biological features is challenging especially for outcomes prone to differential misclassification and significant biological complexity. Misclassification can lead to bias in describing, framing, selecting, and comparing cases and controls. Consequently, researchers across the globe, are facing challenges in making data-driven decisions and recommendations. This is particularly relevant for studies aimed at identifying signatures (correlates of protection) of naturally acquired immunity against malaria due to challenges in defining a clinical malaria case (1,2). The potential misclassification of cases may influence the identification of the important immune signatures and the estimation of their effect (3,4).

1.1.1 Malaria background

The human malaria parasite *Plasmodium falciparum* causes approximately 215 million malaria cases that occurred in the World Health Organization (WHO) African region in 2019 resulting in approximately 0.4 million deaths each year (5). Sub-Saharan Africa

accounts for 94% of malaria cases worldwide , with 6 countries accounting for 51% of global malaria deaths (5). The South-East Asian region ranking second with approximately 6.3 million cases. *Plasmodium falciparum* is the species that is responsible for 99% of the reported cases in sub-Saharan Africa (5).

Different organisations in the globally and within the African continent are working towards malaria elimination (6). The WHO Global Technical Strategy for Malaria 2016–2030 strategy emphasizes a need to strengthen regional collaboration towards elimination. Regionally, SADC with the Elimination Eight (E8) and MOSASWA regional initiatives are working towards a malaria-free Southern Africa with the E8 malaria elimination strategic plan among others(6,7). Great Lakes Cross Border Malaria Initiative a collaboration of the Eastern African Community (EAC) countries and Democratic Republic of Congo (DRC) working towards cross-border malaria control and prevention initiatives (8). The WHO aims to eradicate malaria by 2030, although recent estimates have shown a decline in malaria mortality rates (5), in certain regions, the incidence of malaria continues to rise despite the different approaches being used to tackle malaria (9,10).

The *Anopheles* mosquito *species* are responsible for the spread of malaria and are widespread in sub-Saharan Africa region which also has the highest entomological inoculation rates (EIR) (11,12). These vector species are dominant and well adapted to transmit malaria, leading to a high burden of malaria morbidity and mortality (12). Consequently, various malaria control approaches aim to reduce human-vector contact.

The most commonly used methods to prevent mosquito bites are Insecticide-Treated Nets (ITNs) and Indoor Residual Spraying (IRS) (5).

Nevertheless, despite these different malaria control strategies malaria continues to impose a considerable burden of morbidity and mortality (5). The logistics of distribution, budget constraints and the durability of ITNs are key challenges to control strategy (13). It is widely accepted that novel control measures will be required to help the fight against malaria and subsequently move from control to elimination (14). To help in achieving this goal of elimination and eradication, an effective malaria vaccine is thought to be important, given the emerging resistance to vector control measures (15).

A strong rationale for developing and improving malaria vaccines comes from epidemiological studies in malaria-endemic countries. These studies show that individuals who are continuously exposed to malaria gradually develop natural immunity that initially prevents death from severe malaria, and subsequently recurrent clinical episodes of mild malaria. Therefore, identifying the targets underlying this naturally acquired immunity is of importance but remains challenging for a variety of reasons (2).

The first of these is the potential for misclassification bias in defining clinical malaria. Individuals from malaria-endemic areas are often asymptotically infected. Thus, using fever (as an objective measurement of illness) plus parasitaemia as a definition of a malaria case has the risk of falsely attributing symptoms by coincident infections to malaria. Misclassification bias is thus a problem for both mild and severe malaria that has an impact on attempts to characterize any aspect of the disease.

Secondly, defining immunity against malaria following natural infections is also not straight-forward. Immunity can be defined in different ways; a) lack of clinical malaria symptoms over a period of observation and b) evidence of malaria challenge (i.e., asymptomatic parasitaemia) without any clinical signs of malaria. One challenge for the former is that participants could either be genuinely immune or may have not been exposed to an infectious mosquito bite (1), leading to a misclassification bias. Furthermore, for both definitions, the same individuals can have both asymptomatic infections and clinical malaria during the same observation period, making it difficult to classify them into either category. It has been recommended that observational studies would have increased power if they differentiated between unexposed and immune individuals in analyses to identify correlates of protection (16).

Additionally, the malaria parasite has a complex molecular and multi-stage life cycle (Figure 1-1) with over >5,300 proteins expressed as the infection develops. Antibodies against malaria antigens have been shown to play a vital role in naturally acquired immunity (2,16). However, identifying the specific antibodies that confer protection has been challenging and is a major barrier to vaccine development. Cohort studies are powerful tools that could be used to detect potentially protective associations, but most have been single cohort studies focusing single or limited immune responses. These individual cohort studies are not sufficiently powered to robustly analyse hundreds of individual responses or combinations of response (16). Longitudinal cohort studies in which the endpoint of clinical malaria is compared amongst individuals with or without specific antibodies have been considered the gold-standard for identifying the correlates

of protection. Conflicting evidence has been reported from these studies due to the

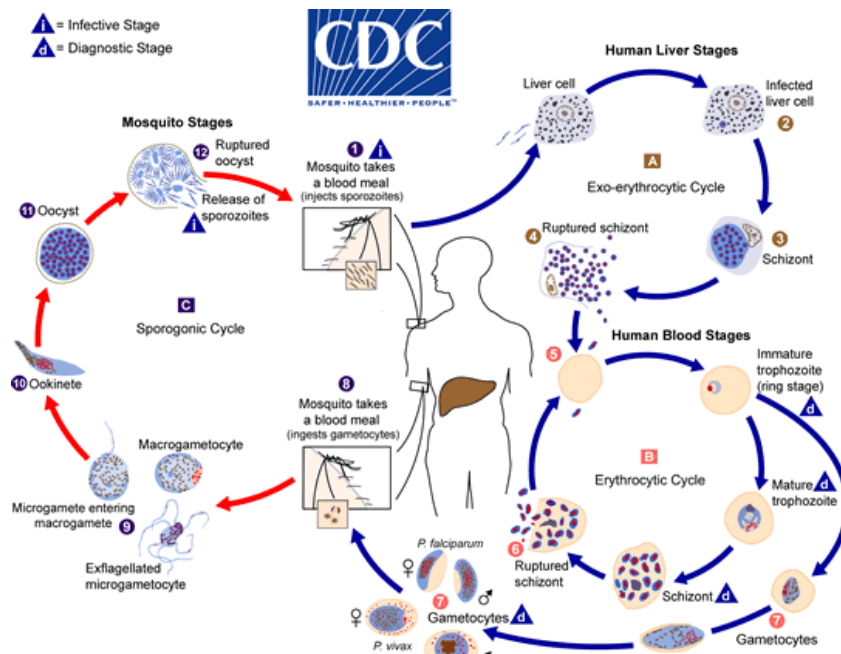


Figure 1-1: Malaria life cycle: Source CDC.org

differences of malaria transmission, classification bias, and case definitions. This lack of clarity stems in part from a range of methodological differences between studies ranging from the study design and populations to detailed laboratory methods and analytical approaches (16).

1.1.2 Malaria case definition

Many diagnostic assays classify samples into two groups according to whether some output (i.e., parasite density or optical density) exceeds a given cut-off. The two groups are mainly positive or negative and this in turn define the clinical statues of an individual, case - disease positive or non-cases (disease negative). However, some of the diseases are prone to differential misclassification (17,18). Such misclassification can lead to biased estimates of a disease frequency and inferences drawn from epidemiological studies.

Different studies have looked at the effect of misclassification on inferential statistics, one of them is that it biases relative risk estimates. These evaluations have shown the importance of addressing this problem during analysis (19). When defining a disease outcome, these errors mainly occur due to false positives or negatives.

1.1.2.1 Misclassification in malaria case definition

In malaria, to select important biological features or evaluating the effectiveness of different vaccine interventions, clear case definitions are needed. Asymptomatic parasitaemia are very common in areas of high malaria transmission; individuals have parasites but do not show clinical signs of malaria. The main reason for this is individuals living in malaria endemic areas develop immunity naturally under continuous parasite exposure (20). This may lead to misdiagnosis of malaria cases, where an individual who has asymptomatic malaria may have a fever due to another infection but is diagnosed to have malaria. Thus, parasitaemia accompanied by fever may not be adequate to define an episode of clinical malaria and may lead to over-estimation of malaria burden (19,21).

Additionally, sensitivity and specificity of case definitions of malaria in endemic areas are variable with time and age (19) due to change in factors that affect malaria immunity. This necessitates a review of commonly used malaria case definitions to preserve sufficient power in epidemiological studies and vaccine trials (22).

Although increase in parasite density is generally associated with a higher risk of developing fever, the association is not linear (21). To overcome this problem of potential misdiagnosis especially in case of low parasitemia and fever due to other non-malaria causes, different studies have based the clinical definition of malaria on fever with

parasite density above a certain threshold. Computation of malaria attributable fractions; that is determination of the proportion of fevers due to malaria have been used to define different cut-offs for parasitaemia (18,19). These cut-offs are utilised as thresholds to categorise an individual as malaria positive or negative.

A range of case definitions exists due to different factors (19,22,23). The cases definitions range from any parasitaemia count to around 7,000 parasites per microliter, depending on age (19), malaria season (24); malaria transmission and topology (25) and change of transmission over time. Significant reductions in malaria transmission have been reported over the last decade in endemic countries in Africa (26)

1.1.3 Identifying correlates of disease protection

Correlates of protection or important immune signatures are measurable parameters that predict whether or not an individual is protected against developing a disease. Modern data often have far too many variables for classical model building especially in selecting important immunological parameters of disease. This is attributed to the growth in Biology to simultaneously measure many parameters of immunity and to generate large datasets over different parameters for a single individual. The main assumption in defining correlates of protection is that the outcome or disease is defined correctly. In contrast, if there is a misclassification, the point estimates will be biased towards the null hypothesis. Additionally, this data could lead to biased estimators and incorrect inference. Therefore, the process of selecting the essential parameters is challenging, especially for problems which have significant complexity and multiple potential correlates of protection.

Malaria provokes complex immunological response. A wide range of immunological parameters for malaria has been proposed as correlates of protection against clinical episodes of malaria but there is not a single universally accepted correlate. These immunological parameters can be considered in three broad groups: a) cellular parameters such as T-cells (27–30) b) humoral parameters which include antibody binding measurements or functional assays (31) and c) soluble parameters such as cytokines.

The main challenge in understanding the targets of, and mechanisms underlying, naturally acquired immunity is the disease complexity and interacting biological and environmental factors. Identifying the important biological parameters depends on several interacting genetic, molecular, environmental and other factors spanning through many temporal and spatial scales (32). This has led to a tremendous growth in understanding the biological complexity leading to massive generation of datasets over different parameters especially antibodies against different disease antigens. In the example of malaria, the completion of the *Plasmodium falciparum* genome led to the identification of numerous potential vaccine candidates (33).

1.1.3.1 Measuring Immune response

Different discovery systems and technologies are used to detect and quantify antibodies in samples. Some of the technologies use colorimetry which quantifies the antibody concentration by measuring the light absorbed after forming a biological reaction that changes colour. Indirect enzyme-linked immunosorbent assay (ELISA) is one of the colorimetry techniques which is a two-step procedure that involves two binding processes

of primary and labelled secondary antibody. The secondary antibody is tagged to an enzyme that reacts with its substrate. Light absorption of the colour formed after substrate addition is measured and converted to optical density values. However, colorimetry techniques require high amounts of antibodies and samples, consequently, they are time consuming and expensive (34).

Other technologies use fluorescence approach which quantify the signal emitted from a biological reaction. This technology is combined with a Photo-Multiplier Tube (PMT) system which is responsible for amplification of the fluorescent signals. Researchers adjust the PMT gain to have a balance between signal bleaching and enhancing lower signals. Protein micro array is one of the fluorescent techniques that quantify the absolute fluorescent emission detected following antibody binding to a single protein. There are mainly two standard types of protein micro array formats, namely the forward phase and the reverse-phase (35,36).

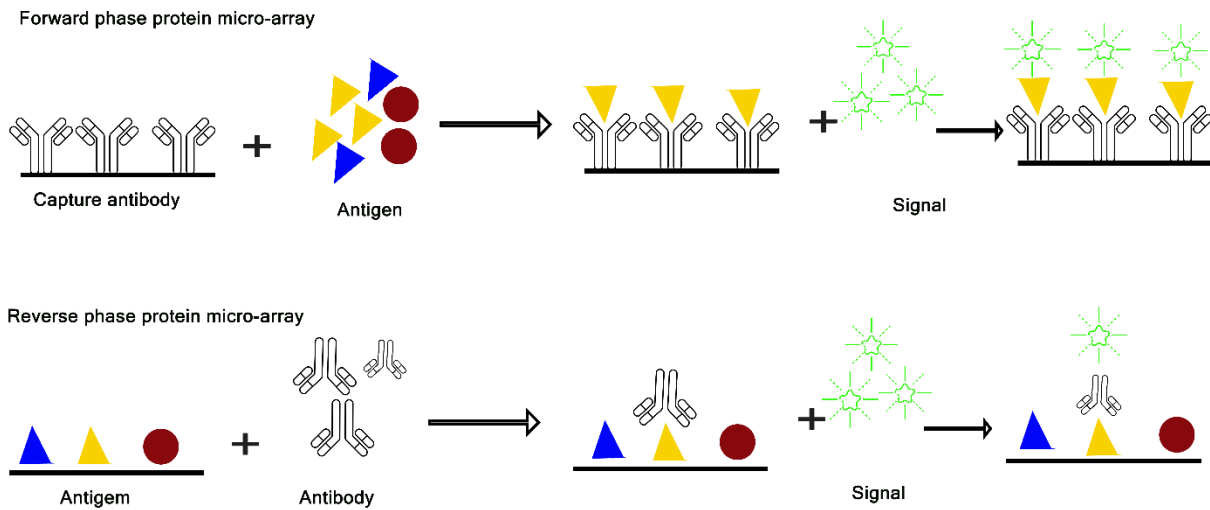


Figure 1-2: Forward phase and reverse phase protein micro-array

Fluorescence techniques are a little more sensitive than colorimetric assays and they allow for a high reading to be measured by widening the dynamic range. These techniques facilitate a high-throughput approach to parallel quantitation of specific proteins in a rapid, low-cost, and low-sample-volume format. However, variations in experimental conditions can introduce systematic and non-systematic sources of bias that have the potential to derail the intended outputs. Antibody detection systems that rely on fluorescence therefore, require careful calibration to minimize background signals arising from the non-specific binding of analytes or spill-over fluorescence from reference materials or high-signal intensity neighbouring spots (37,38). Also, fluorescence and the PMT gain adjustment are significant source of non-systematic and systematic biases (38). Rigorous data pre-processing steps which involve quality control and normalization are of importance before doing downstream analysis especially of selecting important disease parameters.

1.1.3.2 Analysing association between antibody measurements and protection against disease

Different analytical and modelling approaches have been used in the identification of correlates of protection. Binomial models with logit link (logistic models) have been used to identify correlates associated with protection (1,39,40). The binomial models report log odds which have been reported not to have an intuitive interpretation as the relative risk. Similarly, modified Poisson models (Poisson model models with robust error variance) have been used over logistic models (16,41,42). Modified Poisson models provide relative risk as the parameter of primary interest and assist in convergence problems that sometimes arise with binomial regression models (43). Logistic and Poisson regressions are of the generalized linear models (GLM) family and use the iterative weighted least squares method to obtain the maximum likelihood estimates (44).

However, maximum likelihood-based regression models are impacted by multiple random effects and are also slow due to the large covariate set which are computationally demanding. A single complex model incorporating many co-variables carries the risk of high variance leading to overfitting or underfitting. On the other hand univariate tests while adjusting for common markers of exposure, ignore the potential interplay between different covariates.

Major strides in the analytical approaches towards a better understanding of the targets of, and mechanisms underlying, biological immunity (33) to identify important markers have been made with the continued growth of data (45). In particular, when dealing with data from a high-throughput proteomics technology, a large number of combinations and interactions must be tested which is a shortcoming for many regression approaches.

Models based on machine learning approaches offer a systematic way to intergrate possible combinations of immune responses against various antigens (46). Machine learning (46–48) and systems biology (48) have become popular choices for analysis of high throughput datasets in biology. Machine learning and systems biology approaches assists in handling unknown interactions across multiple variables, that may fail to conform to the assumptions of many classical statistical methods. The machine learning methods are categorised into either supervised or unsupervised approaches. Supervised machine learning involves building a statistical model for predicting, or estimating, an output (outcome) based on one or more input variables. With unsupervised machine learning, there are input variables but no supervising (outcome) output.

Supervised learning approaches are categorised mainly into two groups, regression analysis and classification approaches. The regression approaches predict a continuous outcome variable y based on multiple predictors X , i.e., predicting the probability of having malaria (having the outcome as a probability) using different immune parameters or predicting life expectancy based on socio-economic factors. Classification approaches predict a class or a group, i.e., malaria case yes/no being over different immune parameters.

Selecting correlates of protection against malaria depends on some interacting biological and spatial factors. For this reason, using analytical methods well suited to handle multiple biological factors is of great importance in identifying the key antibodies associated with natural immunity (48).

A strong rationale for developing and improving malaria vaccines comes from epidemiological studies in malaria-endemic countries. These studies show that individuals continuously exposed to malaria gradually develop “natural immunity” that initially prevents death from severe malaria, and subsequently recurrent clinical episodes of mild malaria. Therefore, identifying the targets underlying naturally acquired immunity is of importance.

However, this remains challenging from a biological point of view due to the complex life cycle of malaria parasites and their interaction with the human immune system have, to date, proved very difficult to unravel. For this reason, data arising from malaria immunity studies are so closely entwined with confounders that it is difficult to dissect the useful information from noise using classical analytical approaches (2).

1.1.4 Research problem and justification: What is novel?

Many biomedical assays are expected to classify samples into two groups according to whether some output (i.e., parasite density or optical density) exceeds a given cut-off. However, some of the diseases especially malaria are faced with bias due to differential misclassification (2,17). Conflicting evidence has been reported from these studies due to the differences in disease transmission, classification bias, case definitions and heterogeneity of disease exposure (32,49). Additionally, significant reductions in malaria transmission have been reported over the last decade in endemic countries in Africa (26) but little effort has been made to estimate the case definitions with changing transmissions or define a standard way of handling cases (17,18).

On the other hand, a tremendous growth of disease-related data sets has been observed over time. This is attributed to the growth in biology and high throughput technologies to identify parameters associated with protection to improve vaccine discovery and development. Data from these high-throughput technologies especially protein microarray are compounded with systematic and non-systematic biases. On top of that, these datasets are often have far too many variables for classical model building especially in selecting important immunological parameters of disease. Different analytical approaches for case-control and cohort studies have been used to identify correlates of naturally acquired immunity against malaria with different case definitions (39–41,47,50–52).

However, with the different interacting components in the malaria immunity definition, classical regression approaches, mostly ignore the potential interplay between different factors. Mostly, people rely on univariate tests, while adjusting for common markers of exposure. Such approaches do not handle the differential misclassification bias of quantifying immune and non-immune candidates in observational cohort studies well. Unfortunately, there are limited tools available to handle and compare different pre-processing techniques to handle the biases from high through-put protein micro-array data.

Data from a high-throughput proteomics technology have many combinations and interactions that must be tested which is a shortcoming in many regression approaches. Predictive modelling frameworks based on machine learning offer a systematic way to consider all possible combinations of immune responses (46).

The novelty of this study is the extension of malaria case definition approaches to calculate propensity scores depending on parasite density and fever together with results to show the impact of changes in of malaria case definitions over time. Furthermore, the study develops a novel one-stop-shop protein micro-array data processing suite coupled with machine learning approaches to select correlates of disease protection from multiple sites in Sub-Saharan Africa. The use of machine learning approaches assists in handling unknown interactions across multiple disease parameters, that may fail to conform to the assumptions of many classical statistical methods.

Packages and findings from this study will enhance wide uptake of modern analytical approaches in the infectious disease and vaccine development community. Additionally, the packages produced from the study are open source and distributed under GNU General Public License version 3 (GPL-3) for the analysis and comprehension of protein micro array data generated by wet-lab experiments in molecular biology. The GPL-3 license ensures contribution to the free, open-source and open development software projects.

1.1.5 Thesis research questions

- 1) Are case definitions for malaria in endemic areas likely to vary over time as factors affecting immunity change?
- 2) Which are the important biological parameters associated with malaria protection in multi-centre cohort studies?
- 3) Do the correlates of disease protection vary in different subpopulation groups (age, transmission intensity, high vs low immune responders)?

1.1.6 Thesis broad and specific objectives

The primary aim of this study was to develop analytical approaches that are sensitive to the challenge that the outcome is prone to differential misclassification and to identify the correlates of protection against disease using machine learning approaches.

The specific objectives were:

- 1) To develop case-definition models for outcomes prone to differential misclassification using maximum likelihood techniques and compare with the Bayesian latent class models.
 - **Paper 1 title:** Estimating malaria attributable fractions with changing transmission intensity: Bayesian latent class vs logistic models.
<https://doi.org/10.21203/rs.3.rs-916409/v1>
- 2) To develop an R based package for handling protein-microarray data pre-processing.
 - **Paper 2 title:** protGear: A protein microarray data pre-processing suite;
<https://doi.org/10.1016/j.csbj.2021.04.044>
 - Package repository: <https://keniajin.github.io/protGear/> or
<https://github.com/Keniajin/protGear/>
- 3) To use machine learning approaches to identify important biological parameters associated with protection in multi-centre cohort study.
- 4) To develop network models to visualize the interplay across different correlates of protection against malaria.

- **Paper 3 title:** Defining the merozoite targets of protective immunity against *Plasmodium falciparum* malaria through multi-center cohort studies.

1.2 Literature Review

1.2.1 Evolution of case definition approaches in malaria

In malaria-endemic areas, the case definition for symptomatic malaria involves fever together with a parasite density exceeding a given cut-off. The cut-off is derived from estimating the fraction of fevers attributable to malaria (denoted as λ). In 1994, Smith et. al, explored different case definition techniques using data from Kilombero District in Tanzania. Kilombero then was a high endemic area where over 80 per cent of young children were parasitaemic. They compared the classical case definition approach;

$$\lambda = \sum_{j=1}^j p_j (R_j - 1) / R_j ; \text{ where } R_j \text{ is the odds ratio of fever in for the } j^{th} \text{ category relative}$$

to the baseline category $j=1$ with aparasitaemic children, with model-based approaches.

The modelling approaches were of the form $\text{logit}(\pi_i) = \alpha + f(x_i)$ where π_i is the probability that observation i with parasite density x_i is a (fever) case and $f(x)$ is a smooth monotonic function of X .

Table 1-1: Logistic models of case definition

Models		Odds ratio of fever for parasite density x_i	Attributable fraction λ
Model 1 - Linear in x	$f(x_i) = \beta x_i$	$R_i = \exp[f(x_i)]$	$\lambda = (1/N) \sum_1^i (R_i - 1) / R_i$
Model 2 - Linear in $\log(x)$	$f(x_i) = \beta \log(x_i)$		
Model 3 - Power of x	$f(x_i) = \beta x_i^\tau$		
Model 4 - Recoding low parasitaemia	$f(x_i) = \beta (x_i')^\tau$		

The classical approach had limitation of assuming monotonic relationship between parasite density and fever thus they modelled the relationship between the parasite density and malaria fever as a smooth function as shown in Table 1-1. The logistic regression with the power function (Table 1-1 Model 3) was reported to perform better than the other models as assessed by the deviance and residuals (21,53).

The logistic approach has been used in different malaria-endemic areas to define a range of case definitions that exist due to different factors (19,22,23). The case definitions range from any parasitaemia count to around 7,000 parasites per microlitre, depending on age (19), malaria season (24), malaria transmission and topology (25).

Additionally, Smith et al. proposed a non-parametric monotonic regression approach with latent class models which provided precise and close to unbiased estimates compared to the logistic approach. The approach involved dividing the data into K ordered categories of parasite density and count numbers of test samples $\mathbf{n} = (n_0, n_1, \dots, n_{k-1})$ and controls

(non-fever cases) $\mathbf{m} = (m_0, m_1, \dots, m_{k-1})$. This is followed by estimating the category-specific attributable fractions. The overall attributable fraction was obtained by the weighted mean, $(1/N)\lambda^T \mathbf{n}$, where n is the number of test samples per category and λ^T is the category-specific attributable fraction. In category one, the attributable fraction was assumed $\lambda_0 = 0$. Two limitations were reported, one, assuming an attributable fraction of $k = 0$ to be $\lambda_0 = 0$ and secondly, not estimating the maximum likelihood estimates for the overall or category-specific attributable fraction. To overcome the limitations, the authors proposed a latent class model which provides maximum likelihood estimates of the attributable fractions using the EM algorithm. The latent class approach was reported to produce biased estimates where there was an extreme overlap of cases and controls in a given category (18).

A Bayesian latent class model of two-component mixture distributions was proposed to improve the estimation of attributable fractions (17). The data was divided into K ordered categories over the range of the parasite density like the latent class approach described above. A mixture distribution of two multinomial distributions was assumed consisting of parasite densities in children with fever due to malaria or fever due to other causes. Specifically, one component of the mixture corresponds to children without clinical malaria (training component) and the second component with children with clinical malaria. A posterior density of the joint probability density of the multinomial distributions with a uniform prior was then used to estimate the parameters using Gibb's sampling.

Importantly, this approach was reported to give comparable results with the other modelling techniques while offering the best statistical fit (17).

1.2.2 Protein microarray data pre-processing tools

Protein microarray technology is increasingly utilised for vaccine candidate discovery among other range of applications in the 'omics era' with hundreds to thousands of antigen-specific antibodies analysed simultaneously (37,54–56). However, most of the published guidelines and tools favour DNA-based microarrays that often differ fundamentally in their experimental design with protein micro-array.

Linear Models for Microarray Data (Limma) is a Bioconductor/R package that specifically developed differential expression analyses of RNA-sequencing and DNA-microarray data (57). However, this tool has been applied in different protein micro-array studies by adopting its statistical principles (58–60). The RPPAnalyzer statistical tool developed to read reverse-phase protein array data adopts most of its data processing techniques from Limma and provides a downstream analysis option using heatmaps, boxplots, time course plots and correlation plots (58). Limma and RPPAnalyzer are open sources and freely available as R-Packages in Bioconductor and The Comprehensive R Archive Network (CRAN) respectively.

Another tool is Protein Array Analyser (PAA) a Bioconductor/R package build on top of Prospector's core functionality and was specifically developed for biomarker discovery with protein microarrays. The tool was last updated in 2016 and had six specific work flows namely data import, pre-processing and quality control, differential analysis, feature

preselection, feature selection and biomarker candidate's selection. However, high-level R programming skills are required to utilise PAA effectively (61).

The Protein Microarray Analyser (PMA) was developed in 2018 as a multi-platform desktop application. The tool is built in Java programming language however no vignette or software documentation is available (62). The only fully web tool available is Protein Microarray Database (PMD) but it publicly prioritises uploading and archiving of protein microarray datasets. The analysis tool offered by PMD lacks proper user interaction and a vignette for clear guidance is not provided(63).

Protein Array Web Explorer (PAWER) is another web tool build on R for handling protein micro array data(64). The tool automatically performs the data pre-processing and is linked with the Clustering analysis in ClustVis platform (65). The PAWER package and the web server code are available under GNU GPL v2 license. The tool does not offer a comparison of the background and normalisation methods and a clear vignette is not provided. One of the major limitations observed with the tools is the lack of proper documentation and comparison of the different data processing techniques which is very important in this era of reproducible research (64).

1.2.3 Machine learning approaches and correlates of malaria protection.

A strong rationale for developing and improving malaria vaccines comes from epidemiological studies in malaria-endemic countries. These studies show that individuals continuously exposed to malaria gradually develop “natural immunity” that initially prevents death from severe malaria, and consequently recurrent clinical episodes of mild malaria (2). Therefore, identifying the targets underlying naturally acquired immunity is

of importance but remains challenging. For instance, lack of enough statistical power and small effect sizes in many studies as well as the challenge of using classical analytical approaches to analyse hundreds of individual responses or combinations of responses (1,47).

A study done in Kilifi in 2008 by Osier et.al established that natural immunity to malaria results from high antibody titres to several targets (16). In addition, the same group further advocated that high levels of protection against *P. falciparum* can be achieved through a vaccine consisting of multiple components. To support this, the group used modified Poisson regression on 36 individual antibody measurements to identify the top targets of malaria protection adjusting for age and schizont extract. Next, they performed combinatory analysis and demonstrated that a combination of the antibodies most strongly associated with protection gave a high potential protective efficacy. To ensure sufficient statistical power, the combinations were restricted to five proteins (41).

Finney et al in 2014 constructed a protein array with 1922 *Plasmodium falciparum* and 1936 *Plasmodium vivax* recombinant proteins expressed by the blood stage of the malaria parasites. A total of 225 serum samples from Papua New Guinea were used to study naturally acquired immunity against single-species and mixed-species infections of *Plasmodium falciparum* and *Plasmodium vivax*. They utilised the evolutionary trees using the *evTree* algorithm with eightfold cross-validation to measure disease classifier accuracy. Finally, they utilised the mProbes (45) feature ranking algorithm with Random Forest feature selection to investigate the relationship between antibody responses and malaria symptoms then selecting the most informative biomarkers of disease immunity.

mProbes was used to estimate the false discovery rate of the features selected using the Random Forest. The study was highlighted as a proof of concept for applying machine learning algorithms to the identification of important malaria vaccine targets (66).

These multi-dimensional analytical approaches have also been applied in HIV studies to investigate the association among humoral responses elicited in HIV vaccine trials and drug development (67,68). One of the studies done by Chung et al. was inspired by the rationale of multifactorial vaccine development strategy used both supervised and unsupervised machine learning approaches to investigate relationships among antibody responses and HIV immunity (68). In particular, the least absolute shrinkage and selection operator (LASSO) followed by partial least squares discriminant analysis (PLS-DA) were used to identify distinctive features that greatly contributed to HIV profile differences. The PLS-DA method evaluated the biomarkers selected using LASSO after mean centering and variance scaling the data. Orthogonal signal correction and correlation networks were then used to demonstrate model visualization and pairwise correlation, respectively.

More recently, a random forest machine learning approach was utilised to select signatures that highly predicted an individual's immunity status (47). This study used data of malaria antibody profiles against several *Plasmodium falciparum* antigens from three longitudinal cohort studies that were done in Kenya, Kenya/Tanzania and Mali. The Kenyan dataset had 286 individuals with ELISA-based immune profiles from 36 *P. falciparum*-specific antigens. The second cohort immune profile dataset was based on luminex-derived IgG levels against 46 individual PfEMP1 domains of 447 children. Mali

dataset had protein microarray-based antibody reactivity data of 186 participants against a panel of 2320 *P. falciparum*-specific epitopes of the 3D7 line, representing 1204 unique proteins (47). In this study each of the dataset was analysed separately using the random forest package (69,70) using R software (71). For the Mali cohort, since the features tested were more than the serum samples, they first excluded collinear responses (Pearson correlation coefficient >0.8) to focus on features providing a high predictive power (72). The misclassification and model performance was tested using the so-called out-of-bag (OOB) errors and a five-fold cross-validation cross-validation for the Kenyan and Mali cohorts, respectively. Receiver operating characteristic (ROC) curve and area under the curve (AUC) were used to assess the model predictive accuracies (47).

Most of these studies which have utilised the machine learning techniques have not compared the performance of different machine learning techniques. Moreover, the models have not considered utilising machine learning approaches with random effects and handling imbalance data problems by applying the synthetic minority oversampling technique (SMOTE). An imbalance occurs when one or more categories have very low proportions in the training data as compared to the other categories which might be true for malaria-based studies (73).

CHAPTER 2 : METHODOLOGY

2.1 Study population and design

We used data from the South-South Malaria Antigen Research Partnership (SMART) hosted at KEMRI-Wellcome Trust, Kilifi, Kenya. The data comprises fourteen cohorts from different endemic sites in Sub-Saharan Africa (SSA). The cohorts comprised largely children (ages 0 – 10 years; N=6825), as well as older children and adults (N=1297). The data includes processed lab data and epidemiological data shared from the multiple malaria-endemic locations in Africa. Data on the clinical episodes of malaria is therefore prospective with respect to the blood sample collection at the beginning of the cohort study as shown in Figure 1 below. This is a retrospective study that uses serum samples that were originally collected for prospective longitudinal cohort studies of malaria immunity.

2.2 Epidemiological data

Data from a single cross-sectional plasma sample from each participant was collected either at the start of a malaria transmission season, at birth, or at one year of age. After the cross-sectional sample, longitudinal data on the occurrence of malaria is done for six months to a maximum of one year (except in samples from Senegal which was a passive followup) following blood sampling. Data on the clinical episodes of malaria is therefore prospective relating to the blood sample collection as illustrated in the figure below.

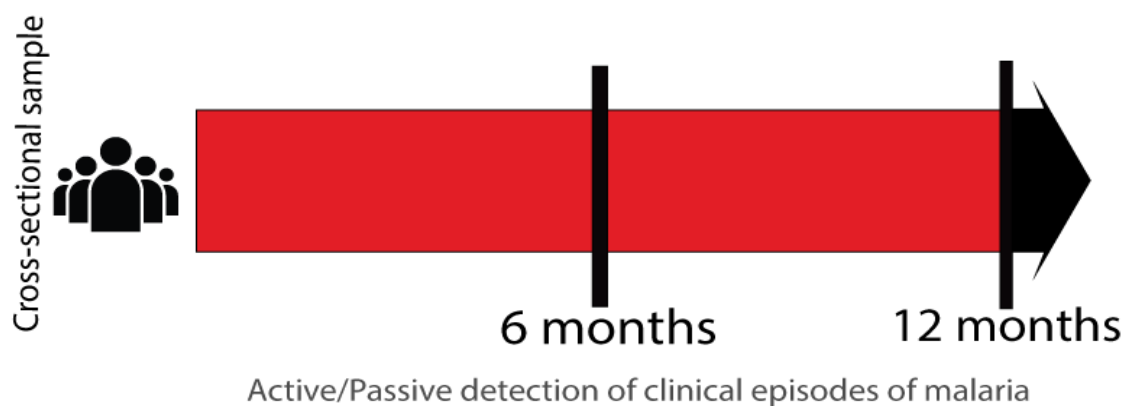


Figure 2-1: Epidemiological design of the cohort studies

Table 2-1: A summary of the study sites

Study Site	Malaria transmission	Samples Included	N	Age	Longitudinal surveillance	Case definition (T°C) +
Kombewa; KENYA	EIR 237-288	2003-2004	-273	12-47m	Active, weekly; Passive	bi- ≥37.5°C + >50,000 parasites/μl
Kilifi; KENYA	EIR 22-53	2005-2017	3607	0-14yrs	Active, weekly; Passive	≥37.5°C + ≥2500 parasites/ μl
Nyamisati; TANZANIA	Moderate EIR unavailable	1994-1995	323-304	0-30yrs	Passive	≥37.5°C + ≥5000 parasites/ μl
Muheza; TANZANIA	EIR 400	2002-2005	294	1yr	Active, monthly; Passive	Illness + any parasitaemia
Bandiagara; MALI	EIR 25	2008	189	1-5yrs	Active, weekly; Passive	≥37.5°C + ≥2500 trophozoites/mm ³
Sotuba; MALI	EIR <15	2008-2009	245	1-4yrs	Active, weekly; Passive	≥37.5°C + ≥2500 trophozoites/mm ³
Samako; MALI	EIR < N/A	2011-2013	197	1-15yrs	Active, weekly; Passive	≥37.5°C + ≥2500 trophozoites/mm ³
Dielmo; SENEGAL	EIR 295.5	2002	253	1-80yrs	Active, daily; Passive	≥37.5°C + age-dependent pyrogenic threshold

Ndiop; SENEGAL	EIR 17.9	2002	245	1-80yrs	Active, daily; Passive	≥37.5°C, illness + >30 trophozoites/100 leucocytes
Kintampo; GHANA	EIR 269	239- 2008- 2011	906	Birth-1yr	Active, monthly; Passive	≥37.5°C + any parasitaemia
Balonghin; BURKINA FASO	EIR 44.4/month	2007	242	0 -5yrs	Active, bi- weekly	≥37.5°C + any parasitaemia
Kanungu; UGANDA	EIR 310	2011- 2012	352	0.5-10yrs	Passive, Active monthly	3 ≥37.5°C + any parasitaemia
Jinja; UGANDA	EIR 310	2011- 2012	287	0.5-10yrs	Passive, Active monthly	3 ≥37.5°C + any parasitaemia
Tororo; UGANDA	EIR 310	2011- 2012	290	0.5-10yrs	Passive, Active monthly	3 ≥37.5°C + any parasitaemia

2.2.1 Kombewa, Kenya

We will utilize data from Kombewa in Kisumu County of Nyanza region in Kenya. Kombewa is a high transmission malaria site with two seasonal peaks during the long (March to June) and short (November to December) rains. The monthly incidence of clinical episodes of malaria defined as fever $\geq 37.5^{\circ}\text{C}$ and parasitemia exceeding $\geq 50,000$ parasites/ μL . The study was done in 2003-2004 to as part of a randomized, double-blind, Phase IIb, a comparator-controlled trial of the FMP1 (MSP142) malaria vaccine. The participants were followed up for six months post-vaccination. The aim was to evaluate the safety, immunogenicity, and efficacy in 400 children aged 12-47 months, with follow-up to 6 months post-vaccination. The vaccine was not effective (74).

2.2.2 Nyamisati, Tanzania

The data used is from ongoing malaria surveillance since 1985 in Nyamisati village, in the Rufiji delta of the Tanzanian coast. Malaria was holoendemic with perennial transmission and some seasonal fluctuation for the period used. Clinical episodes of malaria were

defined as >400 parasites/ μl with reported or confirmed fever (temperature $\geq 37.5^{\circ}\text{C}$) within the last day (with no other signs of an alternative cause for the fever). Data from this project are drawn from study participants in the years 1994 and 1995 and include ~ 218 children below the age of 10 years and ~ 360 older children and adults.

2.2.3 Muheza, Tanzania

Data from this site was originally collected from mother-infant pairs recruited at Muheza Designated District Hospital (DDH) in North-eastern Tanzania in the Mother-Offspring Malaria Studies Project (MOMS). Malaria transmission is intense and occurs throughout the year but is higher from May to October and lower in November to April. Clinical episodes of malaria in infants were defined as parasitemia with any signs or symptoms suggestive of malaria, most commonly fever (temperature $\geq 37.5^{\circ}\text{C}$). We had approximately 300 children aged between 8 months to 14 months (75).

2.2.4 Kilifi, Kenya

The study was conducted in the Junju sub-location of Kilifi county as a vaccine trial on children. Malaria transmission occurs at low levels throughout the year with seasonal high transmission peaks. Malaria was defined as fever (temperature $\geq 37.5^{\circ}\text{C}$) and parasitemia of >2500 parasites/ μl . Children from this study were consented to and recruited into a larger project “Integrated studies of natural immunity to malaria in children in Kilifi district” where they have been continuously monitored for clinical episodes of malaria. We use samples and data from approximately 300 children annually kept under surveillance between 2005 and 2017 (76).

2.2.5 Bandiagara, Mali

We utilize data from 189 children monitored actively in a longitudinal study to validate putative asexual blood-stage antigens as vaccine candidate molecules using standardized assays. The study was conducted in Bandiagara, a Savannah-Sahel area of Mali where malaria transmission is strictly seasonal, occurring between July and December each year (77).

2.2.6 Sotuba, Mali

Sotuba is a peri-urban area on the outskirts of Bamako and the banks of the Niger River with approximately 2600 inhabitants. Malaria is hypo endemic, highly seasonal with transmission occurring between June and December every year. For this study, we utilize data from 245 children in the control arm of a vaccine trial with 400 children (78).

2.2.7 Samako, Mali

Data from this site was from samples originally collected for a study of the effect of malaria infection on host macrophages. The study was conducted in Samako, a peri-urban area approximately 70km South of Bamako, Mali. Malaria transmission is hyperendemic but seasonal and runs from June to December each year. A total of 210 children were recruited into the study in July 2011 and were monitored over two consecutive malaria transmission seasons (to the end of 2012). A clinical episode of malaria was defined as fever (temperature $\geq 37.5^{\circ}\text{C}$), accompanied by a parasitemia of ≥ 5000 trophozoites/mm³ (79).

2.2.8 Dielmo, Senegal

Surveillance for malaria episodes has been ongoing since 1990 to identify all febrile episodes and investigate the relationship between malaria host, vector and parasite in this site in Senegal. Malaria transmission in Dielmo was intense and perennial but has declined over this period from an Entomological Inoculation Rate (EIR) of 142.5 bites per person per year in 1990 to 0.3% in 2012. We utilise data from 253 individuals recruited in July 2008. Following sampling, study participants were monitored for clinical episodes of malaria; defined as a fever (temperature $\geq 37.5^{\circ}\text{C}$) associated with a parasite density exceeding an age-dependent pyrogenic threshold determined locally, however, for this study we used any parasitemia (80).

2.2.9 Ndiop, Senegal

Ndiop is a village of approximately 350 inhabitants in Salaoum, in the Sahelo-Soudanian region of Senegal. It is located only 5km from Dielmo (above) but has a markedly different malaria transmission pattern, described as meso-endemic, and strictly seasonal occurring between July and October each year. We utilize data from a cohort comprising children and adults. Their samples were collected in July 2008, just before the start of the malaria transmission season. Surveillance for clinical malaria was active and involved daily visits (except Sundays) to the participant's homes (81,82).

2.2.10 Kintampo, Ghana

We utilize one-year follow-up data from 906 participants collected to compare the incidence of malaria in children born to primigravidae or multigravida mothers, with or without placental malaria. The data for this study was conducted in Brong Ahafo region

of Ghana, where malaria transmission intensity is high and perennial, but peaks between April and October. The area has an EIR rate of 269 infective bites per person per year and children under five experience 4.3 attacks of clinical malaria every year. Clinical episodes of malaria were defined as fever plus malaria parasitaemia of any density (83,84).

2.2.11 Balonghin, Burkina Faso

The data from this site was generated from samples that were originally collected as part of a study to describe the epidemiology of malaria before the conduct of drug and vaccine trials. The study was conducted in the Saponé health district, 40km south of the capital Ouagadougou. Malaria transmission is markedly seasonal and intense during the rainy season from June to September when the EIR is estimated at 44.4 infective bites per person per month. The ~400 participants were monitored fortnightly for clinical malaria. Clinical malaria was defined as a fever (temperature $\geq 37.5^{\circ}\text{C}$) and any parasitaemia, but analyses were done using the alternative parasitemia thresholds of $\geq 5000/\mu\text{l}$ (85).

2.2.12 Jinja, Kanungu and Tororo, Uganda

We used data that were originally collected as part of a study aimed at describing the epidemiology of malaria in Uganda. The study was conducted in Jinja, a peri-urban area with the low transmission in South-eastern Uganda; Kanungu, a rural area with the moderate transmission in western Uganda; and Tororo, a rural area with a higher transmission than others in eastern Uganda close to the border with Kenya. Malaria transmission is high and intense with two seasonal peaks, May to June and October to December. Surveillance for episodes of clinical malaria was passive, with children

presenting to designated free medical clinics. Also, routine evaluations were carried out every three months for two years (86).

2.3 Laboratory data

Cross sectional serum samples from the above study sites were processed with a standardized *Plasmodium Falciparum* micro array chip. Protein arrays allow for higher throughput at a significantly reduced cost, requiring smaller quantities both antigen and serum. A functional protein micro array chip was used here to facilitate high throughput multiplexed antibody detection from the individual samples. Antigens expressed on or associated with the surface of *P. falciparum* merozoites proposed to be key targets of blood stage immunity against malaria and were selected for this study. Specifically, 115 distinct and correctly folded merozoite antigens were expressed with the same laboratory reagents and assays (54).

The individual samples were processed using slides divided into twenty-one mini-arrays. A mini array is defined as the region allocated to a discrete sample and can be further divided into blocks. The *Plasmodium falciparum* proteins were printed in triplicate on each mini array. The slides were scanned and analysed using the Genepix 4000B scanner. The GenePix Pro & Microarray Acquisition and Analysis Software (Molecular Devices) with a Photo-Multiplier Tube (PMT) gain of 400 and a Scan power of 100 was used for the acquisition and amplification of the fluorescent signal. The quantification software captures multiple parameters in relation to antibody intensity in a pre-prepared template mostly referred to as a “.gal” file and stores the data in a “.gpr/.txt” file (54,87).

2.4 Case Definition methods

Many biomedical assays are expected to classify samples into two groups according to whether some output (i.e., parasite density or optical density) exceeds a given cut-off. The two groups are mainly cases (clinical disease positive) and non-cases (clinical disease negative). However, some of the diseases are faced with bias due to differential misclassification (2,17). In malaria, for example, to select important biological features or evaluating the effectiveness of different vaccine interventions, non-biased outcome definitions of malaria are needed. Using parasite density as a function of fever has been endorsed by the WHO Study Group on Measures of Malaria Vaccine Efficacy for obtaining case definitions for use in pivotal trials of malaria vaccine (88).

2.4.1 Classical case definition

This approach uses the parasite prevalence in febrile (children with fever) children p_f compared to the children without fever, afebrile children p_a using the population attributable fraction formula,

$\lambda = p(R-1)/R$; where p is the proportion of the exposed and R is the relative risk of the disease associated with the exposure.

Thus, the malaria attributable fraction λ taking $p = p_f$ and $R = \frac{p_f(1-p_a)}{p_a(1-p_f)}$ is estimated as;

$$\lambda = \frac{(p_f - p_a)}{(1 - p_a)} \quad (2.1)$$

However, this approach is limited in areas of high malaria transmission since the proportion of afebrile children can be low leading to imprecise estimates of λ with very wide confidence interval limits. Furthermore, the classical approach does not account for the non-monotonicity relationship between fever and parasite density (18,21).

2.4.2 Logistic power model.

The distribution forms of $p_a = \theta(x)$ and $p_f = \phi(x)$ are known and parametric regression approaches have been proposed to estimate the malaria attributable fraction (MAF), λ .

Overall, the proportion of cases over controls; $f(x)/g(x)$ is assumed to be a monotone increasing function of the parasite density x . For example, it makes biological sense to assume that the malaria attributable fraction increases with increasing parasite density. Logistic regression has been proposed to be the best parametric regression to fit this constraint. The model of the form $\text{logit}(\pi_i) = \alpha + \beta x_i$ where π_i is the probability that observation with parasite density x_i is a (fever) case is used. To maximise the likelihood estimation and assign weight to high parasite density values, a nonlinear logistic regression of the form $\text{logit}(\pi_i) = \alpha + \beta x_i^r$ was developed by Smith et. al. (18,53).

The malaria attributable fraction (MAF), λ , is then estimated using the slope coefficient of the logistic regression $\lambda = (1/N) \sum_1^i (R_i - 1) / R_i$; where $R_i = \exp[f(x_i)]$ and N is the total number of test samples (18,53).

Additionally, the model can also be used to estimate the probability of individual case of fever being attributable to malaria using $\lambda_{ind} = (R_i - 1)/R_i$ (53).

2.4.3 Non-parametric monotone regression

The monotone regression proposes dividing data into K ordered categories of the parasite density $x(k = 0, \dots, K - 1)$ and count the number of samples $\mathbf{n} = (n_0, n_1, \dots, n_{K-1})$ and the non-fever individuals $\mathbf{m} = (m_0, m_1, \dots, m_{K-1})$ in each category. This is followed by estimating category-specific attributable fractions $\lambda_k = R_k - 1/R_k$ from the category-specific $R_k = \exp(\beta_k - \beta_0)$. The overall $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_{K-1})$ is estimated by the weighted mean $\frac{1}{N} \lambda^T \mathbf{n}$.

This estimation procedure has two major limitations, one, the $k=1$ category lambda is assumed to be zero, $\lambda_0 = 0$, and the maximum likelihood estimates for category-specific and overall attributable fraction are not estimated (18).

2.4.4 Latent class model.

This method was proposed to overcome the limitations of non-parametric monotone regression using the expectation-maximization (EM) algorithm. Like the non-parametric approach, the data is divided into K ordered categories and $I_k = \text{Bin}(n_k, \lambda_k)$ are the binomial distributed true positives among the test samples in group k . Likewise, m_k are the number of controls in category k . Then, $\mathbf{m}$ has a multinomial distribution

$Mn(\sum m_i, p_0, \dots, p_{k-1})$ where $p_k = \frac{n_k(1-\lambda_k)}{\sum n_k(1-\lambda_k)}$ is the probability that the non-febrile controls belong to the k^{th} group.

Then λ_k is estimated iteratively from the likelihood of $L(\mathbf{I}, \mathbf{m} | \lambda)$ which is the product of the multinomial likelihood of the controls and binomial likelihood of the latent data written as;

$$\log L(\mathbf{I}, \mathbf{m} | \lambda) \propto \sum [m_k \log p_k + I_k \log \lambda_k + (n_k - I_k) \log(1 - \lambda_k)] \quad (2.2)$$

The details of the iteration procedure are discussed in detail by Smith et al (18). After the iteration procedure the, λ is finally estimated as a weighted average of the λ_k 's, $\frac{1}{N} \lambda^T \mathbf{n}$ (18).

2.4.5 Bayesian latent class models

Implementing the EM algorithms have been reported to be challenging since they sometimes converge to a local minimum and may have saddle-point problems (17,89). This led to utilisation of the Bayesian approach to mixture distributions to overcome these limitations. The same as the EM approach, the data is ordered into categories. Each category has two components; thus the likelihood is derived as a product of two multinomial distributions.

One component $g_1(.)$ corresponds to non-malaria fever episodes and the other component, $g_2(.)$ to children with clinical malaria episodes (fever and parasites) (18). Observation from $g_1(.)$ are assumed to be smaller than the other and some overlap

between the two groups. Parasite levels during the cross-sectional bleed were available and are used as the training sample, i.e., a sample that comes from the component of the mixture corresponding to children without fever but who may have parasites. The data was divided into K ordered categories over the range of the of the parasite density $\mathbf{X}$. This was followed by counting the number of test samples $\mathbf{n} = (n_0, n_1, \dots, n_{k-1})$ and control samples (non-fever cases) $\mathbf{m} = (m_0, m_1, \dots, m_{k-1})$. Then the MAF, λ , was then estimated from the two multinomial distributions,

$$\begin{aligned}\theta_i &= P(x \in \text{category } i \mid P_1), \\ \phi_i &= P(x \in \text{category } i \mid P_2), \quad (2.3) \\ \lambda &= P(x \in P_2)\end{aligned}$$

The parameters P_1 and P_2 are the distributions functions of the components $g_1(\cdot)$ and $g_2(\cdot)$ respectively (17).

Then

$$(m_1, m_2, \dots, m_k) \sim Mn\left(\sum_{i=1}^k m_i, \theta_1, \theta_2, \dots, \theta_k\right) \text{ and } (n_1, n_2, \dots, n_k) \sim Mn\left(\sum_{i=1}^k n_i, p_1, p_2, \dots, p_k\right) \quad (2.4)$$

Where

$$\begin{aligned}p_i &= P(x \in \text{category } i) \\ p_i &= P(x \in \text{category } i \mid P_1) p(x \text{ from } P_1) + P(x \in \text{category } i \mid P_2) p(x \text{ from } P_2) \quad (2.5) \\ p_i &= (1 - \lambda) \theta_i + \lambda \phi_i\end{aligned}$$

The joint density of the data $\mathbf{m} = (m_1, m_2, \dots, m_k)$ and $\mathbf{n} = (n_1, n_2, \dots, n_k)$ are written as,

$$\ell(\tilde{m}, \tilde{n} | \theta, \phi, \lambda) \propto \prod_{i=1}^{K-1} \theta_i^{m_i} \left(1 - \sum_{i=1}^{K-1} \theta_i\right)^{m_K} \prod_{i=1}^{K-1} \{(1-\lambda)\theta_i + \lambda\phi_i\}^{n_i} \left\{1 - \lambda \left(1 - \sum_{i=1}^{K-1} \theta_i\right) + \lambda \left(1 - \sum_{i=1}^{K-1} \phi_i\right)\right\}^{n_K} \quad (2.6)$$

Where $\theta = (\theta_1, \dots, \theta_k)$ and $\phi = (\phi_1, \dots, \phi_k)$. A uniform non-informative prior $[0,1]$ was assumed $p(\theta, \phi, \lambda) \propto 1$. The parameters of interest are then estimated using Gibbs sampling from the posterior density; $p(\theta, \phi, \lambda | \mathbf{m}, \mathbf{n}) \propto L(\mathbf{m}, \mathbf{n} | \theta, \phi, \lambda)$ (17,90,91)

The category-specific attributable fractions are obtained using,

$$\lambda_i = P(x \in P_2 | x \in \text{category } i) = \frac{\lambda\phi_i}{(1-\lambda)\theta_i + \lambda\phi_i}. \quad (2.7)$$

2.4.6 Sensitivity and specificity of case definitions

Sensitivity and specificity of various cut-off values for parasite density can be estimated

by $n_c \lambda_c / N \lambda$ and $1 - n_c (1 - \lambda_c) / N (1 - \lambda)$ respectively where $n_c = \sum_{i=c}^K n_i$, $\lambda_c = \left(\sum_{i=c}^K \lambda_i n_i \right) / n_c$, n_i

the number of fever cases in the category i and c represents the parasite density category of which it is the selected cut-off in logistic regression or the lower bound for the category in latent class models (53).

2.4.7 Estimation of individual MAF probabilities

Several articles (92–96) have pointed out problems associated with the categorization of data. Among the problems reported are loss of information and power, but also an increased risk of type I error if continuous and poor predictive performance (92,93)

We demonstrate the utilization of probability estimates from Bayesian latent class models and logistic regression approaches over creating malaria outcome categories using a

selected parasite density cut off to perform correlates of disease protection selection. For the logistic regression approach, the probability of individual fever case being attributable to malaria was estimated using $\lambda_{ind} = (R_i - 1)/R_i$ (53).

To estimate the probability, λ_{ind} , of individual case of fever being attributable to malaria for the Bayesian latent class, local and piece-wise cubic polynomial models were used. The models we fitted using category-specific MAF, λ_c , together with the category-specific midpoint of parasite density. This was followed by the prediction of the individual λ_{ind} using their parasite density measurements from the results of various model fitting functions (91).

The local regression estimates of the function evaluated at a particular midpoint x_{k_mid} of parasite density are constructed by fitting a weighted least squares regression model to the λ_c values corresponding to the x_{mid} midpoint values that are given x_{k_mid} . During the estimation process, Markov Chain Monte Carlo allows the calculation of interval estimates of functions. An interval estimate for the prevalence of malaria fever was then obtained and used to adjust the Bayesian individual probabilities for non-febrile cases (17,91).

2.5 Protein microarray data processing

Data generated using a protein microarray has been shown to contain systematic and non-systematic sources of bias (37,38,57). Systematic biases refer to artefacts that may occur due to experimental and handling procedures. Non-systematic biases on the other hand refer to random errors that may occur due to slight variations during the experiment. Careful pre-processing is essential to eliminate this technical bias while

keeping the biological variation. This bias has been attributed to nonbiological sources, which is introduced by small variations in the experimental conditions during the experiment process. Guidelines for pre-processing the protein microarray data before analysis, continue to evolve. Pre-processing involves assessing the gain scanning parameters (97); background correction (57), quantifying within-sample variation, batch correction (98,99), normalisation (100), among other steps.

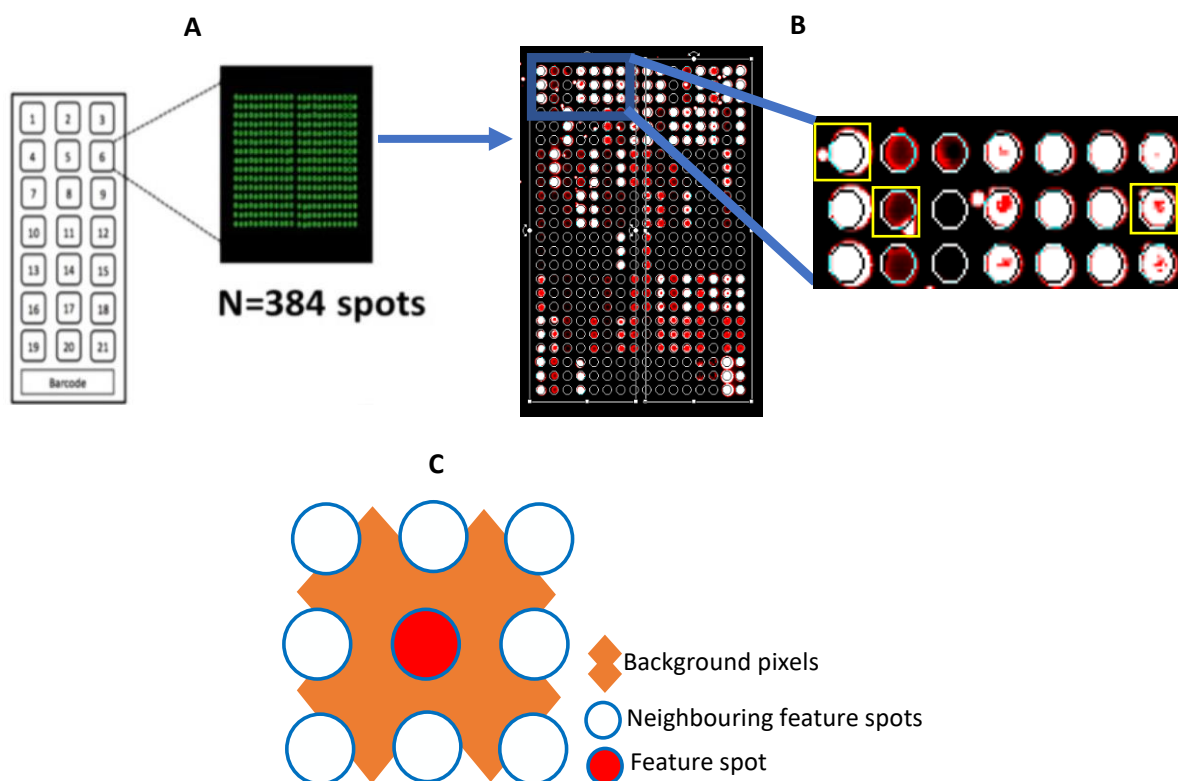


Figure 2-2: Background correction: artefacts add noise to the signal intensity A) A microarray slide with 21 mini arrays and a barcode. Each mini array has a specific number of features represented by a spot. B) Artefacts [spots surrounded with yellow boxes] Kamuyu 2018. C) The total foreground intensity associated with feature spot typically includes the background MFI

2.5.1 Gain scanning for protein microarray

In protein microarray, systematic spatial artefacts can be introduced during the array printing and systematic scanning bias if the alignment of the scanner is not well aligned. The efficiency of antibody signal detection varies across different platforms. To quantify

the antibody measurements, protein microarray relies on fluorescence technique combined with a Photo-Multiplier Tube (PMT) system and razor power responsible for amplification of the fluorescent signals. Researchers adjust the PMT gain to have a balance between signal bleaching and enhancing lower signals. Antibody detection systems that rely on fluorescence require careful calibration to minimize background signals arising from the non-specific binding of analytes or spill-over fluorescence from reference materials or high-signal intensity neighbouring spots (38,97,101).

Scanning laser power and the PMT gain adjustment are a significant source of non-systematic and systematic variability (38) and should be investigated to select the optimal setting. To date, however, there is no agreement on how to determine the optimal settings of microarray scanners and this has to be done empirically. One of the suggested approaches is to scan the slides under different combinations of the laser power strengths and PMT levels and investigate graphically which settings give the optimal Mean Fluorescent Intensity (MFI) quantification. If the gain is set too high, highly immunogenic features, including the control spots will saturate and vice versa if the gain is decreased; .i.e. the sensitivity is lost (102). The Signal to Noise Ratio (SNR) is the preferred measure to use to perform the graphical investigation. The ratio is estimated as $\frac{FG - BG}{sd(BG)}$, where FG is the overall foreground signal, BG is the background signal (102,103). A signal-to-noise ratio of 3 is regularly considered the lower limit for accurate quantification since a higher SNR indicates a higher signal over background noise. The second proposal is to quantify the data under multiple PMT settings and then combines these data during

analysis using a power function described in detail by Feng et. al (97). However, this approach can be costly and time consuming for huge experiments.

2.5.1 Background correction

Background intensity is the signal that is emitted by non-biological sources among other things than sample hybridized to probe or due to nonspecific hybridization (104,105). Background noise is caused by non-specific fluorescence such as the auto-fluorescence of the glass slide or non-specific binding of analytes and reference substances (38). Foreground is the total spot intensity including the background. It is therefore important to handle the background noise to adjust the data from the ambient signals for each spot. There are various methods of background correction.

2.5.1.1 Local and global background subtraction

This method involves subtracting the median of the background pixel intensities from the foreground signal of a specific spot. The local background signal is considered as an unbiased estimate of the non-specific fluorescence around each spot. Therefore, subtracting from the foreground intensity gives an unbiased estimator of the true signal (105). However, in protein microarrays, local background regions can be easily skewed by artefacts such as small speckles (101). The global background subtraction method is provided as an option and involves subtracting a single global background value from each spot. The global value is the median of all the local background values for a given array.

2.5.1.2 Moving minimum background subtraction

This approach involves subtracting the minimum of the local backgrounds for a defined grid. In this tool, the background intensity is replaced by the minimum of the local backgrounds in each block. This is the method of choice when a mini array artefact is observed in the background intensities (57).

2.5.1.3 Neighbourhood background correction

This approach is similar to the moving minimum, however, a three-by-three spot surrounding region is allocated to each protein spot, and the median background of the nine spots subtracted from the central spot (101).

2.5.1.4 Half moving minimum background correction.

The half moving minimum is a novel approach and involves a two-step background correction and is designed to ensure that the background-corrected signals are positive. First, the local background correction is implemented. Secondly, any intensity which is zero or negative following local background subtraction is set to half the minimum of the positive corrected intensities in each block. We created this method by combining the half and minimum background approaches in Limma (57).

2.5.1.5 Log-linear interpolation background correction (Edwards)

This is also a two-step strategy for background correction adopted from one-channel CDNA microarray experiments and provides an alternative approach for dealing with negative background corrected intensities. Following local background subtraction (above), this method adjusts the negative intensities with a log-linear smooth monotonic function when the difference between the foreground and the local background is higher

than a small threshold value. Where the background-corrected value is larger than the threshold value, it is considered to be the true signal of the spot. This method has advantage over the other approaches since it applies a mathematical function on the whole range of data to estimate the threshold value (106).

2.5.1.6 Normal and exponential model (normexp) background correction:

This approach fits a convolution model of normal and exponential distribution for the background signals and background-corrected signals respectively (107). The background intensities are modelled as a covariate and the expected signal given the observed foreground becomes the corrected intensity. This results in a smooth monotonic function of the background-subtracted intensities such that the corrected intensities are all positive (108). For this approach, the shape of the foreground and background must present the distribution characteristics.

2.5.2 Coefficient of variation

Most of the microarray experiments are designed with technical replicates which allow for testing of within-sample variability. Technical replicates help in the reduction of data loss, e.g., if one of the technical repeat spots performs poorly, the other spots may succeed. Additionally, they help in identifying sample-specific experimental variability and allow for potential outlier recognition and removal. Coefficient of Variation (CV) is used to evaluate within-sample variability. Coefficient of variation (CV) is a measure of accuracy expressed in percentage, and is defined as; $CV = \left(\frac{\sigma}{\mu} \right)$, where σ is the standard deviation of the signal divided by the mean of the signal (38). The %CV is one of the

good measures of data reliability and increases with big differences among the replicates (109). On top of that, we developed functionality to identify and filter outliers within the replicates for experiments with more than two technical repeats. The function sequentially calculates the CV's for k-combination of the replicates for each sample per feature and identifies and removes the outliers; replicas with the highest CV among the k-CV's is calculated (110).

2.5.3 Normalisation and Batch correction

Data normalization minimizes the mean-variance dependence that is commonly observed in microarray experiments and may mask true biological variability (37,111). Notably, protein microarrays often use only one colour channel and, therefore, are especially prone to spatial artefacts. Furthermore, many data processing methods for DNA microarray are based on balanced intensities between the two-colour channels. The 2 channels provide inherent controls that are utilised to account for spatial artefacts. However, such controls are missing in one-colour-channel protein microarrays thus we cannot directly apply the DNA approaches in protein micro array experiments. Various normalisation techniques have been proposed and we highlight the most used approaches. The log, variance stabilization normalization (VSN), robust linear model (RLM) and cyclic loess are discussed in detail in Chapter 4.

2.5.3.1 *Log2 Normalisation*

Log transformation reduces the bias in variance between high and low values by altering the distribution of the data moving from a multiplicative function to an additive one. Wider variance is typically observed for higher signal intensities.

The background-corrected values $\hat{y}_{ijkr} = y_{ijkr} - \hat{\alpha}$; where $\hat{\alpha}$ is the estimate of the median background level and y_{ijkr} is the estimated intensity for a spot r of protein feature k in the mini-array j on the i^{th} slide. The transformation is defined as $\log_2(\hat{y}_{ijkr}) = \log_2(y_{ijkr} - \hat{\alpha})$.

2.5.3.2 Global and quantile normalization

Global and quantile normalisation techniques are mostly used in DNA micro-array. Global normalization uses the global median of log intensity ratios as the scaling factor to every value in the array ensures all the features have the same median intensity. On the other hand, quantile normalisation arranges per feature intensities from the highest to the lowest. This is followed by replacing the highest value in feature with the mean of the highest value, the second-highest value becomes the mean of the second-highest values and does this for every rank. Global and quantile normalisation techniques assume a majority of the intensities are not differentially expressed (38).

2.5.3.3 Two-parameter signal normalization

This approach assumes that signal distribution in a grid is consistent across the slide and therefore a spot can be normalized by its neighbouring spots. The median and the median absolute deviation are estimated to represent the distribution of the signals in a grid. This is followed by averaging the estimated measure of central tendency and variation. The average estimates normalise the central spot signal to align with the global distribution of the spot signals. For a given spot, x_{ij} , the signal S_{ij} is normalised as

$\hat{S}_{ij} = \overline{MED} + (S_{ij} - MED_{ij}) \frac{\overline{MAD}}{MAD_{ij}}$, where $\overline{MED}$ and $\overline{MAD}$ are the averaged neighbourhood

medians, MED_{ij} and median absolute deviations MAD_{ij} respectively (62,101).

2.5.3.4 Mean-Variance Regularization

This is a non-parametric method for joint adaptive mean-variance handling using joint regularized shrinkage estimators. The shrinkage estimators take advantage of parameter similarity of measurements done in parallel. The method specifically uses a joint; cluster pooled parameter estimates, and local regularization of the mean and variance parameters to stabilize the mean-variance dependence (112).

2.5.3.5 Cyclic Loess Normalization

Cyclic loess normalization utilises the MA plot and local regression (LOESS) smoothing to transform the data. MA plot in single-colour microarrays is the scatter plot of average log intensity values [A] from both arrays vs. difference in log expression values [M] of the same arrays (113). In the first place, the approach estimates the intensity-dependent differences and then regulates the differences by centering the non-linear local regression (LOESS) line to zero. This approach is suited for data where the mean-variance dependence is non-linear (114). Cyclic loess is defined for both raw and log-transformed intensities. Log cyclic loess is utilised when we have a high number of outliers.

The "fast linear loess" method of Ballman et al. is mostly utilised in protein microarrays since it is relatively fast and specifically defined for a single colour array. It does this by running systematic cycles of smooth-LOESS to normalize each pair of columns in reference to a constructed 'average' array. The intensity level of a spot on the 'average'

array is equal to the average intensity of spot r across all arrays (57,113). In summary, the data is first arranged in a matrix with one column per probe from a given slide, then an average slide is created and finally, the cyclic loess transformation is done using fast linear loess (113).

2.5.3.6 Robust Linear Normalization (RLM)

The assumption of the global normalization method like Cyclic Loess is that features across different samples have a similar signal distribution. However, this has been reported not to be always true for protein microarray data(38). Additionally, the number of reactive proteins and their reactivity can vary significantly from sample to sample, leading to markedly different signal distributions. This led to the implementation of Robust Linear Model (RLM) normalisation by Sboner et al. to reduce both intra-array and inter-array technical variability specifically for protein micro arrays. This approach incorporates a set of control proteins in the mini arrays to be used as normalization factors (38,46).

2.5.3.7 Variance Stabilization Normalization (VSN)

VSN was mainly developed to handle the shortcomings of the log transformation of stabilizing the variance of the low and high intensities over the full range of the data. The proposed approach calibrates for sample-to-sample variations through shifting and scaling and transforms the signals to a scale where the variance, $Var(\hat{y}_{ijk_r})$ is independent of the mean, $E(\hat{y}_{ijk_r})$. This method is ideal with a non-linear mean-variance dependence and is defined for negative intensities. Huber et al. and Durbin et al. independently

proposed the VSN approach which is a variant of the log-transform (asinh). This variant transforms and corrects the inflated variance in low signal intensities (37,100,111).

2.5.3.8 Batch Correction

Samples are typically processed on separate days for a variety of technical and logistical reasons, including the need to rerun samples with unsatisfactory replicate data. Multiple factors contribute to this batch variation such as non-specific day to day differences in laboratory conditions or operators (98,115). Batch correction involves investigating the presence of batch effects using coloured scatter plots, hierarchical clustering, or principal component analysis (PCA). Once identified, subsequent analyses can be adjusted to account for batch effects. A few methods and tools have been developed for microarray experiments (116). For example, single value decomposition (SVD) and principal component analysis (PCA) where the PCA value of the batch effect is subtracted from the data, Distance-weighted discrimination (DWD) which is a modified method based on Support Vector Machines (SVM) algorithm and Mean centering (PAMR) which is a feature-wise based on one-way analysis of variance (ANOVA), Combat which utilises empirical Bayes approach, among others. However, the Empirical Bayes approach is flexible to either small or large batches and can incorporate both parametric and non-parametric data(98,117).

2.6 Feature selection

Feature Selection in machine learning is the process where you automatically or manually select parameters that explain the variation in your outcome variable the most. In health sciences, one of the fundamental motivations for feature selection is to find the most

important correlates that explain the disease outcome. The more correlates of disease protection measured, the larger the hypothesis space. Therefore, feature selection can efficiently reduce the hypothesis space by less relevant and redundant features. The smaller the hypothesis space, the easier it is to find the important correlates of disease. In addition, while building prediction models, finding the most important predictor variables that explain most of the variation in the response variable is key to identify and build high performing models (118,119).

Identifying the important features can be achieved either by ranking the predictor variables according to some criteria and select the top features or selecting a subset of the features empirically without learning performance declining (119).

2.6.1 Modified Poisson

This approach uses the classical regression approach with a categorical outcome adjusting for known confounders followed by ranking the important features using the risk ratio estimated. The modified Poisson regression approach proposed by Zou et. al. is a Poisson model with a robust error variance with a dichotomous outcome. Modified Poisson regression has an advantage over logistic regression in epidemiological studies since it estimates the relative risk directly rather than the odds ratio which is estimated by the latter (43).

When Poisson regression is applied to a dichotomous outcome, the error of the relative risk calculated will be overestimated. However, this error is corrected by sandwich estimation of robust error variance.

Given that x_i is a binary exposure with an underlying risk function $\pi(x_i)$, then the risk function is modelled as $\log[\pi(x_i)] = \alpha + \beta x_i$ since $\pi(x_i)$ has to be positive.

The relative risk (RR) is estimated by $\exp(\beta)$ if a Poisson distribution is assumed for the outcome y_i then,

$$l(\alpha, \beta) = C \times \sum_{i=1}^n [y_i (\alpha + \beta x_i) - \exp(\alpha + \beta x_i)] \quad (2.8)$$

Applying the likelihood theory; $RR = \exp(\hat{\beta}) = \frac{an_0}{cn_1}$, and the estimated variance RR is expressed as $\text{var}(RR) = \frac{1}{a} + \frac{1}{c}$ followed by correction using the sandwich estimator explained in detail by Zou et. al (43).

With several biological features tested, the RR's are ranked according to the criteria of interest, for example from highest to lowest or vice versa and the top k features are selected. However, this "univariate" multiple testing approach ignores the potential interplay between different features (47).

2.6.2 Penalized regression

Penalized regression uses the standard least squares regression, $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$ with a penalty to fit a multivariable model involving p features/predictors. However, with many features used, some of them are not associated or do not explain the outcome and leads to unnecessary complexity in the model. The

penalized regression is preferred over least squares regression due to overfitting if $p > n$ and the unnecessary complexity that comes with non-important features in the model. Thus, penalized regression imposes a penalty to shrink the coefficients of the less important variables towards zero or equal to zero. This shrinkage (also known as regularization) has the effect of reducing variance. Other methods that utilise the least-squares approach are subset selection and dimension reduction (118,120–122). The most used penalized regressions are the ridge and the lasso regression.

2.6.2.1 Ridge regression

The least-square approach estimates the coefficients using the values that minimize the residual sum of squares (RSS)

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (2.9)$$

The ridge regression is similar to the least square approach except that the coefficients are estimated by minimizing the RSS with a penalty.

$$\min_{\beta} \text{RSS} + \underbrace{\lambda \sum_{j=1}^p \beta_j^2}_{\text{ridge shrinkage penalty}} \quad (2.10)$$

The ridge regression coefficients $\hat{\beta}^R$ are the values that minimize the above equation where $\lambda \geq 0$ a tuning parameter (lambda) which determines the amount of shrinkage and requires estimation separately. The estimation is done using cross-validation discussed in detail in Chapter 5. Where lambda is zero, the penalty does not affect the coefficients,

but as $\lambda \rightarrow \infty$ the shrinkage penalty grows hence the coefficients will approach zero (118,120).

2.6.2.2 Lasso regression

The ridge penalty $\lambda \sum_{j=1}^p \beta_j^2$; l_2 penalty, will shrink the coefficients towards zero but does not set any of the coefficients to be zero. This is a challenge in feature selection but may not be a problem in prediction models. The lasso overcomes the disadvantage of ridge regression by using a different penalty $\lambda \sum_{j=1}^p |\beta_j|$, mostly referred to as the l_1 penalty. Thus, the lasso coefficients, $\hat{\beta}^L$, are estimated by

$$\min_{\beta} \text{RSS} + \lambda \sum_{j=1}^p |\beta_j| \quad (2.11)$$

lasso shrinkage penalty

As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression hence the former is mostly used for variable selection. As in ridge regression, selecting a good value λ for the lasso is also important. The elastic net approach which is a combination of ridge and lasso regression can also be applied (118,120).

It has been empirically observed that the performance of the lasso for variable selection works well in the presence of low-to-moderate multicollinearity among the predicting

features. Moderate to high multicollinearity can be quite common in modern-day applications, it is important to consider this during variable selection (123,124).

2.6.3 Bayesian regularization methods

Motivated by the limitation of lasso in 'high multicollinearity' settings, variants of Bayesian lasso have been proposed for computing the lasso solution (123,124). The Bayesian approach assumes the coefficient vector β has some prior distribution, $p(\beta)$. The likelihood of the data can be assumed to be $f(y|x, \beta)$ where $x = (x_1, \dots, x_p)$. Thus, the posterior will take the form,

$$p(\beta|x, y) \propto f(y|x, \beta) p(\beta|x) = f(y|x, \beta) p(\beta) \quad (2.12)$$

With the above posterior form, assuming that $p(\beta) = \prod_{j=1}^p g(\beta_j)$ to be a Gaussian distribution with mean zero and standard deviation a function of λ or a double-exponential (Laplace) distribution with mean zero and standard deviation a function of λ , then the posterior mode for β given the data is a ridge regression or a lasso regression respectively. It was shown through Bayesian interpretation by Gareth et al. that the lasso prior peaked at zero while ridge prior was flatter at zero hence the lasso expects a priori that many of the coefficients are (exactly) zero (120).

Therefore, we note that the Bayesian variable selection approach depends on the specification of the priors in the model space. The priors as described by Matthew et al. falls into four major categories priors on the model space, spike and slab priors, shrinkage priors and projection methods (125). Here we discuss the spike and slab prior which in

its original form was a mixture of a Dirac measure concentrated at zero and a uniform diffuse component. However, it can be used as a mixture of where two distributions where one component is peaked at zero and the other is diffuse (120,125).

The spike and slab prior takes the form,

$$\beta_j | \gamma_j \sim (1 - \gamma_j) G_0(\beta_j) + \gamma_j G_1(\beta_j) \quad (2.13),$$

For $j = 1, \dots, p$ where G_0 and G_1 are probability measures on the response variable and $\gamma \sim p(\gamma)$ where $p(\gamma)$ is the prior of the model space. For example, if an independent Bernoulli prior is assumed $\gamma_j = \text{Bern}(w_j)$, and $w_j \in [0, 1]$, then the spike and slab prior can be written as

$$\beta_j | w_j \sim (1 - w_j) G_0(\beta_j) + w_j G_1(\beta_j) \quad (2.14).$$

However, for very high dimensional data, the computational cost is high for the spike and slab methods. This also motivated the development of Bayesian Lasso (least absolute shrinkage and selection) which is a spike and slab priors with thicker tails (125,126). In particular, James et al. (120) and Ročková et al. (127) considered the spike and slab of the form,

$$\beta_j | \gamma_j \sim (1 - \gamma_j) \text{Lap}(\lambda_0) + \gamma_j (\lambda_1) \quad (2.15),$$

Where $\text{Lap}(\lambda) = \frac{\lambda}{2} e^{-\lambda|\beta|}$ denotes the Laplace (double exponential) distribution as discussed above. Selecting a small λ_1 and a large λ_0 mimic the original spike and slab prior and $\lambda_1 = \lambda_0 = \lambda$ is equivalent to a single Laplace with parameter λ . The Laplace shrinkage has a better computational advantage over the other spike and slab priors.

Sometimes using the Laplace prior does not make the posterior concentrate close to the true value for large samples (128) and has led to the introduction of Bayesian priors with good shrinkage properties. Some of the priors include the normal-gamma (129), Horseshoe prior (130), generalised double Pareto (131), Dirichlet-Laplace (132) and the Horseshoe+ prior (133,134).

The broad hierarchical form of the global shrinkage priors can be expressed as

$$\begin{aligned}\beta_j | \eta_j, \omega &\sim \text{Norm}(0, \omega \eta_j) \\ \eta_j &\sim \pi(\eta_j) \\ \omega &\sim \pi(\omega)\end{aligned}\tag{2.16}$$

Where η_j are known local shrinkage parameters and control the intensity of shrinkage for each individual coefficient β_j and ω is the global parameter that affects the overall shrinkage. The parameter η_j has a distribution that can be written as (135)

$$\pi(\eta_j) = K \eta_j^{-q-1} L(\eta_j) \tag{2.17}$$

Where $K > 0$ and $a > 0$ and L is a positive measurable constant. Quantification of the selected variables is often achieved using the regression coefficients point estimates and their estimated credible intervals (125).

2.6.4 Random Forest

Random forest (RF) is a supervised machine learning technique built on the decision tree-based algorithm. Random forest aggregates a collection of random decision trees. The best split is selected from optimizing several features or predictors. The best split is the

predictor variable and cut point that minimizes the cost function of interest from a forest of decision trees. The forest is then used to identify a classification consensus by selecting the most common feature (mode) (69,118,136).

The general definition of random forest given by Breiman in 2001, let $(\hat{h}(., \Theta_1), \dots, \hat{h}(., \Theta_q))$ be a collection of tree predictors with $\Theta_1, \dots, \Theta_q$ i.i.d. random variables independent of the learning sample L_n . The random forest predictor $\hat{h}_{RF}$ is obtained by aggregating the collection of random trees.

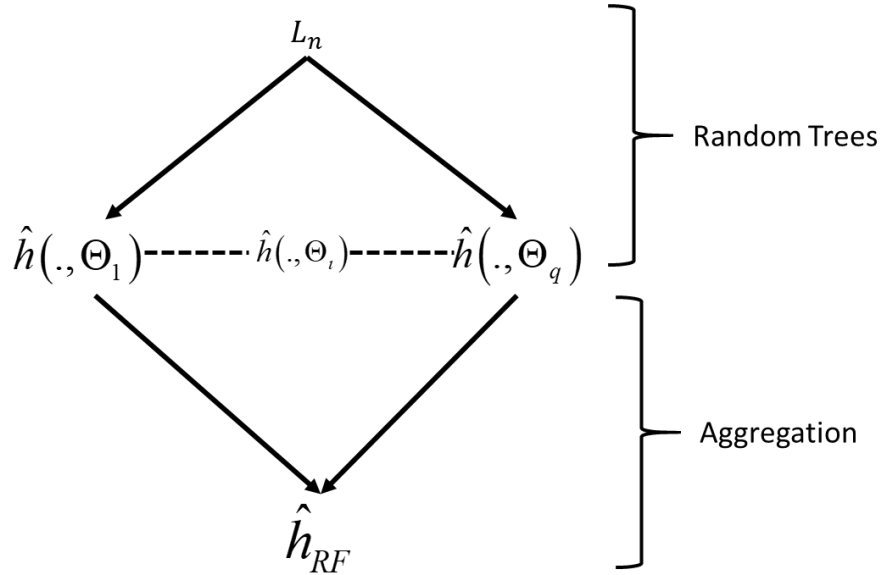


Figure 2-3: General schematic flow of the random forest algorithm (136)

$\hat{h}_{RF}(x) = \frac{1}{q} \sum_{\ell=1}^q \hat{h}(x, \Theta_{\ell})$ - average of individual tree prediction in a regression approach.

$\hat{h}_{RF}(x) = \arg \max_{1 \leq c \leq C} \sum_{\ell=1}^q \mathbf{1}_{\hat{h}(x, \Theta_{\ell})=c}$ majority vote among individual tree predictions in a classification approach (136).

To fit the RF, a bootstrap sample of a learning sample L_n of size n is obtained by randomly drawing n observations from L_n with replacement. Each observation (X_i, Y_i) of L_n having a probability $1/n$ of being selected in each draw. The process of aggregating the draws is called Bagging, an acronym for "Bootstrap aggregating". Bagging predictor is obtained by aggregating a set of tree predictors, each constructed on a different bootstrap sample. Aggregating is achieved by a unit vote for the most popular class, then combining these results get the final sort of result. An input of the number of randomly sampled candidates/features in each split the most important parameter in the RF model. This parameter is optimised using cross-validation. The random selection of features is done by a uniform draw of "mtry" features without replacement among the p available variables Figure 2-4. Therefore, the optimal "mtry" value depends greatly on the total number of variables p .

Bagging brings out an out-of bag/ Bootstrap (OOB) data which we used to estimate the OOB error and the prediction errors. Each tree had an OOB estimate for its error, and the overall generalization error of RF is the average of estimations of all tree error for every tree contained in the RF. The OOB error avoids the need for a test and train data since the samples not used in the bootstrap are used as the test data. The number of features randomly sampled as candidates at each split is an important parameter for the RF model (136).

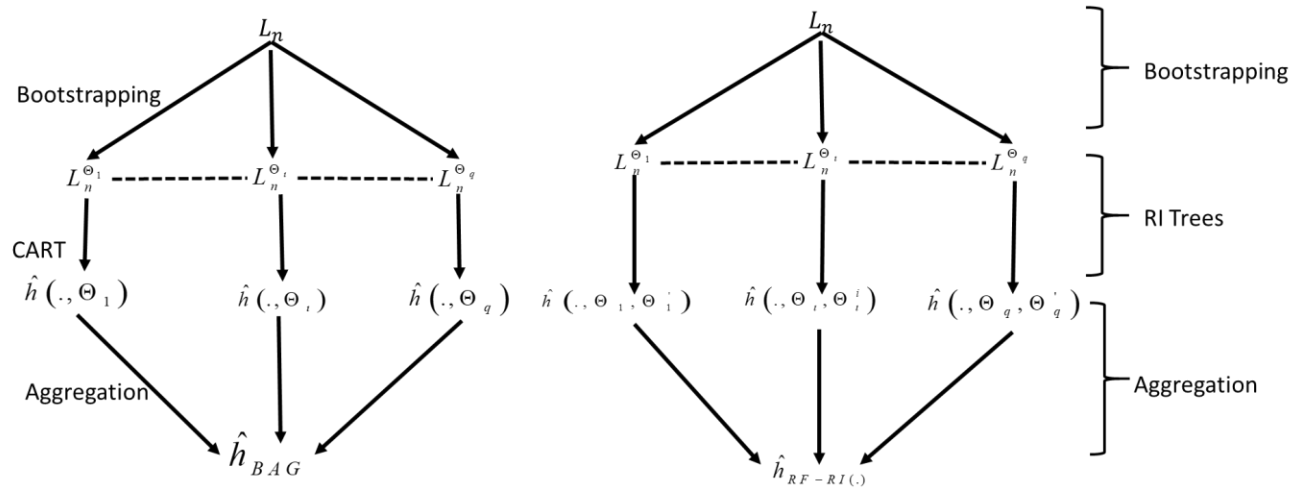


Figure 2-4: Schematic flow of Bagging on the left and RF-RI algorithm “mtry” on the right. The RF-RI stands for Random Forest Random Inputs to show the random inputs in a RF model. Where Θ refers to the bootstrap draw and Θ' refers to the random draw of variables (136)

The OOB error is estimated from predicting the Y_i of the i^{th} observation using predictors that do not include the used bootstrap sample i.e., bootstrap samples not containing (X_i, Y_i) . The OOB error is then calculated as follows in a classification approach:

$$OOB\ error = \frac{1}{n} \sum_{i=1}^n 1_{Y_i \neq \hat{Y}_i} \quad (2.18)$$

Another parameter but not very crucial is the number of trees, since the larger the number of trees, the better. However, due to computational cost, the number of trees is selected up to the point where the OOB error stabilizes. While doing feature selection with random forest, correlated features will be given equal or similar importance, but overall reduced importance compared to the same tree built without correlated counterparts.

Class imbalance, where one category has very low numbers compared to the other, have been observed to bias the OOB error towards the minimum category in classification problems. Synthetic Minority Oversampling Technique (SMOTE) has been proposed to handle class imbalance in binary classification. This technique is a sampling procedure which oversamples the class with lower numbers, by using randomly selected k nearest neighbours of the minority class followed by generating a synthetic minority class instance (73,137).

2.6.5 Boosting trees

While random forest builds an ensemble of deep independent decision trees, boosting approach grows shallow successive decision trees with each tree learning and getting better from the previous. Boosting approach will continue learning to minimize errors and this may cause overfitting by emphasizing the outliers, however, cross-validation is used to overcome this. Similarly, like random forest boosting uses Bagging (bootstrap aggregation) which involves bootstrapping the training dataset, building a discrete decision tree to each copy, and then assembling all of the trees to create a single predictive model (118,120,138).

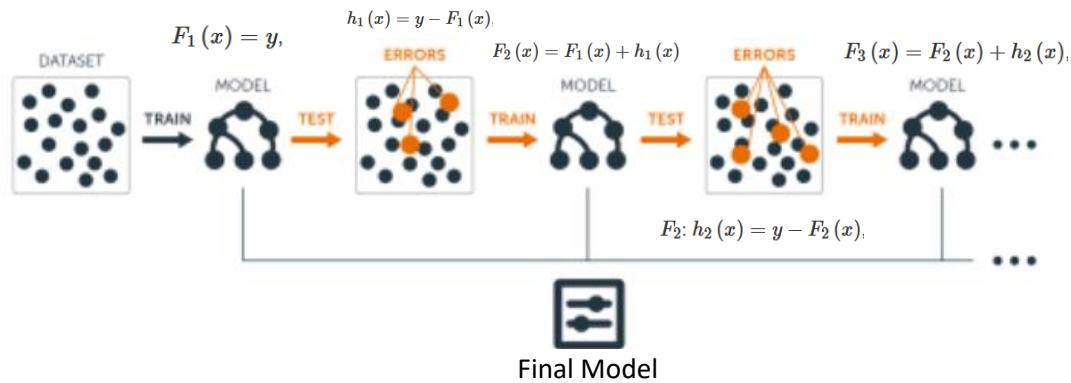


Figure 2-5: Sequential boosting approach workflow, where the h^* are the residuals: Improved from [bradleyboehmke.github.io](https://github.com/bradleyboehmke)

Recently a new algorithm called GPBoost of combining boosting and mixed effect models was introduced by Fabio et al. GPBoost (138) also allows building tree-boosting independently, inference, prediction, and selection of important features for mixed-effects models. Mixed effects model is a modelling technique for grouped data or data with repeated measures. The independence assumption made in most boosting approaches is also maintained for more efficient learning of the mean function and obtaining probability predictions. Importantly, this approach takes advantage of boosting in handling complex non-linearities and it is robust to multicollinearity in predictor variables. In this approach, the response variable is assumed to be the sum of a potentially non-linear function $F(X)$ and random effects $Z\beta$:

$y = F(X) + Z\beta + \varepsilon$, where y is the response variable, x are the predictor variables, $Z\beta$ are the random effects and follow a multivariate normal distribution and ε is an independent error term (118,138).

To build better Boosting models the following parameters are tuned using Cross Validation approach with training and test sets. The first and most important parameter is the

“nrounds” which is number of trees in a single boosting iteration. Secondly, the “learning rate” which is the shrinkage parameter that controls how quickly the model proceeds down the gradient descent (learns) whereby lower values usually lead to higher predictive accuracy, but more boosting iterations are needed. The “learning rate” ranges between 0 to 1 and also determines the contribution of each tree on the final outcome. The other parameter is the “min_data_in_leaf” which is the minimal number of samples per leaf in a boosting iteration (138,139). The values in a grid that minimize average Root Mean Squared Error (RMSE) on the test sets are selected as the optimal values.

2.6.6 sparse Partial Least Squares discriminant analysis (sPLSDA)

Partial Least Squares (PLS) originally developed as dimension reduction has been improved to be applied in classification and discrimination problems (140,141). The PLS approach uses a linear combination of the original variables to explain most of the variance/covariance structure of the data (142). A variant of sparse Partial Least Squares regression (sPLS) called sparse Partial Least Squares discriminant analysis (sPLS-DA) was developed to improve on computational efficiency and interpretability of results (143). The sparsity of sPLS-DA approach is achieved with a Lasso penalization (described above in regularization methods) of each of the PLS loading vectors when computing the Singular Value Decomposition. This introduces sparsity in the LDA to obtain a parsimonious model (141,142).

The sPLS-DA in a supervised way identifies a new set of features utilising the response variable. The identified features must approximate all the old features without bias and

be related to the outcome variable, which is a limitation of other dimension reduction techniques such as Principal Components Analysis (PCA) (141,143). This makes it another good approach to utilise when you have a dataset with a high degree of multicollinearity in the predictor variables (140). Importantly, this approach puts a strong focus to visualizations to support the explanation of the results and computational efficiency thus giving it a strong advantage over other types of dimension reduction feature selection approaches (143,144).

2.6.7 Block Forest and Networks

During the multi-omics era, quantifying various types of data such as gene expression, methylation or copy number variation data, protein micro-array data has become fast and less costly. This has led to more patient data with several types of omics data available for the same individual. Similarly, powerful modelling techniques have been developed to assist on feature selection and prediction using multi-omics data. In some health studies with complex patterns combining different types of data would be important to identify important features in different “blocks” or sets of multi-omics data. The interactions between variables across the different blocks of data motivates the importance of modelling the data together (145).

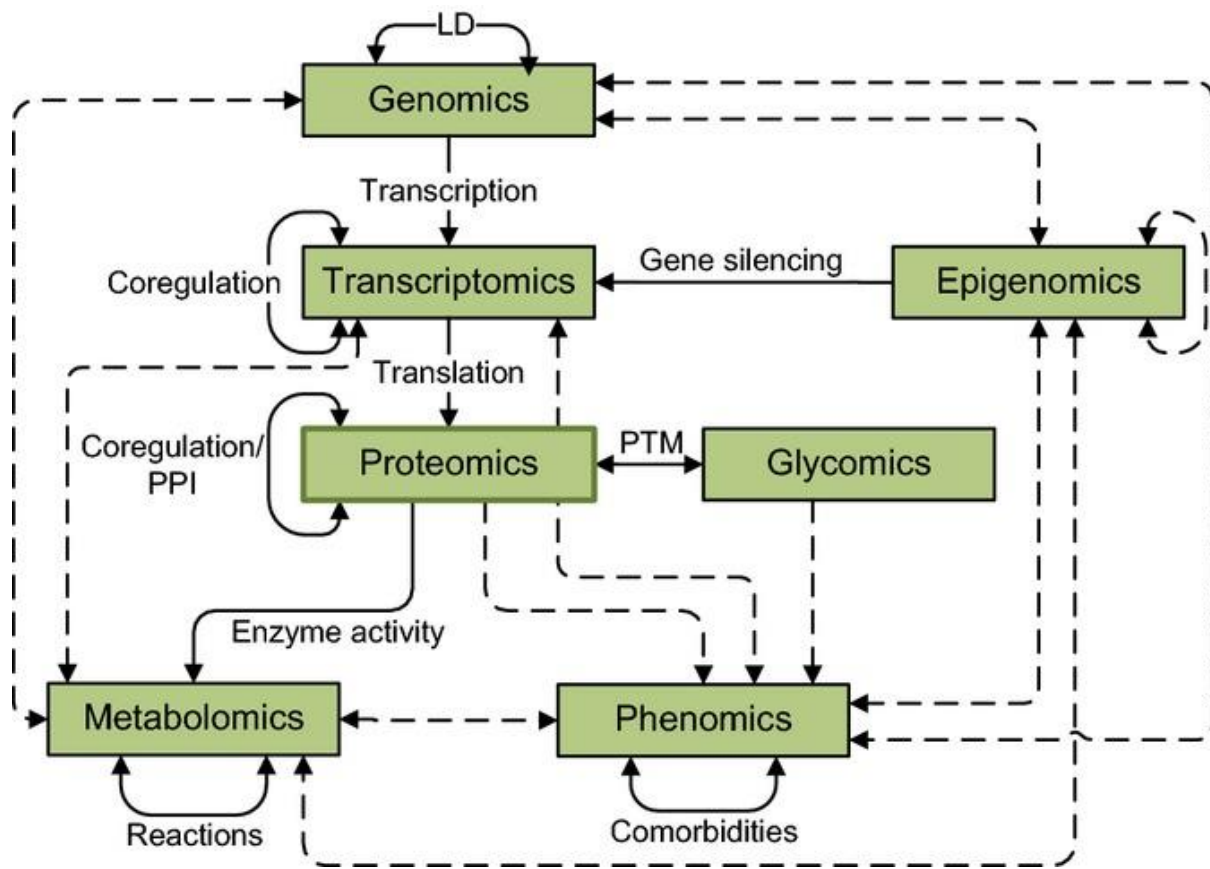


Figure 2-6: Interdependencies of omics data from different blocks. Source: Zierer et al. (146)

A block forest algorithm was developed by Hornung et. al to assist in selecting features from different blocks using block-specific weights of split criteria. The algorithm is built on the random forest algorithm which is a powerful feature selection tool and captures complex patterns between the outcome and the predictor variables (145,147);. The results from these mutli-omics or multi-site data can be presented in networks. The networks have nodes or vertices which are the important features which are connected using edges and links (146).

2.7 The inter-linkages of the different methodologies

The inter-linkage of the different methodologies discussed from Section 2.4 to Section 2.6 is as shown in the Figure 2-7 shown below. The "blue" section depicts the case definition using the classical logistic models and the Bayesian latent class models and is described in detail in Chapter 3. The "orange" color depicts development of tools and functionalities to assist in protein micro array data management and is discussed in detail in Chapter 4. The "brown" depicts Chapter 5 which involves utilising different machine learning approaches to identify correlates of disease protection. The data used in Chapter 5 is generated using the methodologies in Chapter 3 and Chapter 4.

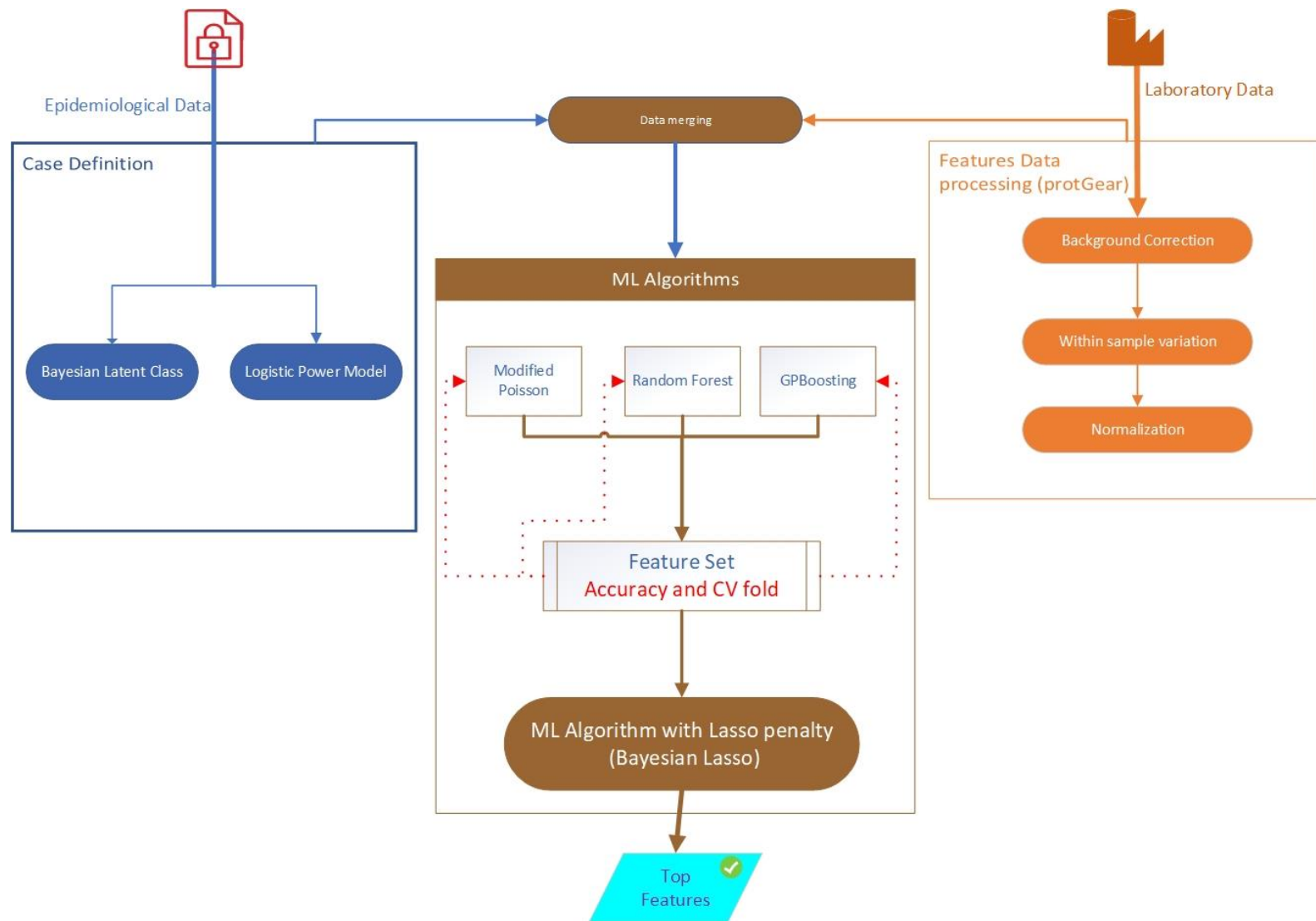


Figure 2-7: The schematic presentation presenting the linking to of the thesis objectives. A) Case definition to reduce misclassification bias. B) Handle non-systematic and systematic protein micro-array bias, C) Identifying important correlates of clinical malaria

2.8 Ethical considerations

The data used in this study had no personal identification; however, we used the individual unique identifier for the analysis. Permission to use the dataset was sought and granted by the KEMRI- Scientific and Ethics Review Unit (SERU) – Approval numbers KEMRI SSC No. 3139 for the SMART study. The data have also been published by Kamuyu et al. 2018 (54). The data have been anonymised and stored in the KEMRI servers in Kilifi, Kenya and managed by Kilifi Data Governance. The Kilifi Data Governance Committee has provided approval to utilise the data to test the PhD statistical models.

The doctoral research was given ethical approval by the University of Witwatersrand's Human Research Ethics Committee (Medical) (Clearance Certificate No. M190121), and a copy of the ethical clearance letter is attached in the thesis appendices.

CHAPTER 3 : CASE DEFINITION

Paper 1. Estimating malaria attributable fractions with changing transmission intensity: Bayesian latent class vs logistic models.

Kennedy Mwai ^{1,2*}, Irene Nkumama ^{2,3}, Amos Thairu ², James Mburu ², Dennis Odera ^{2,3}, Rinter Kimathi ², Lydia Nyamako ², James Tuju ^{2,4}, Samson Kinyanjui ^{2,4,5}, Eustasius Musenge ¹, Faith Osier ^{2,3,4,5}

¹Epidemiology and Biostatistics Division, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa.

²Centre for Geographic Medicine Research (Coast), Kenya Medical Research Institute-Wellcome Trust Research Programme, Kilifi, Kenya

³Centre of Infectious Diseases, Parasitology, Heidelberg University Hospital, Heidelberg, Germany

⁴Department of Biotechnology and Biochemistry, Pwani University, Kilifi, Kenya

⁵Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom

3.1 Abstract:

Background

Asymptomatic carriage of malaria parasites is common in high transmission intensity areas and confounds clinical case definitions for research studies. This is important for investigations that aim to identify immune correlates of protection from clinical malaria. The proportion of fevers attributable to malaria parasites is widely used to define different thresholds of parasite density associated with febrile episodes. We investigated whether the varying intensity of malaria transmission had a significant impact on parasite density thresholds. We used the same dataset to explore an alternative statistical approach of

using the probability of developing fevers as a choice over threshold cut-offs as the former has been reported to increase predictive power.

Methods

Data from children monitored longitudinally between 2005 and 2017 from Junju and Chonyi in Kilifi, Kenya were analysed. We compare the performance of Bayesian-latent class and logistic power models in estimating malaria attributable fractions and probabilities of having fever given a parasite density threshold with changing malaria transmission intensity. Zero-inflated beta regressions were used to assess the impact of using probabilities to evaluate anti-merozoite antibodies as correlates of protection compared with multilevel binary regression.

Results

The malaria transmission intensity declined from over 49% to 5% between 2006 and 2017 respectively. During this period, the attributable fraction varied between 27%-59% using logistic regression compared to 10%-36% using the Bayesian latent class approach. Both models estimated similar patterns of fevers attributable to malaria with changing transmission intensities. The former performed well in estimating the probabilities of having fever, while the latter was efficient in determining the parasite density threshold. However, compared to the logistic power model, the Bayesian algorithm yielded lower estimates for both attributable fractions and probabilities of fever. In modelling the association of merozoite antibodies and clinical malaria, both approaches resulted in comparable estimates, but the utilization of probabilities had a better statistical fit.

Conclusions.

Malaria attributable fractions varied with an overall decline in the malaria transmission intensity in this setting but did not significantly impact the outcomes of analyses aimed at identifying immune correlates of protection. These data confirm the statistical advantage of using probabilities over binary data.

Keywords: case definition, malaria, attributable fractions, probability

3.2 Background

Asymptomatic carriage of malaria parasites is highly prevalent in areas with high malaria transmission as a result of naturally acquired immunity (148). It is therefore likely that, in such areas, an individual with a non-malarial fever has coincidental parasitemia. Since the likelihood of having fever generally increases with parasite density (17,149,150) the assumption is that fever in the presence of parasitemia necessarily constitutes a clinical malaria. However, in high transmission settings (151) parasitemia accompanied by fever may not be adequate to define an episode of clinical malaria and may lead to differential misclassification. Besides causing an over-estimation of malaria burden in an area (18,149), the misclassification complicates immunological and clinical trials where clinical malaria cases are an endpoint or one of the outcome variable. As an outcome variable it is particularly important for identifying correlates of protection from clinical episodes to inform vaccine development.

To overcome this problem of misdiagnosis, different studies have based the case definition of malaria on fever with parasite density above a locally defined threshold.

Computation of malaria attributable fractions (MAF) or the proportion of fevers due to malaria parasites has been used to define different thresholds for parasitemia (17,150).

The classical method for deriving the attributable fraction is a simple numerator denominator approach (152) which is prone to bias when applied in high malaria transmission areas (18). In high transmission settings, individuals may have parasites and not show clinical signs of malaria. Logistic regression models are typically used to handle this bias. They determine the risk of the outcome as a continuous function of parasite density (149,150), and have been widely used to obtain attributable fractions against a range of outcomes with parasitemia as the exposure variable (24,25,149,150,153). Additionally, a Bayesian latent class model of two-component mixture distributions was proposed to improve the estimation of attributable fractions (17). The latent class model was developed to handle the limitation of imprecise or negative attributable fractions occasionally observed in standard logistic regression models (18).

Malaria transmission intensity has been found to strongly influence the attributable fractions. In a study conducted in two areas with different transmission intensities in Kilifi at the coast of Kenya, a MAF of 50.2% derived for Ngerenya, the low transmission site and 47.9% for Chonyi, the high transmission site. In this study, the logistic regression method was applied and derived a parasite density threshold of 2500 parasites/ μ L of blood as the most appropriate to distinguish malaria attributable fevers from fevers due to other causes in both settings. Following this study, this threshold has been widely

applied in the definition of malaria cases in various studies conducted along the Kenyan coast (41,153–156).

Significant reductions in malaria transmission and admissions have been reported over the last decade in endemic countries in Africa (26) and in particular at the Kenyan coast (155,157,158). Based on this observed reduction in transmission and the influence of transmission intensity on the malaria attributable fractions, we conducted the present study to determine the variation of malaria attributable fractions over time.

We also estimated the probability of fever as a function of parasite density and the optimal parasite thresholds using logistic regression (17,159). The estimated probabilities of fever have been used in determining risk of developing clinical episodes in malaria vaccine trials. In these trials, the probabilities estimated from a Bayesian latent class model were proposed as a better approach to compare the placebo and control groups (88,160).

Several articles (92–96) have pointed out problems associated with the categorization of data. These include not only the loss of information on variation and statistical power, but also an increased risk of type I errors and poor predictive performance (92,93,95). In this study, we explore the utilization of probability estimates from Bayesian latent class models as an alternative to dichotomizing individuals using a selected parasite density threshold.

3.3 Methods

3.3.1 Study area and population

We used cohort data from Junju and Chonyi sub-counties in Kilifi County, which is part of the Kilifi Health and Demographic Surveillance System (KHDSS) on the coastal region of Kenya Figure 3-1 (161). The area has two malaria transmission seasons May-July and November-December. For Junju the data were prospectively collected from participants aged 1 to 15 years old between 2005 and 2017 (inclusive) who were initially recruited into a malaria vaccine trial (FP9 ME-TRAP and MVA-ME TRAP, Bejon et al. 2006). The Chonyi dataset had 286 children aged between 0-10years collected in October 2000 and was only used for model comparison (162).

cohort, the cross-sectional malaria survey was done in October 2000. Parasitemia was determined by thin smear microscopy. In both studies, the participants were followed up both actively with weekly home visits by trained field workers and passively at health facilities to identify clinical episodes of malaria. Blood smears were prepared to determine parasite densities for any child who had a fever (axillary temp $\geq 37.5^{\circ}\text{C}$) for the cross-section surveys and follow-up surveillance, respectively. The Government of Kenya-recommended first-line treatment was used for treatment episodes of malaria.

For the parasitemia determination by microscopy, the number of asexual-stage parasites/200 leukocytes was counted, and parasitemia estimated based on actual or assumed (8,000 leukocytes/ μL) leukocyte count measured for each blood smear.

3.3.3 Statistical analysis

A case-control analysis of the relationship between the risk of fever and parasite density was carried out separately for each year using logistic regression and Bayesian latent class models using R (163) and OpenBugs (164) respectively. Additionally, the analysis was repeated in each age group using the overall dataset. Cases corresponded to fever episodes detected by passive or active case detection and controls were those collected during the cross-sectional surveys or the individuals with parasites and no fever during follow-up.

3.3.3.1 Logistic regression

A logistic regression model was fit to the data modelling the risk of fever as a continuous function of the parasite density. The model was of the form $\text{logit}(\pi_i) = \alpha + f(x_i)$ where π_i is the probability that observation i with parasite density x_i is a (fever) case. Along with $f(x_i) = \beta x_i^\tau$, a smooth monotonic function of x_i^τ where τ is the power transformation of the parasite density. This power function τ was tested at different values between 0.10 and 0.90 with a precision of 0.01 and the value that maximized the log-likelihood best was chosen. The malaria attributable fraction (MAF), λ , was estimated using the slope coefficient of the logistic regression; $\lambda = (1/N) \sum_1^i (R_i - 1) / R_i$ where $R_i = \exp[f(x_i)]$ and the standard error estimated using the bootstrap approach with 1000 bootstrap samples. The logistic model was also used to estimate the probability of each individual case of fever being attributable to malaria using $\lambda_{ind} = (R_i - 1) / R_i$ (149).

3.3.3.2 Bayesian latent class

For the Bayesian latent class model, the parasite density was resolved to a mixture of two multinomial distributions. One component $g_1(\cdot)$ corresponds to non-malaria fever episodes and the other component, $g_2(\cdot)$ to children with clinical malaria episodes (fever and parasites). Parasite levels during the cross-sectional bleed were available and were used as the training sample, i.e., a sample that comes from the component of the mixture corresponding to children without fever but who may have parasites. The data was then divided into K ordered categories over the range of the parasite density $\mathbf{X}$. This was

followed by counting the number of test samples $\mathbf{n} = (n_0, n_1, \dots, n_{k-1})$ and control samples (non-fever cases) $\mathbf{m} = (m_0, m_1, \dots, m_{k-1})$. Then the MAF, λ , was then estimated from the two multinomial distributions,

$$\begin{aligned}\theta_i &= P(x \in \text{category } i \mid P_1), \\ \phi_i &= P(x \in \text{category } i \mid P_2), \quad (3.1) \\ \lambda &= P(x \in P_2)\end{aligned}$$

The parameters P_1 and P_2 are the distributions functions of the components $g_1(.)$ and $g_2(.)$ respectively. The category-specific attributable fractions were obtained using,

$$\lambda_i = P(x \in P_2 \mid x \in \text{category } i) = \frac{\lambda \phi_i}{(1-\lambda)\theta_i + \lambda \phi_i}. \quad (3.2)$$

To estimate the probability, λ_{ind} , of each individual case of fever being attributable to malaria local and piece-wise cubic polynomial models were used. The models were fitted using category-specific MAF, λ_c , together with the category-specific midpoint of parasite density. This was followed by predicting the individual λ_{ind} using their parasite density measurements from the results of various model fitting functions.

Sensitivity and specificity of various cut-off values for parasite density were estimated by

$\frac{n_c \lambda_c}{N \lambda}$ and $\frac{1 - n_c (1 - \lambda_c)}{N (1 - \lambda)}$ respectively where $n_c = \sum_{i=c}^K n_i$, $\lambda_c = \frac{(\sum_{i=c}^K \lambda_i n_i)}{n_c}$, n_i the number of fever cases in the category i and c represents the parasite density category

of which it is the selected cut-off in logistic regression or the lower bound for the category in latent class models.

3.3.3.3 Association with protection.

Multi-level logistic and zero-inflated models were used to investigate the association between high versus low merozoite antibodies and clinical malaria. Various antibody concentrations were applied as cutoffs to define the high and low responders (162). The results were used to compare the performance of probability and binary outcomes. The zero inflated approach was used since we had a probability mass at zero due to the non-febrile participants. Specifically, for the probability outcome, we compared results from the Maximum Likelihood (MLE) and Bayesian inference estimations (162,165,166).

3.4 Results

3.4.1 Study population

A total of 4,722 participants were recruited in Junju, Kilifi County from 2005 to 2017. Approximately 300 or more participants were followed up each year with average recruitment age of 6.5 years (ranging between 1 month old to 16 years) as shown in Table 3-1. Each child had on average 2.94 test occurrences during follow-up giving rise to a total of 14,404 events during the entire study period.

3.4.2 Temporal distribution of fever and parasite density

Table 3-1 shows the distribution of fevers (axillary temperature of $\geq 37.5^{\circ}C$) at all the cross-sectional surveys. Approximately 1,034 (2.19%) occasions of fever were reported during the cross-sectional surveys. A decreasing trend of fevers was observed over the study period except for 2005 and 2006 where the samples were collected specifically for a vaccine trial (1). The

prevalence of *P. falciparum* was also artificially high during this period since the participants were recruited during the malaria season. A decline in the prevalence of *P. falciparum* parasite was observed between 2006 to 2013 from 30.21% to 8.78%. This was followed by a slight increase in 2014 and 2015 then another decline in 2016 to 4.32% in 2017.

Table 3-1: Prevalence's of Plasmodium falciparum positivity, fever, and presumptive malaria (fever + parasitemia) in the pre-transmission season cross-sectional survey and the number of active follow-up events in the Junju Cohort. ¶ Participants were recruited for the original vaccine study in 2005 and 2006 (ref) and sampling of participants extended into the high transmission season each year. *Excluded in the main analysis due to health workers' strike.

Year	Samples (n)	Mean age in years (min-max)	<i>P. falciparum</i> (%)	Fever (%)	Presumptive malaria (%) †	Active follow-up events (n)
2005¶	372	3.9 (1-9.0)	49.05%	0.81%	0.54%	1029
2006¶	300	4.5 (1.5-9.5)	31.21%	0.67%	0.33%	615
2007	339	4.8 (<1-11.0)	15.93%	3.24%	1.18%	1816
2008	341	5.4 (<1-12.0)	29.62%	3.52%	3.52%	884
2009	352	5.8 (<1-13.0)	20.17%	2.56%	1.14%	958
2010	377	6.5 (<1-14.0)	27.59%	3.98%	2.65%	957
2011	377	7.1 (<1-12.7)	23.08%	1.86%	1.33%	891
2012*	399	7.1 (<1-13.7)	16.79%	1.75%	0.50%	696
2013	410	7.5 (<1-14.7)	8.78%	1.46%	0.24%	1483
2014	404	8.2 (<1-15.7)	14.25%	1.49%	0.50%	1619
2015	400	8.3 (<1-16.5)	17.36%	1.25%	0.25%	1319
2016	316	7.8 (<1-15.0)	11.61%	1.27%	-	1087
2017	335	7.1 (<1-15.0)	4.32%	1.19%	-	1050
Total	4722	6.5 (<1-16.5)	20.69%	2.19%	1.06%	14404

3.4.3 Relationship of fever to parasitaemia over time

The probability that a fever case was malaria attributable at a given parasite density λ changed gradually over the study period as shown in Figure 3-2. The MAF was estimated using the Bayesian Latent class model and logistic regression. The Bayesian latent class gave a lower MAF estimate, Bland-Altman bias =0.20 (0.16-0.24), compared to the logistic model. After estimating the

sensitivities and specificities of different parasite densities, the optimal parasite cut-off was selected using the logistic regression for the different years [Supplementary Table 3-1]. However, the number of malaria positive individuals did not vary significantly with the new thresholds compared

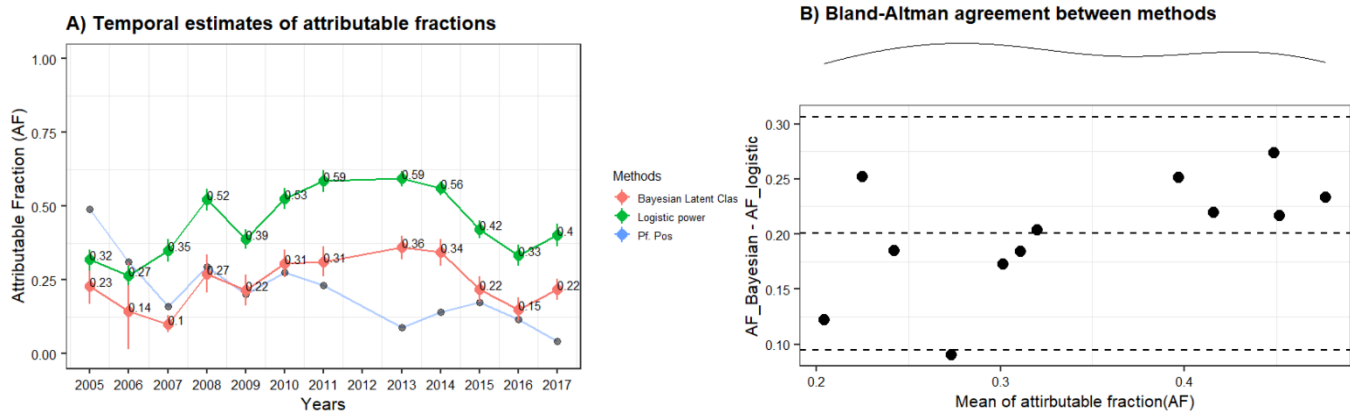


Figure 3-2: A) Temporal estimates of attributable fraction (AF) from 2005 to 2017 using Bayesian Latent class models and logistic power models. Pf. Pos is the prevalence of parasite positivity during the cross-sectional bleed. B) Bland-Altman plot of agreement

to the previously defined 2500p/μl threshold [Supplementary Figure 3-1], despite the changing patterns. Notably, the Bayesian latent class approach and the logistic power models approximated a similar pattern of MAF but the estimates were lower in the former model. Comparable patterns were also observed in the probabilities, λ_i , predicted from the individual parasite densities [Supplementary Figure 3-1].

3.4.3.1 Non-febrile individuals.

An interval estimate for the prevalence of malaria fever was estimated using the Bayesian latent class model. The Bayesian individual probabilities for non-febrile participants with parasitemia were adjusted using the interval estimate for the prevalence of malaria fever. This is shown in Supplementary Figure 3-2 and Supplementary Figure 3-3 where non-febrile cases had a lower, λ_i

,compared to the febrile cases for the parasite positive individuals. Detailed implementation of the methodology is included in the repository as OpenBUGS and R codes.

3.4.4 Impact of age on MAF.

The MAF estimates were higher for older age groups than the children <1 year as shown in Figure 3-3A below. Additionally, the predicted individual probabilities declined with age as shown in Figure 3-3B and likewise the logistic power model had higher estimates and smaller range than the Bayesian latent class predictions [Supplementary Figure 3-4]. This shows as expected that the age groups of 1-5 years and 5-10 years had a higher probability of having malaria compared to the other age groups. The younger age groups had a higher specificity and sensitivity intersection

(Figure 3-3) indicating a lower parasite density threshold for clinical episodes compared to the older age groups.

To assess whether the variation in case definitions had a significant impact on the estimates of correlates of protection. The performance of the predicted probabilities from the Bayesian latent

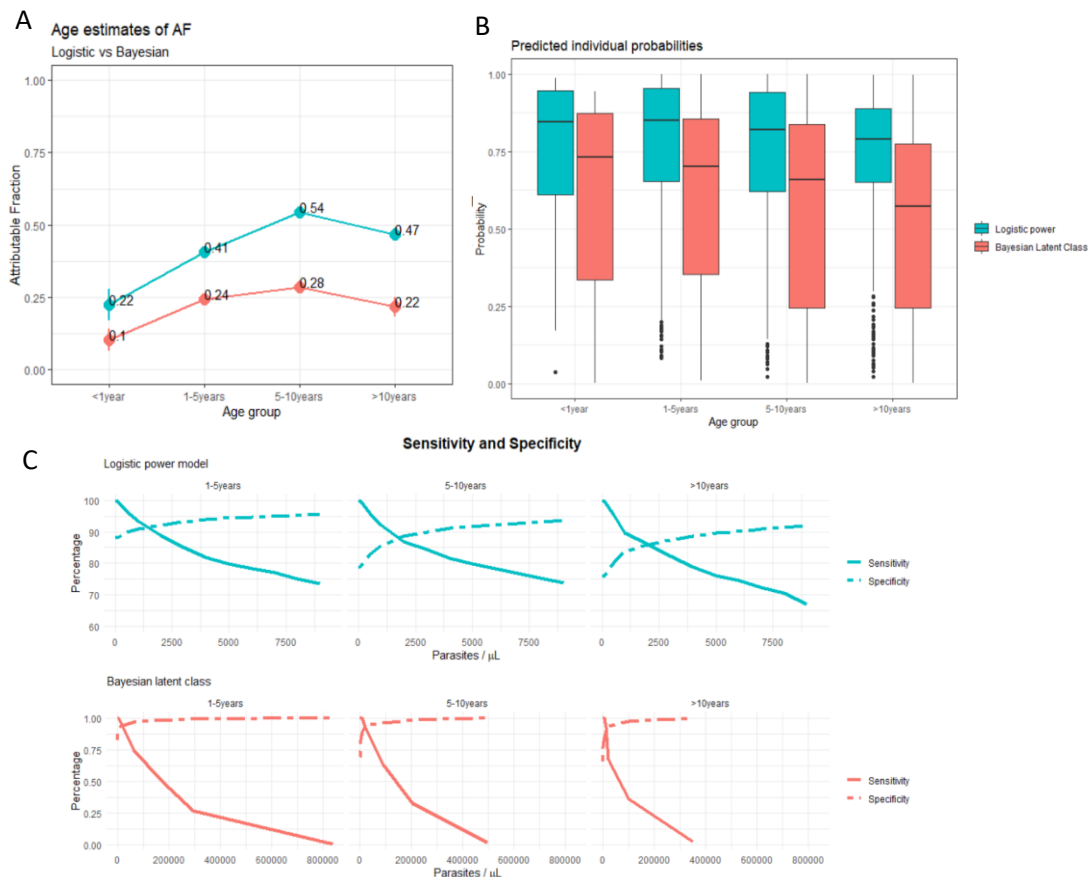


Figure 3-3: Malaria attributable fractions and probabilities over age group for all the study participants

class approach and the binary outcome based on the newly defined parasite density thresholds plus fever were compared. Data on antibody responses to selected *Plasmodium falciparum* merozoite antigens for a study done in Kilifi was used for the comparison. Specifically, the data had antibody measurements for the survey conducted in Junju in the year 2008 and a subset of the Chonyi cohort in the year 2000 (162).

A cut-off of 2500 parasites/ μ L [Supplementary Table 3-1] plus fever was used to define the binary outcome (malaria positive, parasites $\geq$ 2500 parasites/ μ L or negative otherwise). The loess

predicted probabilities from Bayesian latent class models were used as the response variable to fit the zero-one inflated beta regressions. Table 3-2 shows that using the probability as the outcome gave comparable point estimates with the binary outcome. The binomial multilevel models however, had high standard errors and Bayesian Information Criterion (BIC) values.

Table 3-2: A comparison of a binary and probability outcome using high vs low antibody levels in Junju 2008 and Chonyi 2000 cohort. Coef. – Regression Coefficients, SE- Standard Error and BIC – Bayesian Information Criterion,

		Binary Outcome		Probability Outcome		Probability Bayesian
		Coef. (SE)	BIC	Coef. (SE)	BIC	Coef. (SE)
Junju Cohort	AMA1	-0.30 (0.32)	427.73	-0.22 (0.20)	416.026	-0.23 (0.19)
	MSP2	0.09 (0.34)	428.6	-0.04 (0.20)	417.3474	-0.04 (0.21)
	MSP3	-0.24 (0.35)	428.18	-0.16 (0.19)	415.947	-0.19 (0.21)
Chonyi Cohort	AMA1	0.18 (0.34)	426.8972	-0.24 (0.11)	390.0447	-0.26 (0.13)
	MSP2	-0.67 (0.54)	425.4389	-0.33 (0.14)	401.3226	-0.34 (0.15)
	MSP3	0.68 (0.35)	423.5147	-0.11 (0.12)	410.5933	-0.12 (0.14)

3.5 Discussion:

In areas with high malaria transmission, differences in the prevalence of malaria fever can occur among various subsets of the population like age groups due to change in transmission intensities or differences in levels of immunity (150). The present study shows a variation of transmission intensity over time, and how this contributes to variation in the MAF. We observe a decreasing trend in transmission intensity and a lower probability of having malaria in older age groups. Similarly, a previous study done in Kilifi showed that immunity to malaria is affected by age and transmission (150). The study compared Chonyi, a high transmission area and Ngerenya, a low transmission area. The sites had a variable age-specific clinical disease pattern with Ngerenya having a higher MAF compared to Chonyi overall and specifically for the older age group of 5-

19years. Shifting MAF was also observed with changing transmission patterns in the current study. A shift in malaria transmission intensities and malaria epidemiology has been reported in different endemic areas (26,167). Therefore, it is important to review MAF and case definitions with changing transmission settings. Furthermore, transmission intensity correlates with the rate of acquisition of natural immunity (168). A decrease in malaria transmission intensity led to reduced immunity which would result in a higher tendency to acquire malaria attributable fevers at lower parasite densities as was observed in this study.

A strong rationale for developing malaria vaccines comes from cohort studies, which show that individuals continuously exposed to malaria develop immunity that initially prevents death from severe disease, and subsequently recurrent illness (41,148). The main assumption in defining correlates of protection to inform vaccine development is that malaria case definition is non-biased. Many of the studies classify the participants into two groups (clinical malaria case and non-case) using a defined parasitemia threshold plus fever (24,25,149,150,153). The optimal parasite density threshold is selected from maximum combined sensitivity and specificity after fitting the case definition models (149). Additionally, the models estimate the probabilities individual episodes of fever are malaria attributable at a given density of parasitemia (17,149,159).

The logistic power model is the widely used technique for case definition (19,25,53,78) and rarely the Bayesian latent class model (17,159). However, in this study the logistic approach was observed to give higher but comparable pattern estimates with the Bayesian latent class model. Similarly, this was observed in a study done by Vounatsou et. al comparing the logistic power and Bayesian latent class model (17). The logistic model approach, however, has been reported to have a limitation of estimating imprecise standard errors and negative probabilities sometimes (18). Comparatively, in this present study, the logistic approach also gave high probability

estimates with narrow variation in the low parasite densities compared to the latent class model. In vaccine studies, the latent class was reported to help in identifying possible biases in efficacy estimates since it utilizes the whole range of possible parasite density cut-offs (88). An inverse relationship of clinical malaria and age has been shown (155,169), similarly, this was observed with the estimated probabilities which decreased with age.

Continuous variables, like the probabilities we use here have been shown to have more variation information and statistical power and are sometimes preferred over the categorization of data (93). Several articles (92–94,96) have also pointed out problems associated with the categorization of data. In this present study, we compare the performance of using probability and binary outcome model the association with clinical malaria (162). Assuredly, the probability model had a good statistical fit; it had lower BIC estimates and standard errors and gave comparable coefficients with the binary model. Also, the point estimates were similar to what was reported by Murungi et. al in the 2008 study from the same cohort (162). In this study however, they reported risk ratios that were estimated using a modified Poisson regression (170) and here we report coefficients both of them showing a lower risk of disease for individuals with high antibody measurements.

3.6 Conclusion:

The present study shows the importance of reviewing malaria case definitions and attributable fractions with changing malaria transmissions. Additionally, we show the utilization of probabilities estimated from the case definition models in modeling the association of correlates of disease. Results from Junju and Chonyi Kilifi data verify the validity of using the probability outcome to identify correlates of disease protection while still having a better statistical fit. The computational time to fit the zero-inflated models was higher compared to the binary-based regression models

and a training class is required for the latent class models which can be a limitation for some cohort designs.

Approaches to estimate individual's marginal probabilities of clinical malaria over a given follow-up time would be of importance for creating parsimonious models. Further studies to compare the probabilities estimated from models utilizing the quantitative nature of the parasite densities without grouping the data in conjunction with changing transmission would be valuable.

3.7 Ethics approval and consent to participate

The research was given ethical approval by the University of Witwatersrand's Human Research Ethics Committee (Medical) (Clearance Certificate No. M190121) and KEMRI- Scientific and Ethics Review Unit (SERU) (Approval numbers KEMRI SSC No. 3139). Informed consent was collected from parents/guardians before the beginning of recruitment.

3.8 Consent for publication

Not applicable

3.9 Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. The scripts used for the current study are available in the GitHub repository, https://github.com/Keniajin/case_definition.

3.10 Competing interests

The authors declare that they have no competing interests.

3.11 Funding

This research was commissioned by the National Institute for Health Research (NIHR) Global Health Research programme (16/136/33) using UK aid from the UK Government. The views

expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

K.M, F.O, S.K are funded by the National Institute for Health Research (NIHR) Global Health Research programme (16/136/33). F.O. is supported by a Sofja Kovalevskaja Award from the Alexander von Humboldt Foundation (3.2 - 1184811 - KEN - SKP) and an EDCTP Senior Fellowship (TMA 2015 SF1001) which is part of the EDCTP2 Programme supported by the European Union. S.K and E.M are supported by the DELTAS Africa Initiative under Initiative to Develop African Research Leaders (IDeAL) Grant No. DEL-15-003 and DELTAS Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) Grant No. 107754/Z/15/Z-DELTAS Africa SSACAB respectively.

3.12 Authors' contributions

Conception and Design of the work: KM, IN, AT, JM, EM and FO. Analysis: KM, JM and AT. Funding Acquisition: EM, SK and FO. Interpretation of data: KM, IN, SK, EM and FO. Work drafting: KM and FO. Review and Editing: KM, IN, AT, JM, RK, DO, LN, JT, SK, EM and FO. Project administration: LN. Software: KM, JM and AT. Supervision: EM, SK and FO. Validation: SK, EM and FO. All authors read and approved the final manuscript.

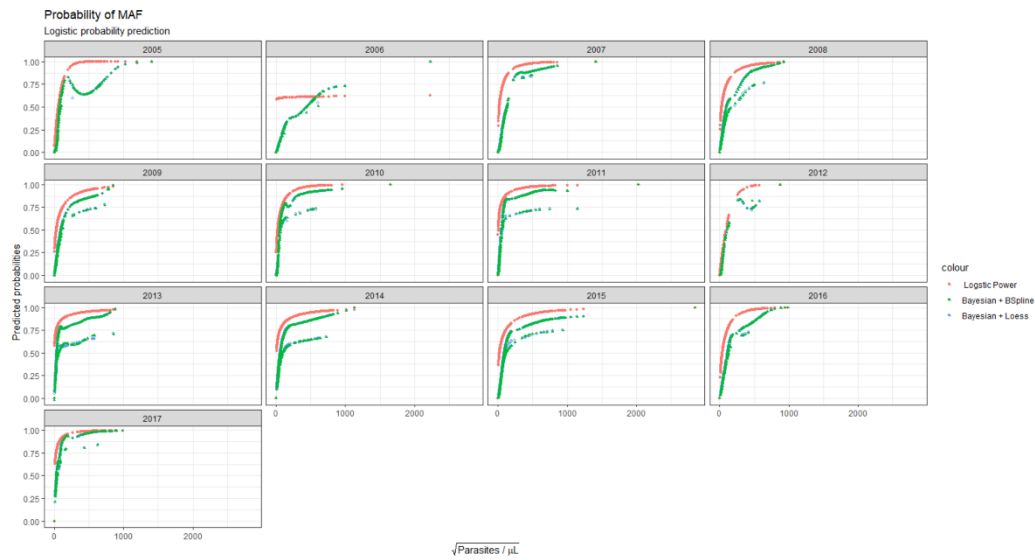
3.13 Acknowledgements

We thank all participants who took part in the cohort studies and the field workers involved in the data generation process.

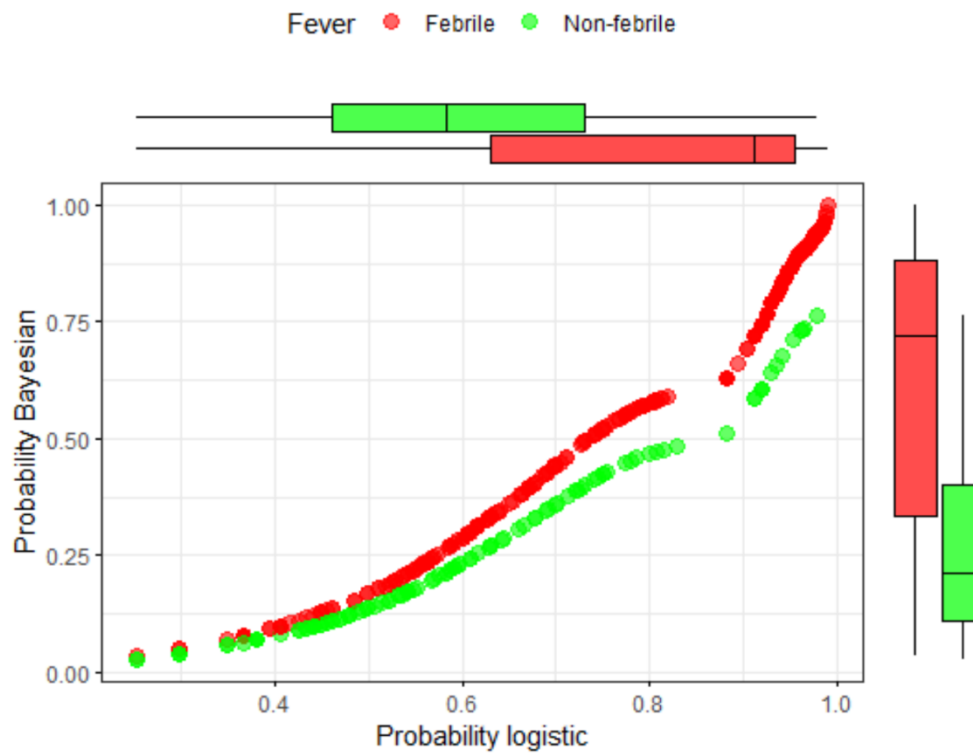
3.14 Supplementary

Supplementary Table 3-1: Parasites/ μL cut off using Logistic regression. **Excluded in the main analysis due to health workers' strike.*

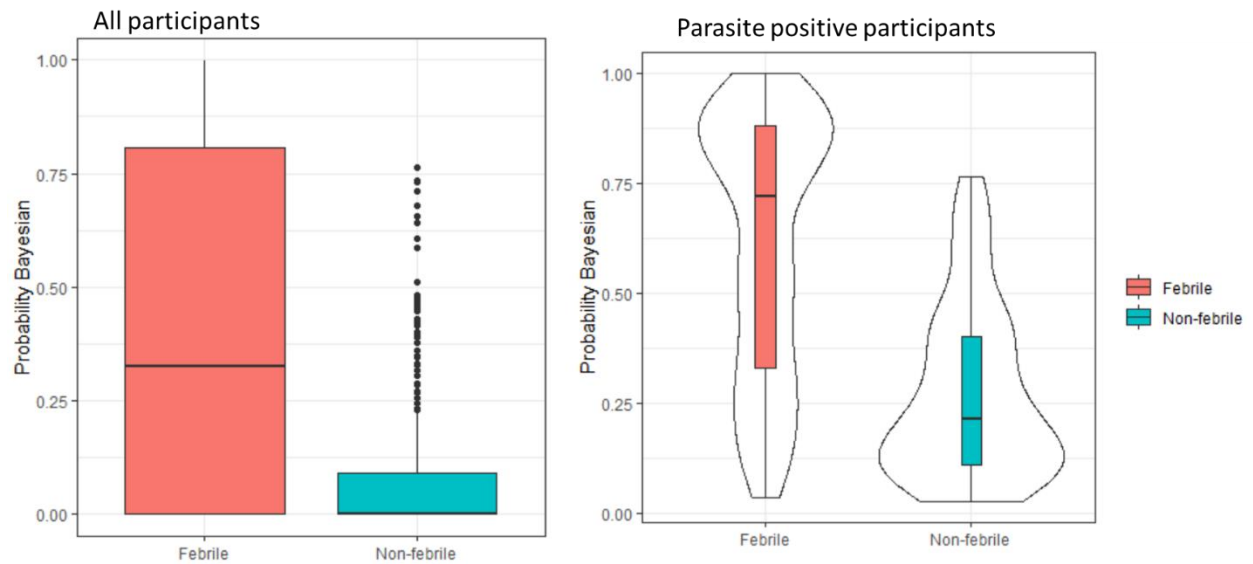
Year	Parasites/ μL cut off	Malaria Positive	Malaria Positive by 2500 p/ μL	Sensitivity	Specificity	Lambda (UL - LL)
2005	2000 p/ μL	74	73	89.66	90.24	81.11 (77.96,84.25)
2006	1300 p/ μL	47	43	81.05	80.95	60.57 (60.44,60.7)
2007	900 p/ μL	56	46	91.86	91.23	84.99 (82.91,87.04)
2008	2500 p/ μL	129	129	87.23	86.08	87.29 (85.67,88.9)
2009	1800 p/ μL	105	102	87.84	88.1	82.38 (80.83,83.93)
2010	3000 p/ μL	133	137	89.06	88.96	89.96 (88.8,91.11)
2011	1200 p/ μL	123	119	87.93	87.81	91.07 (90.27,91.89)
2012*	1000 p/ μL	18	16	95.39	94.44	65.01 (55.3,74.56)
2013	2500 p/ μL	173	173	87.09	86.74	90.56 (90.15,90.99)
2014	3500 p/ μL	207	211	85.25	85.41	88.17 (87.65,88.67)
2015	3500 p/ μL	156	158	87.27	87.47	83.56 (82.38,84.69)
2016	1500 p/ μL	104	101	92.95	91.99	85.37 (83.32,87.44)
2017	500 p/ μL	62	59	94.67	95.04	92.80 (91.73,93.84)



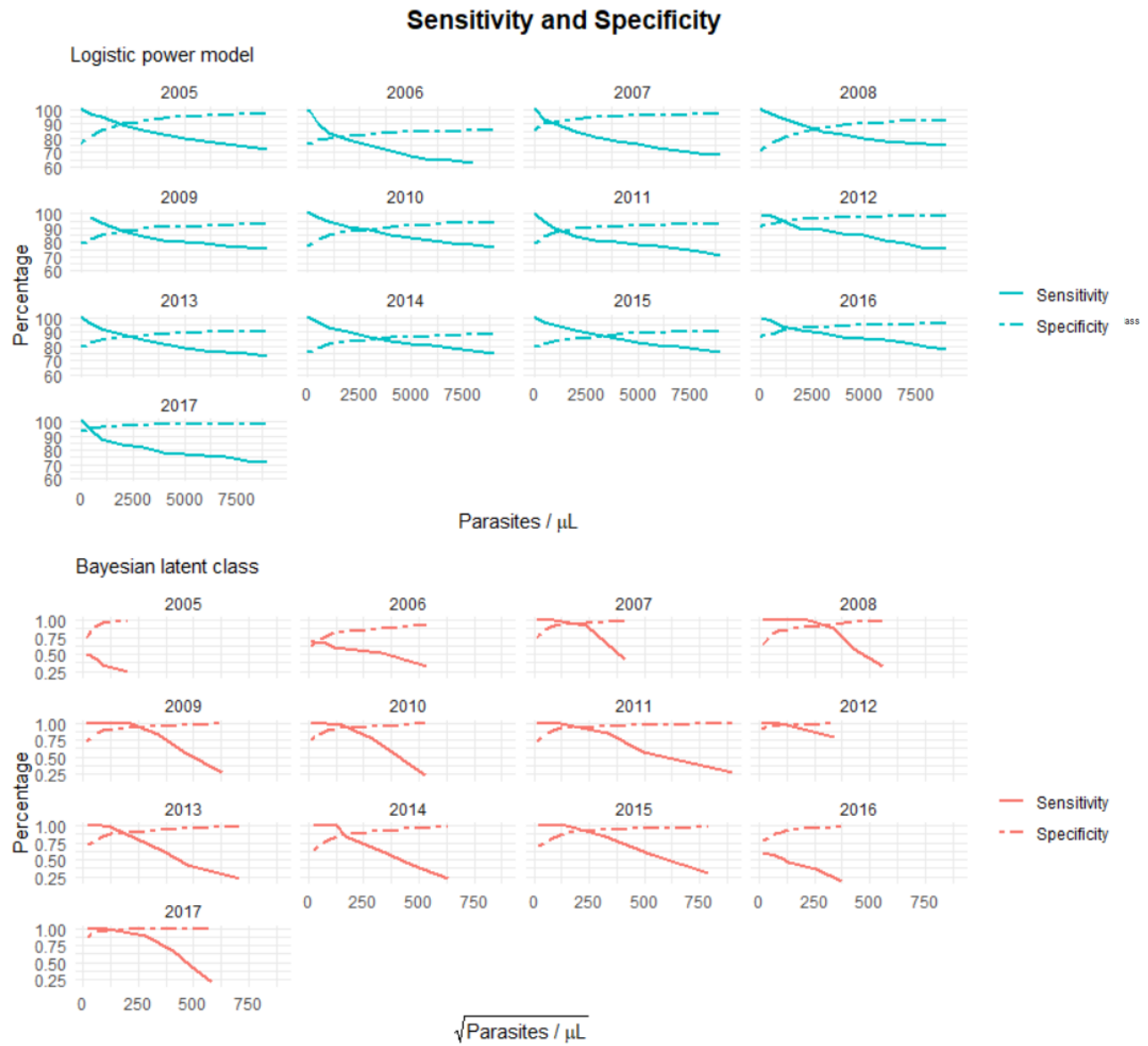
Supplementary Figure 3-1: Distribution of predicted probabilities



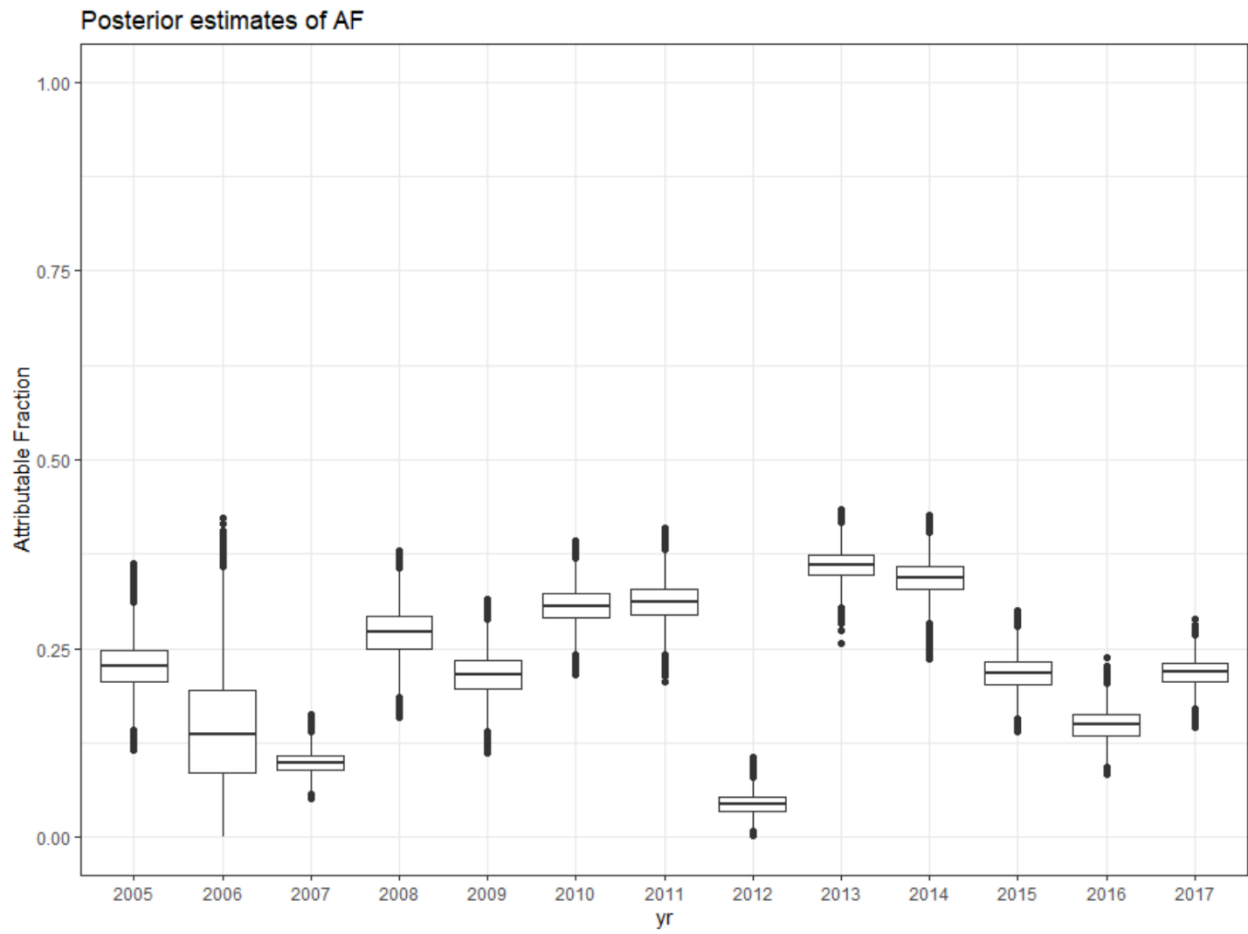
Supplementary Figure 3-2: Comparison of probabilities for Bayesian and Logistic, Junju 2008



Supplementary Figure 3-3 Comparison of probability of febrile and non-febrile



Supplementary Figure 3-4: Sensitivity and specificity of the different years



Supplementary Figure 3-6: Posterior estimates of AF

CHAPTER 4 : PROTEIN MICRO-ARRAY DATA PROCESSING

This study has so far developed and compared methodologies to handle differential misclassification associated with the outcome that is clinical malaria. This chapter introduces methods to handle the systematic and non-systematic biases associated with high throughput laboratory data. Additionally, we create an R package suite to assist other researchers in protein micro-array research in handling the systematic and non-systematic biases. Moreover, we create an interactive dashboard that eliminates the need for prowess in R programming while using the tool. The open-source R interface includes illustrative examples, generates an auditable record, enables reproducibility, and can incorporate additional custom scripts through its online repository (<https://github.com/Keniajin/protGear>).

The chapter is formatted as the paper published in Computational and Structural Biotechnology Journal Volume 19,2021 (<https://doi.org/10.1016/j.csbj.2021.04.044>)

Paper 2: protGear: A protein microarray data pre-processing suite.

Kennedy Mwai^{*1,2}, Nelson Kibinge², James Tuju^{2,4}, Gathoni Kamuyu², Rinter Kimathi², James Mburu², Emily Chepsat², Lydia Nyamako², Timothy Chege², Irene Nkumama^{2,3}, Samson Kinyanjui^{2,4,5}, Eustasius Musenge¹, Faith Osier^{2,3,4,5}

¹Epidemiology and Biostatistics Division, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa.

²Centre for Geographic Medicine Research (Coast), Kenya Medical Research Institute-Wellcome Trust Research Programme, Kilifi, Kenya

³Centre of Infectious Diseases, Heidelberg University Hospital, Heidelberg, Germany

⁴Department of Biotechnology and Biochemistry, Pwani University, Kilifi, Kenya,

⁵Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom

* Corresponding author: email: kmwai@kemri-wellcome.org, phone: +254724211786

4.1 Abstract.

Protein microarrays are versatile tools for high throughput study of the human proteome, but systematic and non-systematic sources of bias constrain optimal interpretation and the ultimate utility of the data. Published guidelines to limit technical variability whilst maintaining important biological variation favour DNA-based microarrays that often differ fundamentally in their experimental design. Rigorous tools to guide background correction, the quantification of within-sample variation, normalisation, and batch correction specifically for protein microarrays are limited, require extensive investigation and are not centrally accessible.

Here, we develop a generic one-stop-shop pre-processing suite for protein microarrays that is compatible with data from the major protein microarray scanners. Our graphical and tabular interfaces facilitate a detailed inspection of data and are coupled with supporting guidelines that enable users to select the most appropriate algorithms to systematically address bias arising in customized experiments. The localization and distribution of background signal intensities determine the optimal correction strategy. A novel function overcomes the limitations in the interpretation of the coefficient of variation when signal intensities are at the lower end of the detection threshold. We demonstrate essential considerations in the experimental design and their impact on a range of algorithms for normalization and minimization of batch effects.

Our user-friendly interactive web-based platform eliminates the need for prowess in programming. The open-source R interface includes illustrative examples, generates an auditable record, enables reproducibility, and can incorporate additional custom scripts

through its online repository. This versatility will enhance its broad uptake in the infectious disease and vaccine development community.

Keywords: protein microarray, normalisation, background correction, batch correction, reproducibility

4.2 Introduction.

Protein microarray technology is increasingly utilised for vaccine candidate discovery among other range of applications in the 'omics era' with hundreds to thousands of antigen-specific antibodies analysed simultaneously (54–56,171,172). Antibody data are correlated with infection or disease outcomes in experimental models and observational studies(173). The platform is also useful for the dissection of variant-specific antibodies induced by polymorphic proteins (174).

Although multiple pipelines of the analysis of DNA microarrays are published (57,111), they are not always suitable for proteins because of fundamental differences in the underlying experimental design. In the former, gene expression levels are typically compared by mixing test and control samples that are labelled with a pair of distinct fluorescent dyes. The emission signal at a defined locus in a test sample is expressed as a ratio, relative to its counterpart in the control. Normalization in this context factors in intrinsic differences between dyes and the relative efficiency of their incorporation into the samples under investigation. In contrast, standard protein microarrays (the reverse and forward phase microarrays) quantify the absolute fluorescent emission detected

following antibody binding to a single protein, therefore other considerations for normalization become more important.

Similarly, although the concordance between replicates rises with increasing signal intensity (mean-variance dependence) for both DNA and protein microarrays, the respective normalization algorithms differ. An expectation of minimal variation in the majority of genes with the exception of the one(s) under investigation is the norm in standard DNA microarray experiments (38). The exact opposite is true for experimental designs where important differences in antibody binding between individuals and proteins underpin the hypothesis (54). Consequently, while scaling down variation for DNA microarrays serves the correct purpose, some algorithms may mask important biological variation in responses to proteins (38).

Tools to guide the rigorous processing of protein microarray data are limited (58,61,63,64), some are time-consuming to optimize and not centralized (62). Some of the available tools are Protein array web explorer (PAWER) (64), Protein Microarray Database (PMD) (63), Protein Microarray Analyser (PMA) (62), Prospector, Protein Array Analyser (PAA) (61). Here, we provide a one-stop data-processing suite that empowers users to determine the most appropriate method for each data-handling step by comparing the different data handling techniques. A detailed comparison of the tools is documented in 123Supplementary D:. We systematically address background correction (105) within-sample variation (38), normalisation (175) and batch correction (98). Our easy-to-use interactive web-based R interface (71) and illustrative examples enable wide utility.

4.3 Methods.

The data processing suite incorporates a range of sequential statistical functions organized into an R package with accompanying guidance notes (Figure 4-1).

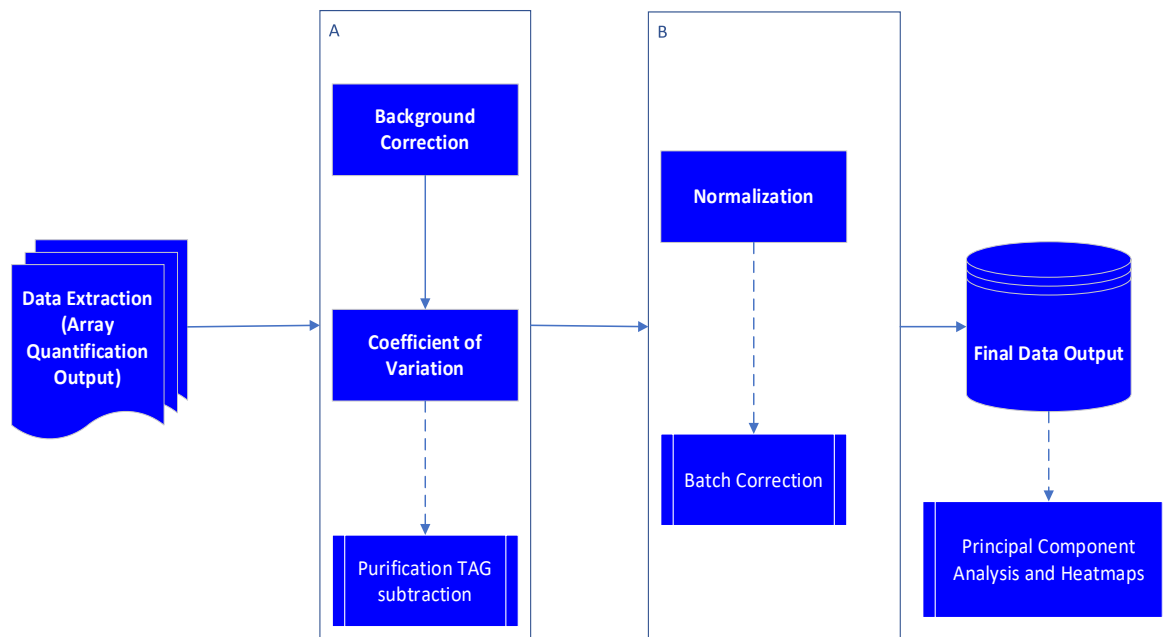


Figure 4-1: protGear data processing scheme. Dotted lines indicate optional steps. Tag subtraction is applied for antigens containing purification tags. Batch correction is relevant when multiple samples from the same sample set are processed in more than one experiment.

We examine the performance of protGear using data from KILchip v1.0, a protein microarray chip designed to enable the simultaneous detection of antibodies against > 100 proteins in large cohort studies. *Plasmodium falciparum* proteins were printed in triplicate on a slide divided into mini-arrays, defined as the region allocated to a discrete sample and can be further divided into blocks (54).

4.3.1 Data extraction.

Microarray image analysis software captures multiple parameters in relation to antibody intensity in a pre-prepared template mostly referred to as a “.gal” file. [Supplementary A: Definition of terms and quantitation results file contents] Quantification softwares for estimating pixel intensities for example Proscanarray Express software (PerkinElmer) (171), QuantArray software (GSI Lumonics) and GenePix® Pro software (Molecular Devices) (176) generate data with similar parameters. Pixel intensities for each spot are typically reported as means or medians with respective standard deviations. We recommend the median as it is relatively insensitive to outliers. We created a versatile tool that can be adapted to load data output format from different quantification softwares mentioned above and this is highlighted in Supplementary A: Definition of terms and quantitation results file contents.

4.3.2 Background correction.

Background intensity is the signal emitted by sources other than the sample under investigation (104,105). In microarrays, it arises either from the glass slide and/or from the non-specific binding of analytes and can vary within and between slides(38,104,105). The foreground is the total spot intensity and includes the background.

The background is typically calculated using a circular region around the spot (Figure 4-2). In GenePix® for example, this is estimated using a diameter three times that of the corresponding spot indicator (176). We adapted subtraction and model-based functions for background correction using a combination of the GenePix® Pro (176) and Linear Models for and Microarray Data (Limma) (57), with minor modifications respectively.

When the background exceeds the foreground, we implement functions to enable mathematical computation as explained under moving minimum background and half moving minimum background correction (105,106). Prior to background correction, it is important to visually inspect the protein microarray data for any spatial biases. protGear provides two functionalities to visually inspect the slides, however, we recommend inspection of the scanned images [Figure 4-2] since the data file might not record all the spatial artefacts. The dashboard includes a function to visualize the printing buffer spots which are used to monitor any background reactivity and detection of any potential protein carry over during printing (54).

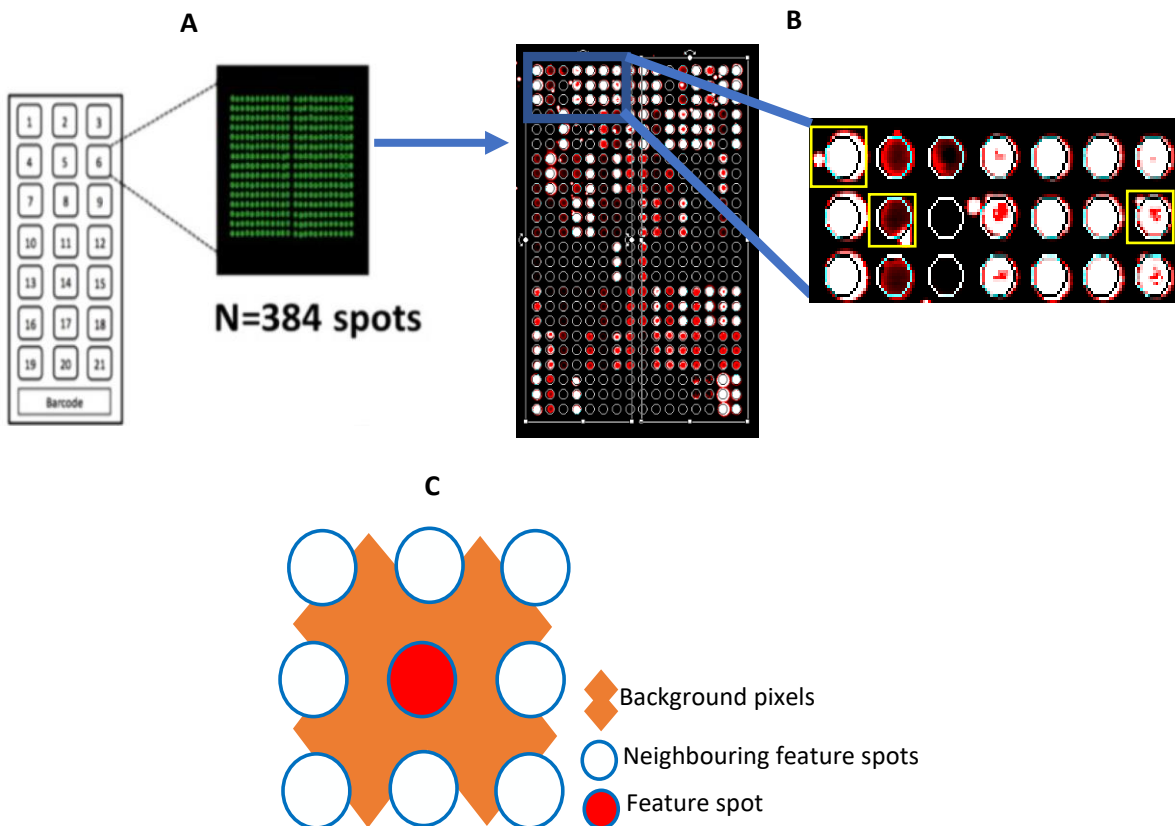


Figure 4-2 Background correction: artefacts add noise to the signal intensity A) A microarray slide with 21 mini arrays and a barcode. Each mini array has a specific number of features represented by a spot. B) Artefacts [spots surrounded with yellow boxes] Kamuyu 2018. C) The total foreground intensity associated with feature spot typically includes the local background.

protGear provides a range of background subtraction methods based on understanding the distribution of the artefacts in specific experiments and supported with graphical outputs. The *local background* is the signal detected in the immediate vicinity of a spot. Subtracting the median local background pixel intensities from the foreground is thought to give an unbiased estimator of the true signal intensity for that specific spot (105). However, uneven variation in the local background across an entire slide may skew the data (101) and can be minimized by implementing a *global background* subtraction. This subtracts the median of all the local background intensities for a given slide from each

spot on that slide. When artefacts are localized to a specific region of a slide often referred to as a block or mini-array, the *moving minimum background* or *half moving minimum background correction* options can be adopted (57). The former restricts the subtraction of local backgrounds to the block within the slide. Since many spots may be affected within a block, it utilizes the minimum rather than the median local background. Zero or negative values are subsequently set to half the minimum of the positive corrected background intensities different from limma implementation that sets any intensity which is less than 0.5 after background subtraction to be equal to 0.5.

The *normal and exponential model (normexp)* method is recommended when the distribution of background intensities is normal, but that of the background-corrected data is exponential (107). The background intensities are fitted as covariates in a convolution model and the expected signal given the observed foreground becomes the corrected intensity. This yields a smooth monotonic function of positive background-corrected intensities and replaces negative values on the entire slide with a single positive co-efficient (108).

The *log-linear background correction* method utilizes the range of positive background-corrected values on the slide to compute a log-linear smooth monotonic function from which negative values are interpolated (106). It can be considered for data that do not fit the above distributions.

4.3.3 Coefficient of variation.

Technical replicates assess within-sample variability and help to quantify the reliability of the experimental procedures. They can minimize data loss when for example a single

spot performs poorly, and others succeed. They are also utilized to detect sample-specific experimental variability linked to outliers or caused by specific reagents. Although many replicates are advantageous, this comes at a cost and may reduce experimental throughput.

A function to evaluate the extent of within-sample variability is implemented using the Coefficient of Variation (CV). This is a measure of accuracy expressed defined as;

$$CV\% = \left(\frac{\sigma}{\bar{x}} \right) \times 100, \text{ where } \sigma \text{ is the standard deviation divided by the mean } \bar{x}, \text{ and}$$

expressed as a percentage (38). Users can define a threshold (cut-off) for the CV using n -replicates and visualize those that do not meet this criterion. Replicate spots that are acceptable are averaged (110). Additional functions enable the visualization and subtraction of signal intensities from purification tags, where applicable. Protein purification tags refer to specific amino acids or polypeptides fused to target proteins to facilitate their subsequent affinity purification. In the case of the datasets used here, the tags include the CD4 hexa-histidine, Maltose binding protein (MBP) and Glutathione S-transferase (GST) tags (54). Purification-tag is specific to different experiments and designs, purification-tag subtraction step is optional as shown in Figure 4-1.

4.3.4 Normalisation.

Data normalization minimizes the mean-variance dependence (MVD) that is common in microarray experiments and may mask true biological variability, Figure 5A (111,177). Normalisation transforms the intensities to a scale where the variance, $Var(MFI)$ is independent of the mean, $E(MFI)$ where MFI is the (Mean Fluorescent Intensity).

Although many normalisation strategies have been proposed (178) we focus on five techniques applicable to protein microarrays.

4.3.4.1 *Log₂ Normalisation*

Log transformations reduce the bias in variance between high and low values by converting the data from a multiplicative to an additive distribution. Wider variance is typically observed for higher signal intensities. Log₂ transformation is readily compatible with the doubling sample dilutions typically employed on the bench. The background-corrected values $\hat{y}_{ijk_r} = y_{ijk_r} - \hat{\alpha}$; where $\hat{\alpha}$ is the estimate of the median background level and y_{ijk_r} is the estimated intensity for a spot r of protein feature k in the mini-array j on the i^{th} slide. The transformation is defined as $\log_2(\hat{y}_{ijk_r}) = \log_2(y_{ijk_r} - \hat{\alpha})$. However, $\log_2(\hat{y}_{ijk_r})$ is not defined for $y_{ijk_r} \leq \hat{\alpha}$ or for intensities where $y_{ijk_r} - \hat{\alpha} < 0$. Additionally, the asymptotic variance of $\log(\hat{y}_{ijk_r})$ is approximately constant for large y_{ijk_r} but approaches infinity as $y_{ijk_r} \rightarrow 0$. Log transformations are unsuitable for negative values, tend to be inflated for low values that are in the range of the background (175), and are not sensitive to other sources of MVD.

4.3.4.2 *Cyclic Loess Normalization.*

This stabilizes the MVD between slides by applying a pairwise non-linear local regression (LOESS). It utilizes a pseudo Bland-Altman (MA) plot defined as the average [A] versus the difference [M] between the intensities on two independent slides, repeated for N number of slides (113). It yields an 'average' array that is used as a reference to adjust

the MVD across all slides. It can be applied to both raw and log-transformed data (114). Cyclic loess performs a pairwise normalization on all distinct pairs of slides utilising the MA plot and LOESS smoothing. The MA plot in single-colour microarrays for a pair of arrays is the scatter plot of average the intensity values [A] from both arrays vs. difference in expression values [M] of the same arrays. The intensity-dependent differences are first estimated and the differences subsequently regulated by centering the LOESS line to zero (110,113).

Given $\hat{y}_{ijk_r}$ for a given slide $i = 1, 2, 3, \dots, n$, $M_r = \log_2 \left(\hat{y}_{1jk_r} / \hat{y}_{2jk_r} \right)$ and

$A_r = \frac{1}{2} \log_2 \left(\hat{y}_{1jk_r} \times \hat{y}_{2jk_r} \right)$ where $r = 1, 2, 3, \dots, p$ are the spot intensities for a specific protein.

A LOESS curve is then fitted for the MA differences and M'_r a normalised value for M_r is generated. The spot for each specific protein intensities is normalized as follows

$\hat{y}'_{1jk_r} = 2^{A_r + \frac{M'_r}{2}}$ and $\hat{y}'_{2jk_r} = 2^{A_r - \frac{M'_r}{2}}$ or the logarithm transformed equivalents (113,179). Here

we use the LOESS method of Ballman et al. (57,113).

The underlying assumption in cyclic loess is that there is minimal variation between individual arrays under the conditions being studied. Its application for protein microarray experiments designed to detect high levels of variation in different arrays may thus be limited. We recommend the randomization of samples during the design of the study to ensure there is minimal variation between the arrays.

4.3.4.3 Robust Linear Normalization (RLM).

This method stabilizes the mean-variance dependence (MVD) by using standardized control spots on each slide to adjust intensities across the entire experiment (38). It assumes that the signal detected from control spots remains constant with the exception of technical differences within or between slides. It is the method of choice when a significant amount of variation between samples is anticipated.

Data from control spots are fitted to a robust statistical model using an iteratively reweighted least-squares procedure with a robust “sandwich estimator”, like the median. Fixed effects for each array or slide and positive control proteins are estimated from the statistical model. Sboner et al. recommended using a linear model applied to log-transformed intensities (38). The model $\log_2(\hat{y}_{ijk_r}) = \alpha * \text{Slide}_i + \beta * \text{Block}_j + \tau * \text{Protein}_k + \varepsilon_{ijk_r}$, $\hat{y}_{ijk_r}$ is the background-corrected intensity for the spot r of protein feature k in the block j on the i^{th} slide. α is the slide effect of the slide i , β is the block j effect and τ is the effect of protein feature k ; this helps account for the spotted protein amount and binding affinity of different protein features and $\varepsilon_{ijk_r} = \text{Norm}(0, \sigma^2)$ is the error term. After estimating the best parameters, the transformed values are estimated as $\log_2(\hat{y}'_{ijk_r}) = \log_2(\hat{y}_{ijk_r}) - (\alpha_i + \beta_j)$ (38). We recommend that the control antigens used for normalization are optimized to avoid saturation to facilitate the identification of true technical variation. We used human Ig (IgG and IgM) as controls in our experiments.

As with other logarithm transformations, RLM is not suitable for negative signals values. These are consequently replaced using the moving minimum positive approach (above).

4.3.4.4 Variance Stabilization Normalization (VSN).

The VSN method overcomes the limitations of log transformations by accommodating negative values and minimizing the inflated variance around low signal intensities. It calibrates between-feature variation through shifting and scaling mechanism in which all the data are adjusted.

Huber et al. and Durbin et al. independently proposed the VSN approach which is a variant of the log-transform ($glog2$). A two-component model to explain the proportional increase in the variance with the mean intensity of the proteins was proposed (111,175,177);

$y_{ijk_r} = \hat{\alpha}_{ijk_r} + \mu_{ijk_r} e^{\eta} + \varepsilon_{ijk_r}$, where $\hat{\alpha}_{ijk_r}$ is the background signal and $\mu_{ijk_r} = \hat{y}_{ijk_r}$ is the actual signal. $\eta = Norm(0, \sigma_{\eta})$ and $\varepsilon_{ijk_r} = Norm(0, \sigma^2)$ are the proportional error and background error respectively. However, with background corrected data this can be modelled as

$y_{ijk_r} \approx \mu_{ijk_r} e^{\eta}$. A transformation h is used to produce values such that $Var(h(y_{ijk_r}))$ is approximately independent of the mean, $E(h(y_{ijk_r}))$. In general, for a matrix, μ_{ijk_r} the

function implemented fits a normalisation transformation $\mu_{ijk_r} \rightarrow h(\mu_{ijk_r}) = glog_2 \left(\frac{\mu_{ijk_r} - a_i}{b_i} \right)$

where b_i is the scaling parameter for array i which is always ensured to be positive with

a parameter transformation $f(b) = \exp(b)$, a_i is the background offset included if the

data is not background corrected and $glog_2(u) = \log_2 \left(u + \sqrt{u^2 + 1} \right) = \text{arsinh}(u) / \log(2)$ is the

generalised transformation h . A robust variant of the maximum likelihood estimator for

the 2 parameters is utilised (171). Each slide is treated independently and slide to slide variation is not considered (38).

4.3.5 Batch Correction.

The processing of samples on separate days introduces batch-to-batch variations due to non-specific day to day differences in laboratory conditions or operators (98).

We implement a selection of tools to identify and visualize batch effects such as coloured scatter plots, hierarchical clustering or principal component analysis (PCA). Subsequent analyses can be adjusted to account for batch effects, but the majority are designed for large experiments of at least 25 batches (116). We utilise the Empirical Bayes approach as it accommodates all batch sizes and can utilize both parametric and non-parametric data (98).

4.3.5.1 Empirical Bayes (EB) batch correction using ComBat.

This uses the EB approach to estimate and correct batch effects. It can be applied to high-dimensional data even when the sample size is small. Suppose we have b batches in the data containing n_s samples within a batch w for $w=1,2,...,b$ and a protein $k=1,...,K$ then a location and scale (L/S) adjustment model is assumed;

$Y_{wsk} = \alpha_k + X\beta_k + \gamma_{wk} + \delta_{wk}\epsilon_{wsk}$. Then, the EB batch adjusted data γ_{wsk}^* is then calculated as

follows $\gamma_{wsk}^* = \frac{\hat{\delta}_k}{\hat{\delta}_{wk}^*} (Z_{wsk} - \hat{\gamma}_{wk}^*) + \hat{\alpha}_k + X\hat{\beta}_k$ (98) [4.9.2 Supplementary B for details]. To

perform this, we utilise a wrapper to *SVA's* function *ComBat()* for the batch adjustment that has both the parametric and non-parametric approaches (98).

4.4 Results.

4.4.1 Implementation.

protGear is an R based suite with a range of functions to facilitate protein microarray data pre-processing. It has a built-in user-friendly Shiny® dashboard [Supplementary C-1] and a video link [Supplementary C-2] to assist in real-time processing, visualization and downstream analysis using heatmaps and Principal Component Analysis (PCA). It provides five sequential steps for handling a data table of fluorescent intensities. Importantly, the package enables the inclusion of additional functions that may be deemed useful. A detailed workflow is included in the *protGear_vignette* document in the supplementary or <https://keniajin.github.io/protGear/>

4.4.2 Background Correction

The protGear *background_correct* function implements five different techniques for background correction that are complemented by diagnostic plots. Figure 4-3 shows the example of a background diagnostic plot produced by protGear. As shown in Figure 4-3 similar local background values were observed across the different blocks. The correlation between the foreground and background intensities (medians) on the same array was also low. Therefore, a local background correction approach was selected and applied.

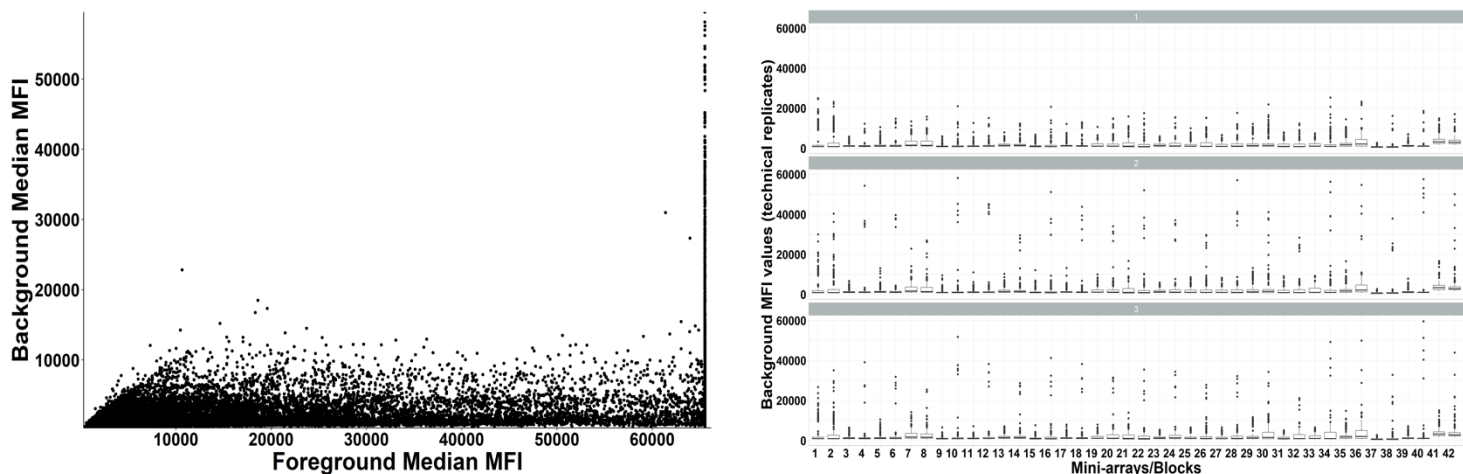


Figure 4-3: Example of background diagnostic plots produced by protGear. (A) is the background MFI vs foreground MFI plot that is useful to assist in selecting the appropriate background correction method. (B) is a boxplot of the blocks/mini arrays categorised into the technical repeats. This plot is important to check whether there is a block artefact in the background MFI values.

4.4.3 Within sample variation and purification tag subtraction.

protGear provides a *cv_estimation* function that is applied to technical repeats to calculate and visualize within-sample variability utilizing a user-defined CV. A filtering algorithm identifies technical repeat spots meeting user-defined criteria e.g., $CV < 20\%$. The function generates a flag variable to enable further scrutiny of non-performing spots. Figure 4-4 illustrates graphics to monitor the CV before and after filtering. A folder is created to store the CV-corrected data in the working directory or the package function, respectively. An average of the replicates is calculated before the subtraction of the intensity measured against the purification tag (optional). Here, we kept 2 of 3 technical replicates with CVs less than 20% and excluded the outlier value.

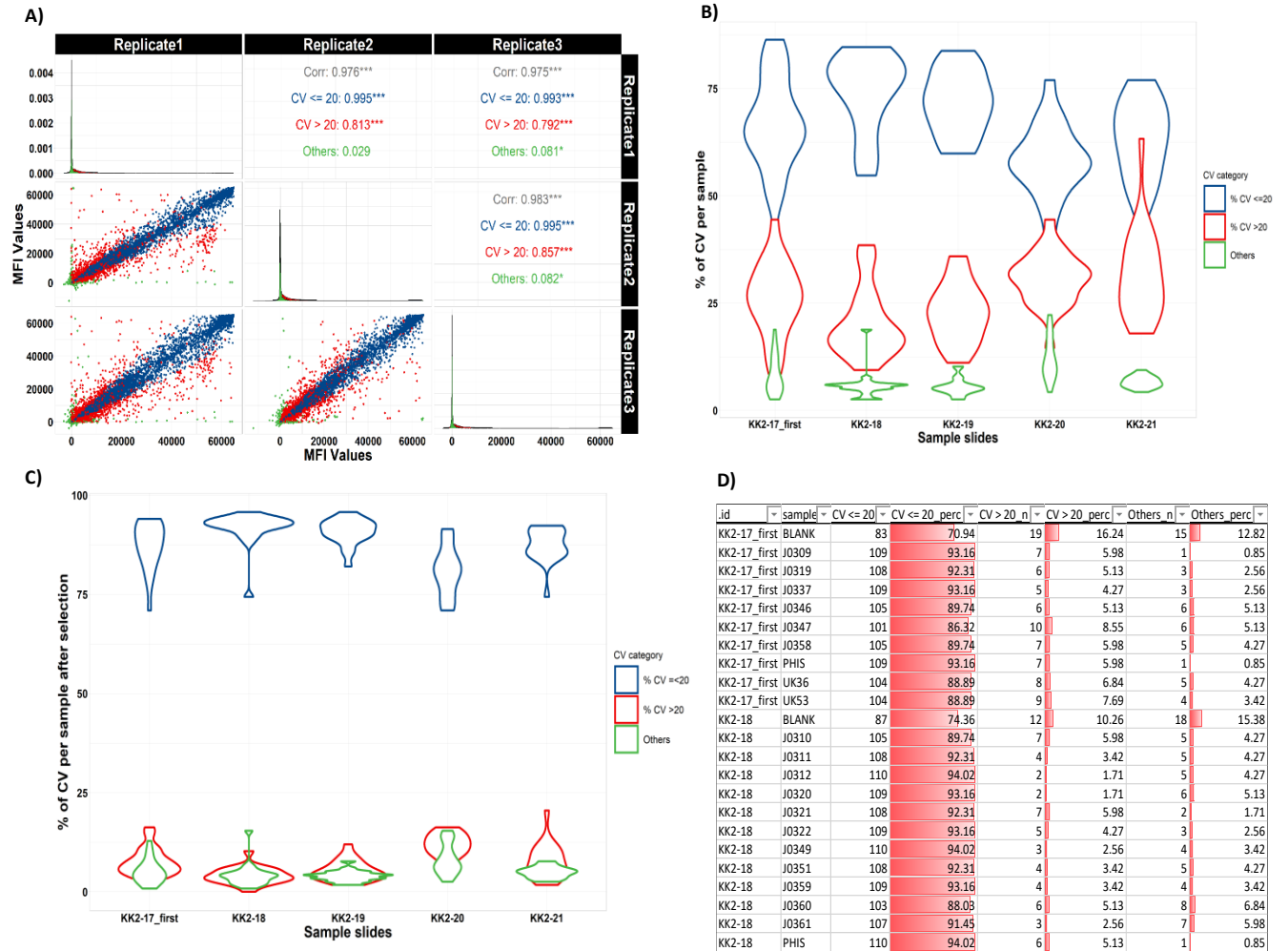


Figure 4-4: The visualization of the CV A) Correlation of the technical replicates, B) Proportion of CV by the CV cut off, C) Proportion of CV after “cv_based_filtering”, D) A static image of an Interactive table to inspect the CV cut off values. The table shows the specific slide id (.id), the serum sample identifier (sampleID), count of CV's < 20% (CV<=20), % of CV's

4.4.4 Normalisation and Batch Correction.

We tested four functions for normalization to identify the one that optimally reduced the MVD. We formally assessed this using the mean versus standard deviation plots (*meanSdPlots*) and coupled this to automatically derived Spearman correlation estimates (*Rho*) and Cox–Stuart or the Mann–Kendall trend tests. The latter quantifies the performance of the normalization. In the example below, we illustrate the versatility of

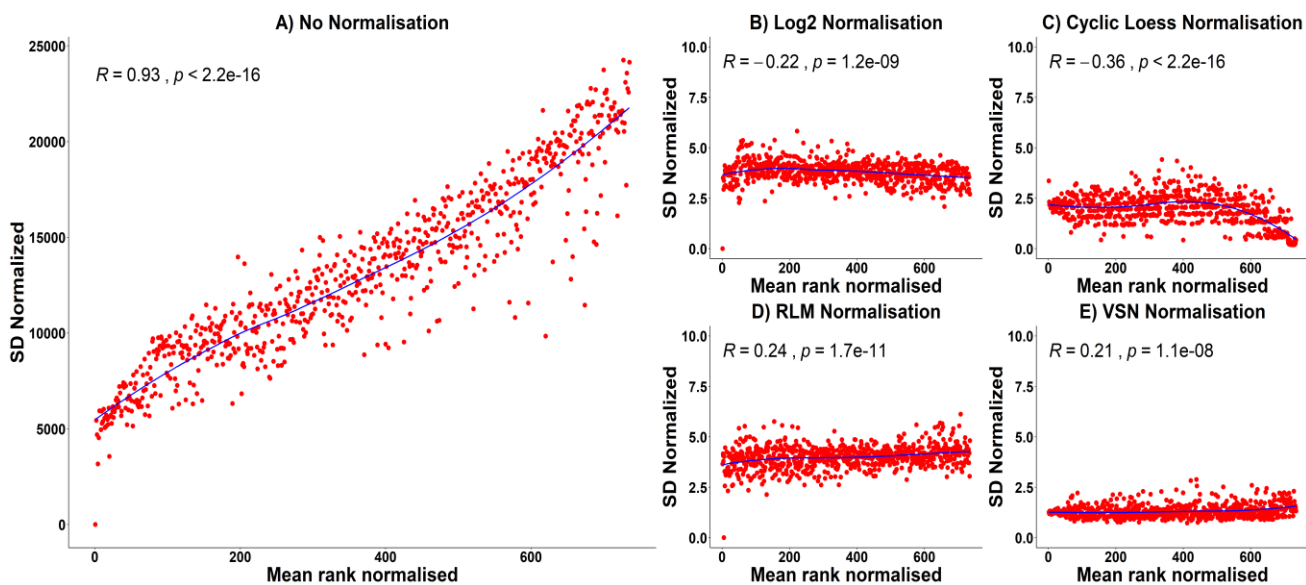


Figure 4-5: Standard deviation vs mean plots (meanSdPlot) of A) Non normalised data B) log₂ normalisation C) Cyclic loess normalisation D) Robust Linear Model normalisation and E) VSN normalisation

the tool across four Methods for normalization. Using these approaches, the MVD reduction led to a drop in the Rho estimate from 0.93 to between 0.2 and 0.3 in this dataset as shown in Figure 4-5 below. Implementation of the log₂ approach led to an inflation of variance for the low MFI (Mean Fluorescent Intensity) values. The cyclic loess and

RLM were not optimal for the experimental design (discussed in the methods). Consequently, the VSN normalization approach was adopted (Figure 4-5) and still

maintained the difference of the controls Supplementary 4-7. We then proceeded to investigate batch effects using *ComBat* from the SVA package. The day-to-day dependence was noticeably reduced by *ComBat* batch correction [Supplementary 4-1] for day m1 and m2.

4.5 Discussion:

protGear is an open-source one-stop integrated data pre-processing suite specifically tailored to address the systematic and non-systematic sources of bias in high throughput protein microarray experiments. It can be adapted to a range of design formats and is compatible with the majority of commercial protein microarray platforms. It provides a choice of functions to address each major source of bias, outlines the theoretical background necessary to guide selection and generates custom graphical and tabular outputs to support data interrogation and interpretation. It is coupled with a user-friendly interactive web platform that eliminates the need for specialized programming skills. In line with other open-source platforms, protGear can be adapted to incorporate new functions either by adding custom scripts or contributing to its online repository.

The localization and distribution of background signal intensities guide the selection of the appropriate correction strategy (104–106). Our experimental design comprised multiple mini-arrays and slides. We did not detect any significant background pattern within the mini-arrays or block and consequently selected the local background correction method. Although unmodified background intensities are subtracted from the foreground resulting in unbiased estimates (105), negative values are often generated, and the consistency of the results across different mini-arrays and slides needs to be assured.

Additional methods including the novel *half moving minimum* we developed that overcome these limitations have been included (57).

The inclusion of two or more technical replicates is vital for the robust analysis of within-sample variability (62). Caution in interpreting the CV is recommended as small differences in signal intensities between replicates at the lower end of the detection threshold can yield misleadingly high CVs. To overcome this, we implemented a novel function that enables users to set signal intensity thresholds below which high CVs can be ignored.

The importance of adopting an appropriate strategy for normalization is critical and requires the coupling of the experimental hypothesis with the mathematical assumptions underpinning the methods(180). Key design issues to consider are the extent of sample biological variation anticipated under the conditions being investigated and technical considerations in the array design such as the inclusion of appropriate controls. Adopting the wrong method increases the probability of detecting false-negative and false positives.

The RLM normalisation approach that was specifically developed for protein microarrays (38) requires that the secondary control normalisation proteins yield constant signals across and within slides. These control proteins are expected to have a low CV and can be used for both normalizations and evaluating the slide variability (38). Although the MVD was significantly reduced with all methods tested, the VSN method was optimal. protGear provides a novel powerful single platform that empowers users to sequentially interrogate all these options to determine the optimal solution for the data in question.

The easy-to-use interface accommodates multiple data formats, saves enormous amounts of time and generates high-quality visual outputs that facilitate rapid decision making.

Batch effects create an additional unwelcome source of variation that could further reduce statistical power. These must be considered when data are processed at different times, by different users and on different instruments, among others (99). An important consideration that particularly applies to large cohort studies is the random processing of samples to ensure that batch effects do not exaggerate pre-existing genuine biological variation. For example, in the context of malaria sero-epidemiology, age and geographical location are critical determinants of antibody levels (16). The processing of samples of young children at one time-point, and those of older children at another, could inadvertently lead to enhanced differences in either group that were unrelated to the true underlying biological variation. A similar effect could occur when samples from settings with differing malaria transmission intensity are analysed in the same experiment but at separate times.

A limited number of batch correction approaches have been proposed and these typically accommodate experimental designs with a minimum of 25 batches. We adopted the ComBat batch correction function from the “sva” package in R since it has been reported to be robust to outliers in small batch sizes (98).

protGear provides a state-of-the-art, one-stop adaptable workflow for protein microarray data pre-processing. It can be coupled to software such as Sweave or knitr (181) for report generation and sharing along with the raw data to promote data reproducibility.

The user-friendly, interactive, web-based and graphical interface requires limited R-experience and will enhance broad uptake in the infectious disease and vaccine development community.

4.6 Availability and implementation

The protGear R package is publicly available in the GitHub repository (<https://github.com/Keniajin/protGear>).

4.7 Funding

This research was commissioned by the National Institute for Health Research (NIHR) Global Health Research programme (16/136/33) using UK aid from the UK Government. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

K.M, F.O, S.K are funded by the National Institute for Health Research (NIHR) Global Health Research programme (16/136/33). F.O. is supported by a Sofja Kovalevskaja Award from the Alexander von Humboldt Foundation (3.2 - 1184811 - KEN - SKP) and an EDCTP Senior Fellowship (TMA 2015 SF1001) which is part of the EDCTP2 Programme supported by the European Union. S.K and E.M are supported by the DELTAS Africa Initiative under Initiative to Develop African Research Leaders (IDeAL) and DELTAS Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) respectively.

4.8 Declaration of Competing Interest

The authors declare no conflict of interest.

4.9 Supplementary

4.9.1 Supplementary A: Definition of terms and quantitation results file contents

4.9.1.1 Quantification results file contents

A quantification results file contains the data for multiple parameters quantified and measured in relation to a given spot. Importantly, it has the location details of each spot as defined in the input file. Below we show examples of parameters generated by the GenePix™ Pro software. The first part is the header which describes basic file information and all block properties. Importantly, it has the location details of each spot, the row (Y coordinate) and column (X coordinate).

```
ATF 1.0
30 41
"Type=GenePix Export 3"
"BlockCount=4"
"DateTime=2017/08/29 13:39:04"
"Settings="
"GalFile=\Multicenter study\gal file\gal file_4`.gal"
"PixelSize=10"
"Wavelengths=635"
"ImageFiles=Z:\Multicenter study\28_08-2017\datascan_28-8\KK2-07.tif 0"
"NormalizationMethod=None"
"NormalizationFactors=1"
"JpegImage="
"StdDev=Type 1"
"RatioFormulations=W1/W2 (635/594)"
"FeatureType=Circular"
"Barcode="
"BackgroundSubtraction=LocalFeature"
"ImageOrigin=0, 0"
"JpegOrigin=1740, 900"
"Creator=GenePix Pro 7.3.0.0"
```

```

"Scanner=GenePix 4300 [141110]"
"FocusPosition=0"
"Temperature=0"
"LinesAveraged=1"
"Comment="
"PMTGain=400"
"ScanPower=100"
"LaserPower=0.66"
"Filters=Standard Red"
"ScanRegion=0,0,2200,7200"
"ScanArea=1"
"Supplier="
"Flags" "Normalize" "Autoflag" "Block" "Column" "Row" "Name" "ID" "X" "Y" "Dia." "F635 Media
n"
0 0 0 1 1 1 Landmark 1F1 2310 1130 150 65535
0 0 0 1 2 1 MSP3.6 1B3 2540 1130 150 13182
0 0 0 1 3 1 ETRAMP10.2 1H3 2770 1130 150 2310
0 0 0 1 4 1 GAMA 1D5 3000 1130 150 18395

```

- **Block** - the block number of the feature. The array has a number of blocks. For example GenePix assigns block numbers row wise from left to right. The top left block is assigned block #1 as shown below.
- **Column** - the column number of the feature.
- **Row** - the row number of the feature/protein.

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21

Table 1: Block numbering of GenePix

- **Name** - the name of the feature/protein derived from the Array List (up to 40 characters long, contained in quotation marks).
- **ID** - the unique identifier of the feature/protein derived from the Array List (up to 40 characters long, contained in quotation marks).
- **X** - the X-coordinate in nm of the center of the feature/protein-indicator associated with the feature/protein, where (0,0) is the top left of the image.
- **Y** - the Y-coordinate in nm of the center of the feature/protein-indicator associated with the feature/protein, where (0,0) is the top left of the image.
- **Dia.** - the diameter in *nm* of the feature/protein-indicator.
- **F635 Median** - median feature/protein pixel intensity at wavelength (635 nm). This changes according to the wavelength used to scan.
- **"B635 Median"** - background pixel intensity at wavelength (635 nm). This changes according to the wavelength used to scan.
- **Flags*** - the type of flag associated with a feature/protein.

4.9.1.2 Definition of terms

- **Scanner:** an instrument used to detect and quantify the intensity of fluorescence of spots on a microarray slide,
- **Mini-array/Sub-array/Block** - a number of spots on the array representing data for a given sample. A number of blocks can form a mini array.
- **Spot** - a small round mark differing in fluorescence from the surface around it.
- **Pixel** - a small area of illumination on a scanned screen. A spot is made up of a number of pixels.
- **Sample ID file** - A file with names of samples with their corresponding blocks as they were arranged on the slide

4.9.2 Supplementary B: ComBat Batch Correction

The processing of samples on separate days introduces batch-to-batch variations due to non-specific day to day differences in laboratory conditions or in operators (98,115). Here we show

how the evaluation of EB batch adjusted data γ_{wsk}^* is calculated from

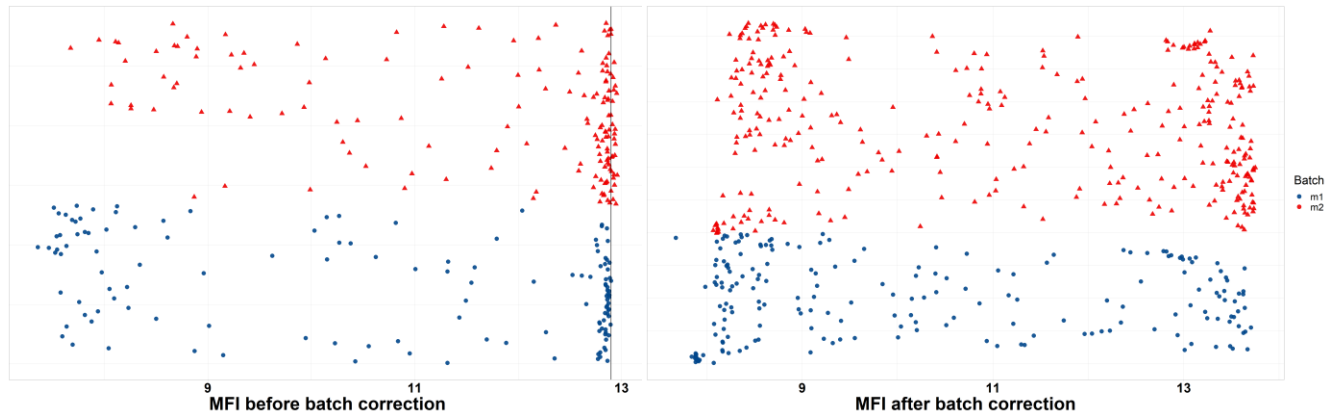
$$\gamma_{wsk}^* = \frac{\hat{\delta}_k}{\hat{\delta}_{wk}^*} (Z_{wsk} - \hat{\gamma}_{wk}^*) + \hat{\alpha}_k + X \hat{\beta}_k \quad (98).$$

Application: Suppose we have b batches in the data containing n_s samples within batch w for $w=1,2,\dots,b$ and a protein $k=1,\dots,K$ then a location and scale (L/S) adjustment model is assumed; $Y_{wsk} = \alpha_k + X \beta_k + \gamma_{wk} + \delta_{wk} \varepsilon_{wsk}$ and that the errors $\varepsilon_{wsk} \sim \text{Norm}(0, \delta_k^2)$. The above model is defined with non-standardized data and this implies that $\alpha_k, \beta_k, \gamma_k$ and δ_k^2 differ across proteins (98). This can bias the estimates of EB estimates of the prior distribution of batch effect and reduce the amount of systematic batch information that can be borrowed across proteins. However, this can be avoided by standardizing the data or applying a non-parametric approach (117).

Assuming, standardized data $Z_{wsk} \sim \text{Norm}(\gamma_{wk}, \delta_{wk}^2)$ with parametric forms for the prior distribution of the batch effects $\gamma_{wk} \sim \text{Norm}(Y_w, \tau_w^2)$ and $\delta_{wk}^2 \sim \text{InverseGamma}(\lambda_w, \phi_w)$ the batch effect parameters, γ_{wk} and δ_{wk}^2 are estimated by the conditional posterior means (98). The EB batch adjusted data γ_{wsk}^* is then calculated as follows

$$\gamma_{wsk}^* = \frac{\hat{\delta}_k}{\hat{\delta}_{wk}^*} (Z_{wsk} - \hat{\gamma}_{wk}^*) + \hat{\alpha}_k + X \hat{\beta}_k \quad (98).$$

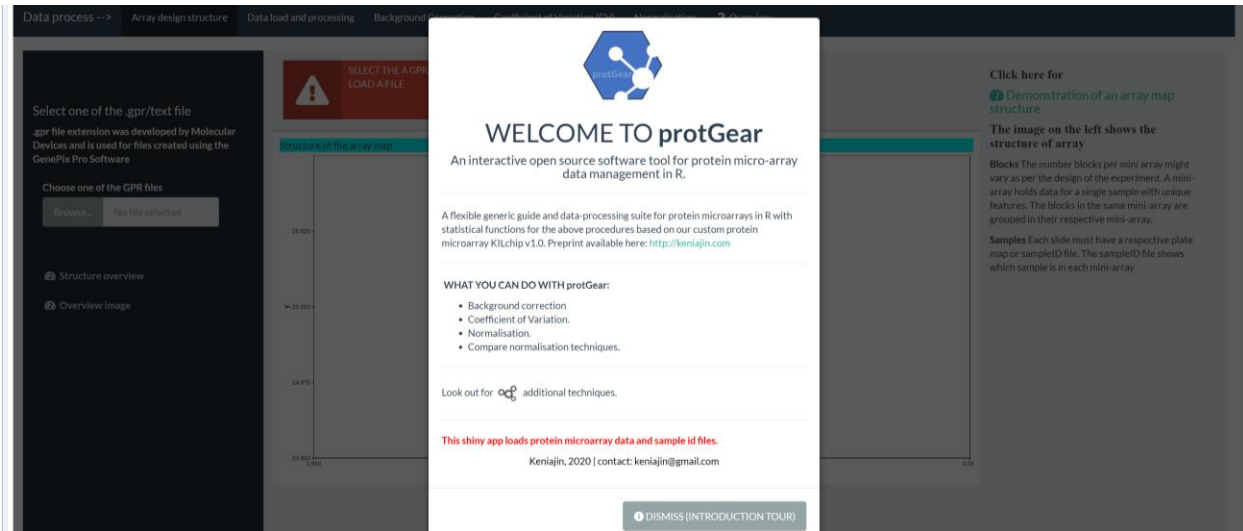
To perform this, we utilise a wrapper to SVA's



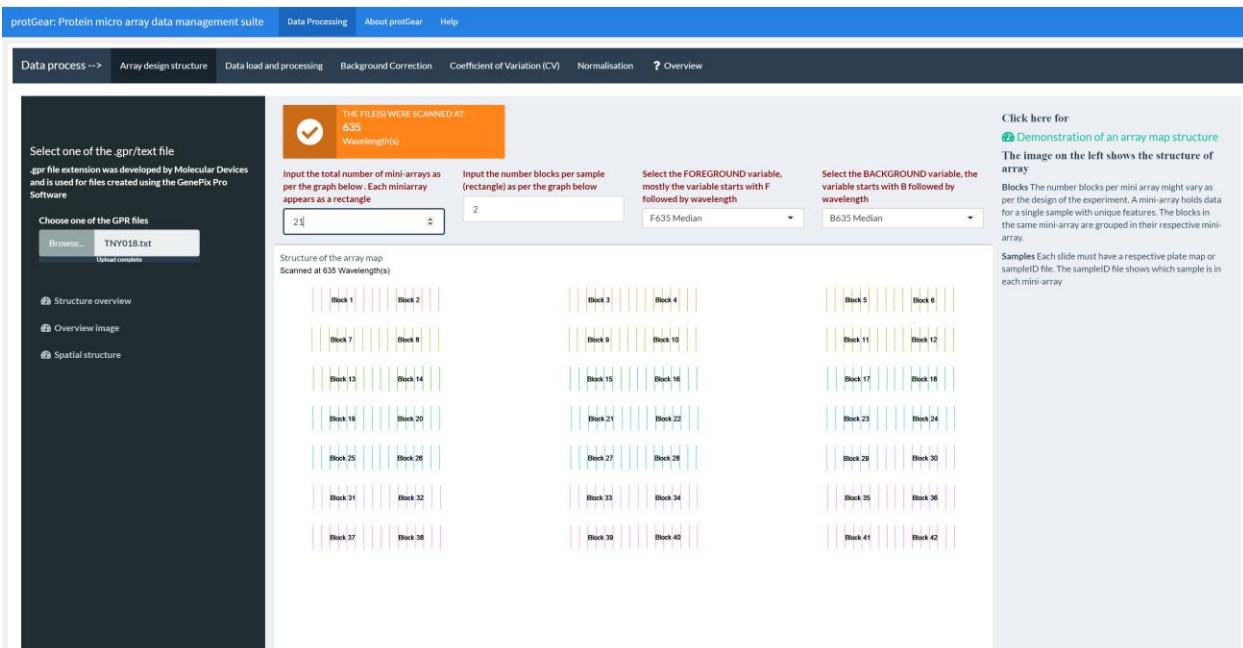
Supplementary 4-1: Comparison of Batch m1 and Batch m2 before and after batch correction using Combat() approach

function ComBat() for the batch adjustment that has both the parametric and non-parametric approaches (117).

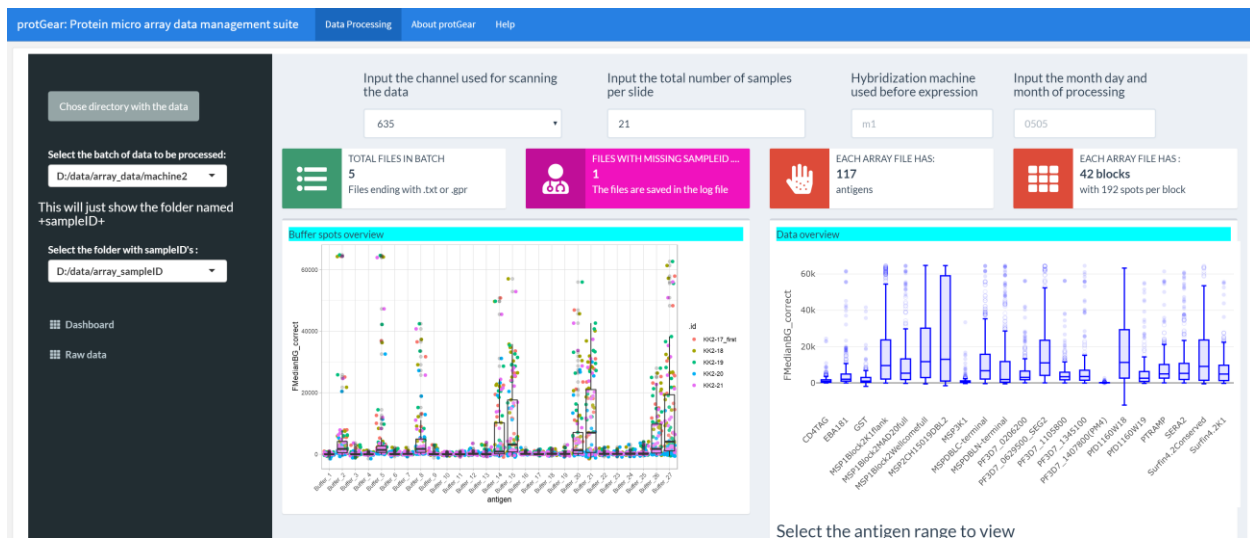
4.9.3 Supplementary C-1



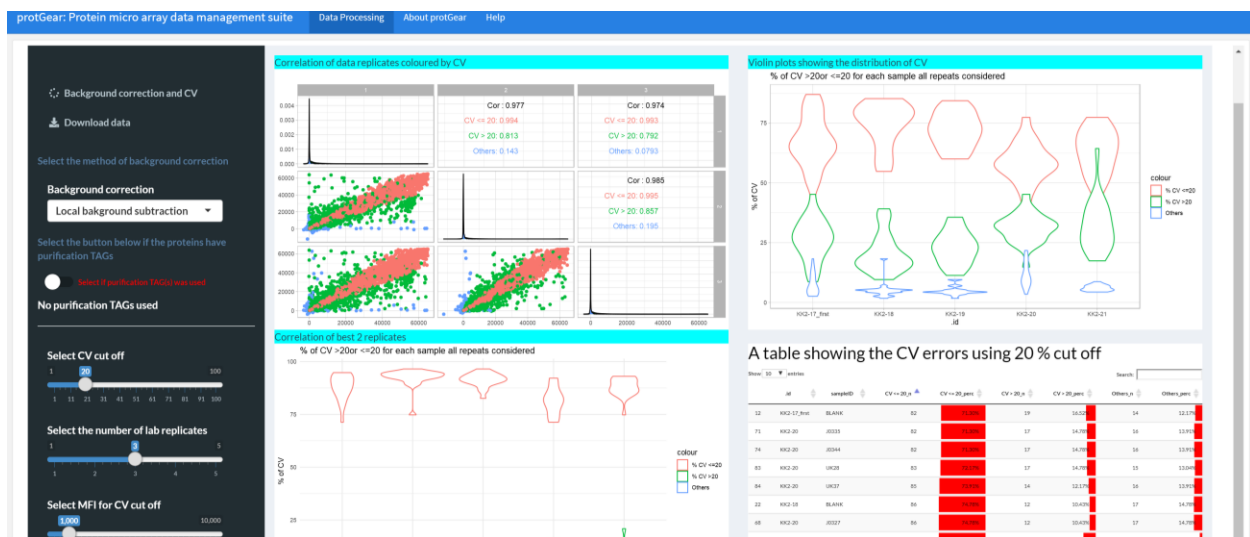
Supplementary 4-2: The protGear Shiny® app interface on load



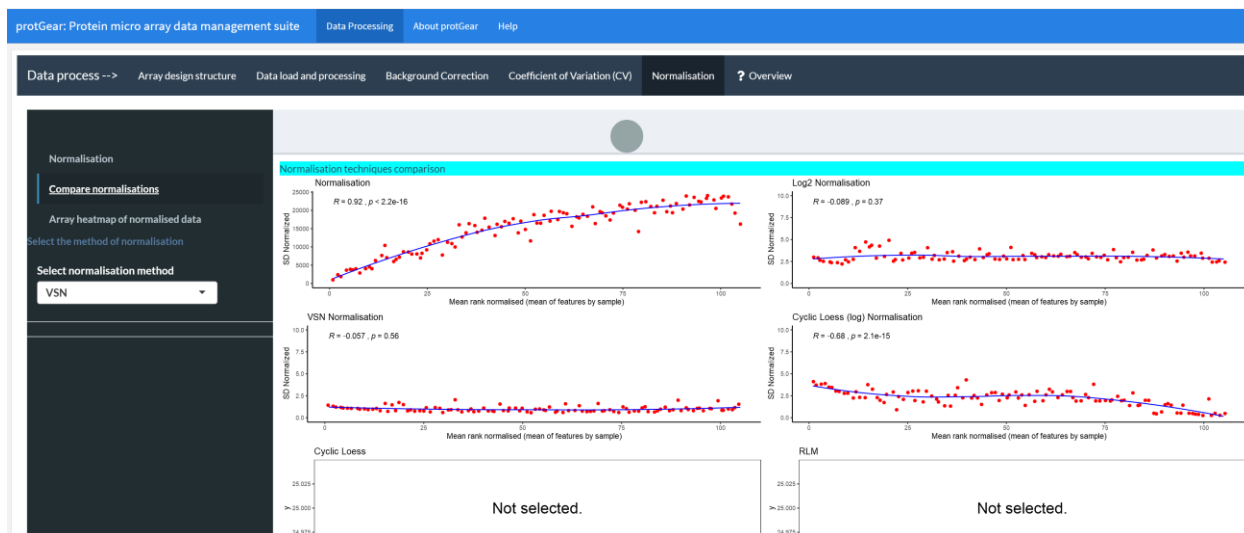
Supplementary 4-3: Checking the array design structure



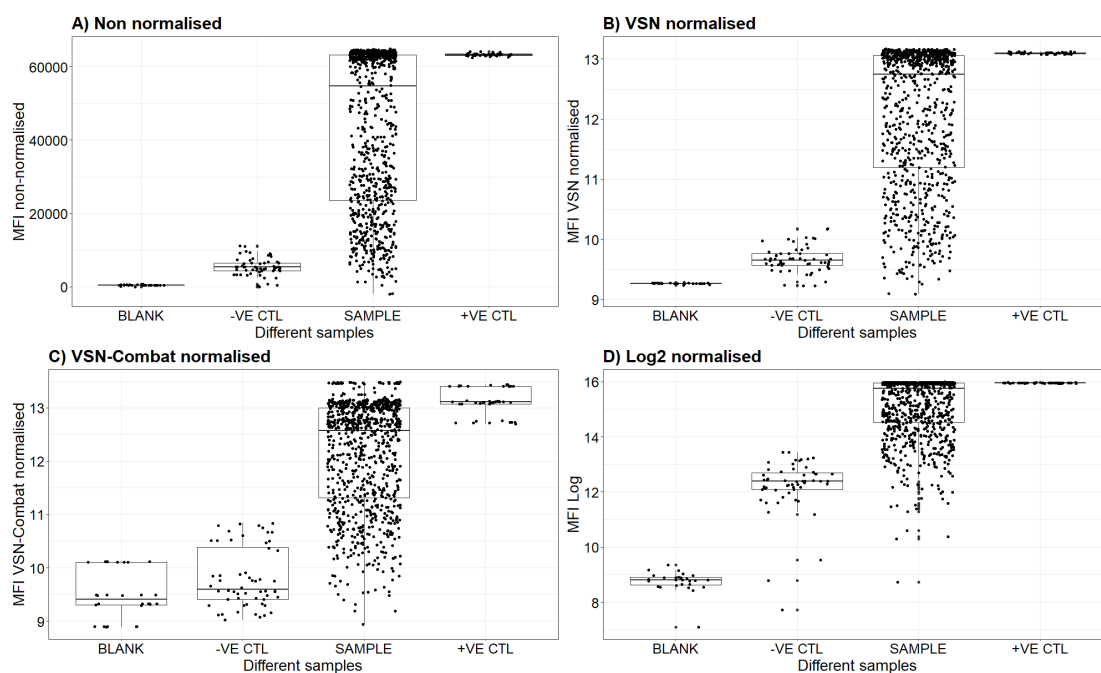
Supplementary 4-4: Loading the array data files



Supplementary 4-5: The visualization of the CV on the dashboard



Supplementary 4-6: Normalisation



Supplementary 4-7: Performance of VSN and log normalisation A) non-normalised data, B) VSN without batch correction, D) Log normalization on blank which is the sample buffer labelled as BLANK , negative (malaria naïve sera) controls as -VE CTL, serum samples labeled as SAMPLE and positive (malaria immune sera) controls labeled as +VE CTL. This is a sample dataset of the KILchip v1.0 which was used as the test data in this tool.

4.9.4 Supplementary C-2

<https://www.sciencedirect.com/science/article/pii/S2001037021001562#m0005>

4.9.5 Supplementary D:

	Comparison between currently available protein microarray analysis tools						
	protGear	Protein Microarray Analyser (PMA)	Protein Array Analyzer (PAA)	protein array web exploreR (PAWER)	RPPanalyzer	RPPASPACE	protein microarray database (PMD)
Language	R and Shiny	Java.	R	R and JavaScript	R	R	PHP and R
Source code contribution	✓	X	X	✓	✓	✓	X
Vignettes: long-form documentation	✓	X	✓	X	X	✓	X
Interactive GUI	✓	✓	X	✓	X	X	✓
Mutiple methods comparison	✓	X	X	X	X	X	X
Background corection	Local and Global Background, moving minimum, half moving minimum, normexp and log linear	Local Background Subtraction	Not specified	Local Background Subtraction	Limma background correct and addmin	Local background subtraction	Not Clear
Normalization	Log, RLM, VSN and Cycli Loess	Pin to Pin and Array to Array Normalization	RLM	RLM	proteinDye,row, housekeeping and extValue	spatial corrections	Not Clear
Downstream Analysis	✓	X	✓	X	X	X	✓
Accessibility	✓	https://sourceforge.net/projects/protein-microarray-analyser/	https://bioconductor.org/packages/release/bioc/html/PAA.html	https://biit.cs.ut.ee/pawer/	https://cran.r-project.org/web/packages/RPPanalyzer/index.html	https://cran.r-project.org/web/packages/RPPASPACE/index.html	http://www.proteinmicroarray.cn/
License	GPL-3	No License Specified	3-Clause BSD License	GNU GPL v2 license	LGPL-3	Artistic License 2.0	Creative Commons Attribution 4.0 International
Last updated	2021	2018	2020	2020	2020	2020	2021

CHAPTER 5 : FEATURE SELECTION

This study so far has introduced methods to handle the misclassification bias in malaria case definition and techniques to handle the systematic and non-systematic biases that come with high throughput protein micro array data. This chapter shows and describes in detail the feature selection approaches for “*Defining the merozoite targets of protective immunity against Plasmodium falciparum malaria through multi-center cohort studies*”.

This chapter covers two main approaches, the first approach involves using dichotomous outcome and secondly, utilizing probability of having fever given a certain parasite density as the outcome in feature selection. For the dichotomous outcome, we use data for all the multicenter sites, but for the probability outcome we use data from Junju, Kilifi only. This is because most of the cohort studies did not have a training data (a sample that comes from the component of the mixture corresponding to children without clinical malaria but who may have parasites) to define the probability outcome as described in detail in Chapter 2 and 3. For all the multicenter analysis, a network model was developed to visualize the interplay of the different top features identified in each site.

This analysis will lead to a shared first co-author paper titled *Defining the merozoite targets of protective immunity against Plasmodium falciparum malaria through multi-center cohort studies*. The code for analysis and further model diagnostics is hosted here https://github.com/Keniajin/multicenter_feature_selection

5.1 Introduction

The process of identifying important biological features is challenging especially for outcomes prone to differential misclassification and significant biological complexity. Misclassification can lead to bias in describing, framing, selecting, and comparing cases and controls. Consequently, researchers across the globe, are facing challenges in making data-driven decisions and recommendations. This is particularly relevant for studies on natural immunity to malaria that has this misclassification complexity due to asymptomatic malaria (1,2).

The last 2 decades have witnessed significant reductions in cases and deaths resulting from malaria in many parts of the world (5). These gains are largely attributed to the scale up of rapid diagnostic tests, use of effective medication and insecticide treated bed-nets. However, despite sustained use of these interventions in the past 5 years, estimates from WHO show a lower rate of decline compared to the first 15 years since the beginning of 2000 (13). This is further exacerbated by increasingly reports of the emergence and spread of parasite drug and mosquito insecticide resistance. The development and deployment of a highly efficacious malaria vaccine will be instrumental if further reductions and even possible elimination and eradication of malaria is to be achieved (48). The most advanced malaria vaccine candidate, RTS,S, that is currently under implementation trials in multiple African countries, offered modest protection against disease 28.3% (95% CI, 23.3 to 32.9) (182), that is short-lived when it was evaluated in efficacy trials. Whilst this marks a major achievement, it falls below the target set by the World Health Organization's Malaria Vaccine Technology Roadmap, of greater than 75%,

when used in malaria-endemic areas. The development of a second-generation vaccine with better efficacy remains a top priority of malaria research (48).

A strong rationale for developing and improving malaria vaccines comes from epidemiological studies in malaria-endemic countries. These studies show that individuals continuously exposed to malaria gradually develop “natural immunity” that initially prevents death from severe malaria, and subsequently recurrent clinical episodes of mild malaria (41,162). Therefore, identifying the targets underlying naturally acquired immunity is of importance. However, this remains challenging from a biological point of view due to the complex life cycle of malaria parasites and their interaction with the human immune system have, to date, proved very difficult to unravel. For this reason, data arising from malaria immunity studies are so closely entwined with complexity and multicollinearity that it is difficult to dissect the useful information from noise using classical analytical approaches.

Here we use various machine learning approaches to adequately handle the heterogeneity of malaria transmission within and between sites to perform feature selection of important correlates of disease protection. To handle the misclassification bias, we compute case definitions using regression models and compare the performance of utilizing a binary and probability outcome in objectively identifying correlates of antibody mediated protection. Correlates of protection or important immune signatures are measurable parameters that a person is protected against developing a disease. As discussed in detail chapter 3, logistic power model and Bayesian latent class models are

used (17,19) to define malaria outcomes to reduce the misclassification bias in high malaria transmission sites.

5.2 Methods

5.2.1 Epidemiological Data

As described in detail in Chapter 2, we use data from 14 cohorts with 8197 samples in Africa as shown below.

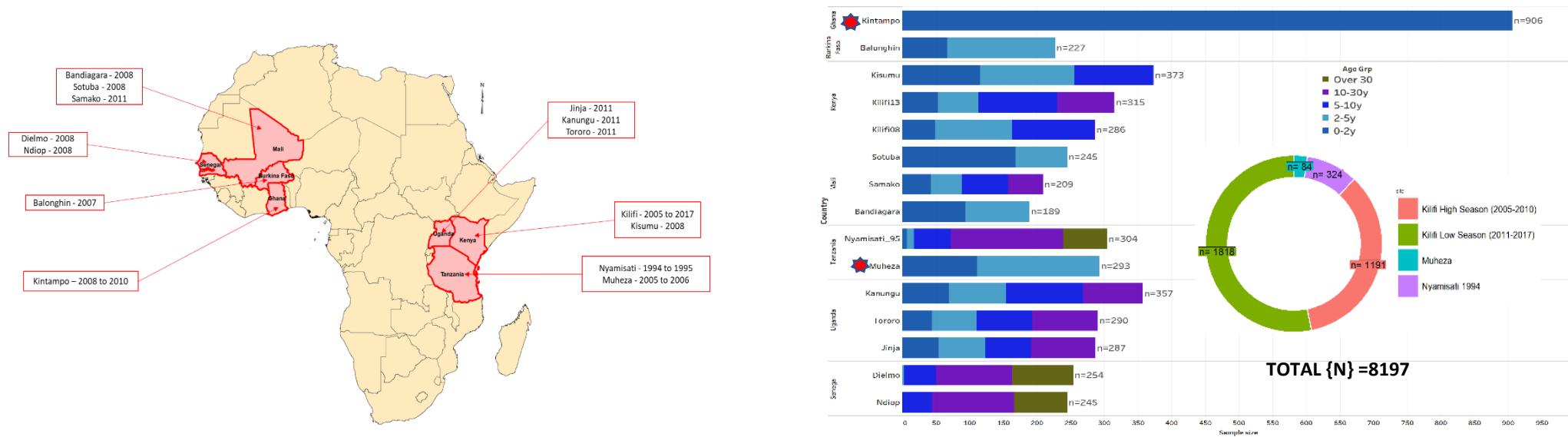


Figure 5-1: The multicentre cohorts from West and East African countries

As highlighted in the Figure 5-1 above, Kintampo and Muheza from Ghana and Tanzania respectively were designed as birth cohorts. The sites had different transmissions and were either categorised into either high, medium, or low transmission depending on parasite positivity during the cross-sectional bleed. Malaria positivity (dichotomous outcome) was defined using a parasite cut-off estimated using logistic regression model. Specifically, for Junju, Kilifi we estimated the probability of fever given a certain parasite density using the Bayesian Latent class models as described in Chapter 3.

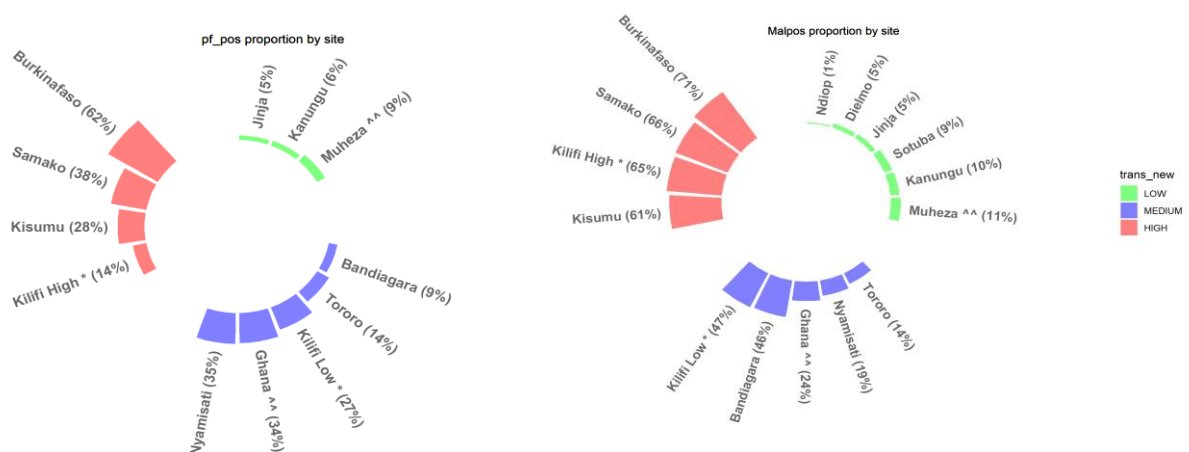


Figure 5-2: Parasite positivity and malaria positivity for the multicentre cohorts

5.2.2 Protein micro-array data

The epidemiological data was then merged with the protein micro-array laboratory data processed in Kilifi using KILchip v1.0, a protein microarray chip designed to enable the simultaneous detection of antibodies. Over 100 *Plasmodium falciparum* proteins were printed in triplicate on a slide divided into 24 mini-arrays, defined as the region allocated to a discrete sample. Each mini array had 2 blocks with 192 spots per block as described

in detail by Kamuyu et al (54). The slides were scanned at a PMT gain adjustment of 400 with a scan power of 100 which gave optimal MFI values. These scanning parameters were selected after graphically investigating the SNR of MFI values from several slides scanned at different combinations of the laser power and PMT levels (102).

To eliminate the systematic and non-systematic bias (37,38,57) associated with protein micro array, empirical data pre-processing was performed (54,87). Systematic biases refer to artefacts that may occur due to experimental and handling procedures. Non-systematic biases on the other hand refer to random errors that may occur due to slight variations during the experiment.

The data pre-processing involved the procedures as shown workflow Figure 5-3 using protGear developed in Chapter 4 (87,99). The local background correction to handle background noise was followed by TAG subtraction and within sample variation check using a coefficient of variation cut-off of 20%. The sample repeats with a CV of greater than 20% were flagged followed by using "*best_CV_estimation()*" to select the best replicates since our experiment had more than 2 replicates (38,87). After CV based filtering, we did VSN normalisation (111) combined with Combat batch (117) correction were used to handle mean variance dependence and batch effects respectively.

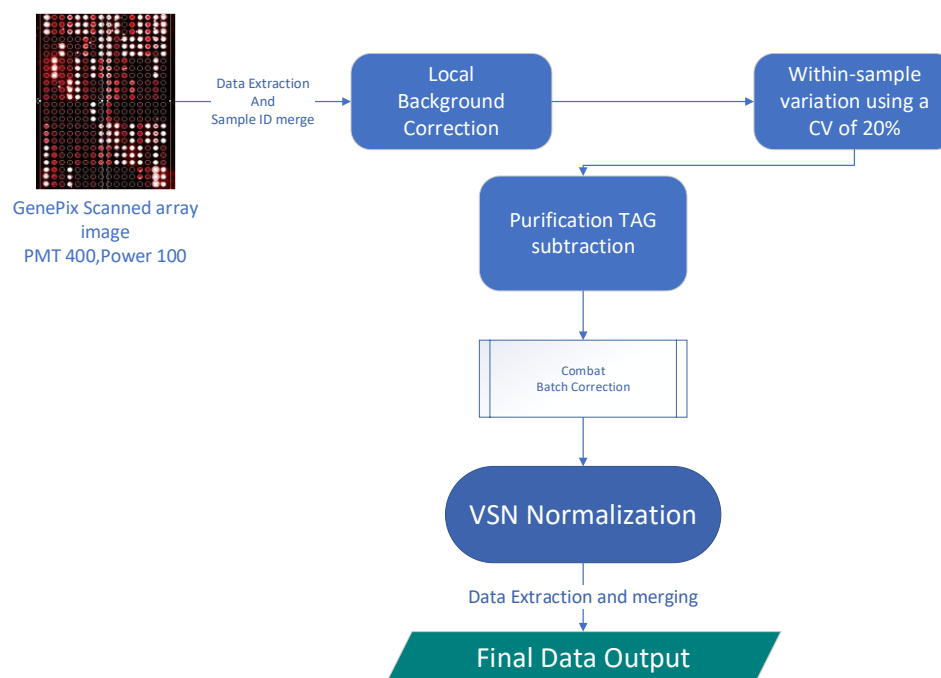


Figure 5-3: Antibody data processing workflow for the laboratory data.

The final data had measurements of 87 proteins Figure 5-4, after excluding the antibodies that were tagged with MBP and GST since they did not perform well after TAG subtraction (TAG MFI values were very high) [Supplementary Figure 5-2]. To check the final data before analysis we empirically and graphically investigated the buffer and blank spots [Supplementary Figure 5-3].

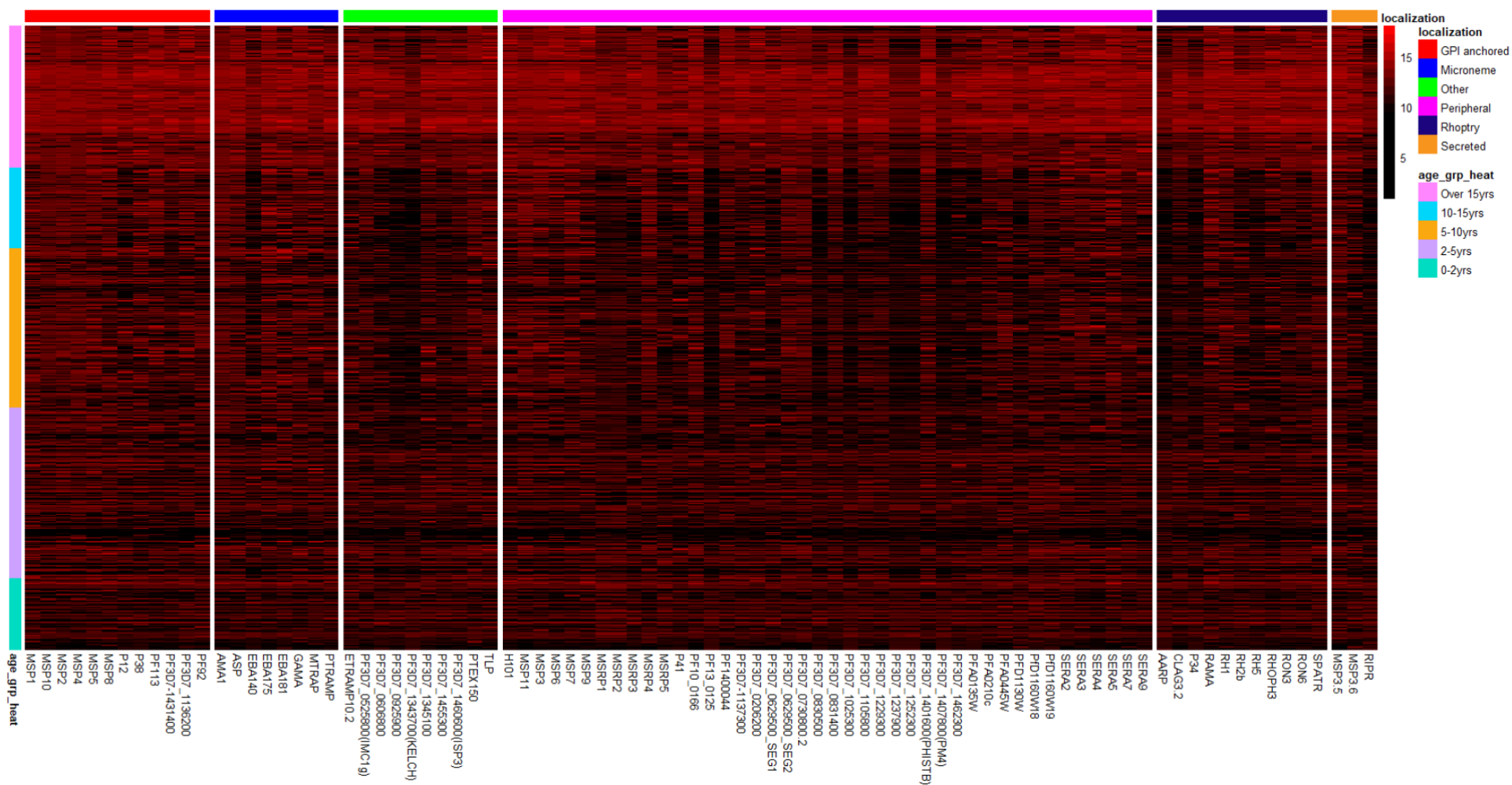


Figure 5-4: Protein micro array data

5.2.3 Feature selection using the Binary outcome

For this section we use the Binary outcome defined using a parasite threshold plus fever defined using the logistic power model. The logistic power model is one of the proposed approaches to handle the misclassification bias that comes with malaria disease (18). Key features identified as correlates of disease protection were selected using a combination of classical modified Poisson regression and supervised machine learning approaches implemented in R v4.0.0 (71) and Python 3.8.5 (183). The supervised machine learning approaches utilised were Random Forest (136) to handle the overfitting bias through Bagging, Tree-Boosting with Gaussian Process and Mixed Effects Models (139); to account for the site random effect and this approach is robust to outliers in and multicollinearity among predictor variables, and Bayesian Lasso feature selection that applies the lasso penalty to select the important parameters (125). After fitting the models, a network model was created to visualize the top feature among the different multicentre sites (184).

5.2.3.1 Modified Poisson Regression (MP)

This approach is a classical regression approach with a categorical outcome adjusting for known confounders followed by ranking the important features using the risk ratio estimated. Given that x_{ij} is the “*malpos*” dichotomous malaria variable from site j with an underlying risk function of $\pi(x_{ij})$, then the modified Poisson regression was defined as,

$$\log[\pi(x_{ij})] = \alpha + \beta x_{ij}; \quad \pi(x_i) > 0 \quad (5.1)$$

We adjusted for the individual's overall response that was estimated using the Principal Component Analysis approach in reference to the PC loading of the negative/blank control spots. The individuals 4 SD's away from the mean of PC1 values of negative/blank control spots were categorized as high responders while the others were categorized as low responders. This variable biologically accounted for the transmission, age and schizont confounding effect and empirically reduced the number of parameters in the model. The relative risk (RR) for each estimated by $RR = \exp(\hat{\beta}) = \frac{an_0}{cn_1}$ (43). The estimated risk ratios of the individual proteins were then ranked in a descending order, and then the top k features were selected. One of the limitations of classical regression based on comparisons between population-level mean responses often fail to convey information about the robustness of the derived associations and how sensitive they are to even small changes in the observed data. Additionally, this approach has a limitation of adjusting the potential interplay between different proteins.

5.2.3.2 Random Forest (RF)

The random forest (RF) approach is built on Classification and Regression Trees (CART) a statistical method for constructing decision trees for both regression and classification problems. Specifically, the RF approach aggregates a collection of decision trees by pooling a set of predictors in this case proteins using Bootstrap aggregating (Bagging) method developed by Breiman et al. (69). The number of trees ("*ntrees*") fit in each bootstrap and the number of candidates/features ("*mtry*") randomly sampled in each split are the main parameters of interest in the RF model (136). The values that significantly

reduced the OOB error using 10-fold cross-validation (CV) were selected as the optimal parameters for *mtry* and *ntrees*. The randomForest v 4.6-14 (70) and caret v 6.0-88 (185) packages in R v4.0.4 (71) were used to select the optimal parameters.

The important variable in a random forest is based on the quantification of the importance of their effects on the response variable. The most used measure of importance of a given variable is the average increase in the error made by a tree among trees in a forest when the observed values of that variable are randomly permuted in the OOB samples (69,136).

The importance index (Mean Decrease Accuracy -MDA) $VI(X_j)$ of variable X^j ; $j \in \{1, \dots, p\}$ was defined by summing the difference between the average error of a tree on the perturbed OOB sample and that on the OOB sample over the repeat operations of all the bootstrap samples (69,136).

$$VI(X_j) = \frac{1}{q} \sum_{\ell=1}^q (errOOB_{\ell}^j - errOOB_{\ell}); \quad (5.2)$$

where $errOOB_{\ell}$ is the error made on OOB_{ℓ} by the tree build on a certain bootstrap sample say $L_n^{\Theta_{\ell}}$. OOB_{ℓ} is the perturbed sample that contains the randomly permuted the values of variable X^j . $errOOB_{\ell}^j$ is the error made on OOB_{ℓ} ; observations that do not belong to the bootstrap sample $L_n^{\Theta_{\ell}}$ (69,136). Similar to the VI a Gini-importance (GI) was calculated and shows how often a particular protein was selected for a split, and how large its overall discriminative value was for classifying the outcome. However, for this analysis we used the MDA importance index.

Thus, the important variable is the one with a higher impact on the error after multiple permutations. Non-informative variables have a very low VI and in the presence of a group of highly correlated proteins limited masking effects appear, thus the proteins will be given the same rank.

5.2.3.3 Boosting Trees (GPBoost)

Boosting is also a tree-based approach but unlike Bagging in Random Forest it is a progressive process in which each next model which is generated is added to improve a bit from the previous model. In this analysis we used a technique called GPBoosting (139), which combines boosting with Gaussian process and mixed effects models. Using mixed effects is important for more efficient learning of the mean function and for obtaining non-biased probabilistic predictions (138). Another advantage of the boosting approach is its capability to handle outliers in and multicollinearity among predictor variables (186). A 20-fold cross-validation was used to tune the "*nrounds*"; number of trees in a single boosting iteration, "*learning_rate*"; shrinkage parameter, and the "*min_data_in_leaf*"; minimal number of samples per leaf. The grid that had the lowest RMSE on the different test sets was used to perform feature selection.

5.2.3.4 Bayesian Lasso

After using the above approaches to pick the top antigens, we then used Bayesian penalization with the aim to shrink features with small effects to zero while maintaining true large effects (121,123). On top of that, this penalization technique helps to remove redundancy and noise and then remain with an optimal antigen subset. Compared to classical penalization techniques, Bayesian penalization techniques performs better in a

data set with multicollinearity, and they also offer additional advantages such as readily available uncertainty estimates, automatic estimation of the penalty parameter (125).

A Bayesian approach with a horse-shoe prior which is a close family of the lasso penalty, derived from the same class of multivariate scale mixtures of normal was used. The horseshoe prior as described in detail in Chapter 2 is reported to perform better than the double-exponential prior at handling sparsity (130). Additionally, we had to use the multilevel strategy to handle the complex dependency structure in the of site-specific repeated measurements (187).

The important proteins was the best subset of explanatory variables selected from the features picked from RF,MP and GPBoost by estimating $\beta = \{\beta_1, \beta_2, \dots, \beta_p\}$ a set of coefficients that describe the relationship between the proteins/antigens $\mathbf{X} = \{X_1, X_2, \dots, X_p\}$ and malaria positivity which is the dependent variable y using a Bayesian Lasso model. In general, our model was defined as;

$$\begin{aligned} y_{ij} &= \beta \mathbf{X}'_{ij} + \alpha_j + \varepsilon_{ij}, & \varepsilon_{ij} &\sim N(0, \delta_e^2) \\ \alpha_j &= \alpha_0 + \alpha_j \end{aligned} \quad (5.3)$$

Where $i = 1, \dots, n_j$ observations, $j = 1, \dots, J$ the 14 sites in our dataset. $\mathbf{X}'_{ij}$ is a $(1 \times p)$ matrix of the regression coefficients $\beta = \{\beta_1, \beta_2, \dots, \beta_p\}$. α_j is the site-specific deviation from the overall intercept α_0 . A smoothing (normal exponential gamma) prior $p(\alpha_1, \dots, \alpha_j)$ was used to tie all the random intercepts together and promoting shrinkage towards the overall intercept (187,188).

$$\alpha_j | \psi_j \sim N(0, \psi_j), \quad \psi_j | \beta \sim p(\psi_j | \beta) \quad (5.4)$$

Then the likelihood was defined as,

$$p(y | \beta, \alpha_j) = \beta^y (1 - \beta)^{1-y} \quad (0 < \beta < 1) \quad (5.5)$$

The horseshoe prior on β was defined as;

$$\beta | \lambda_i, \tau \sim N(0, \sigma^2 \tau^2 \lambda_i^2), \quad \lambda_i \sim \text{Half-C}(0, 1), \tau \sim \text{Half-C}(0, \tau_0), i = 1, \dots, p \quad (5.6),$$

For $i = 1, \dots, n$ and $\text{Half-C}(0, 1)$ is the standard half-Cauchy distribution which puts probability mass at zero and has a fat tail. The global parameter $\tau \sim \text{Half-C}(0, \tau_0)$ shrinks all β towards zero and the local parameter $\lambda_i \sim \text{Half-C}(0, 1)$ allow the important β to escape the shrinkage (125,130).

Given the hyperparameters, the posterior mean satisfied approximately

$$\bar{\beta} = (1 - \kappa_j) \beta^{ML}, \quad \kappa_j = \frac{1}{1 + n \sigma^{-2} \tau^2 \lambda_j^2}, \quad (5.7)$$

Where κ_j is the shrinkage factor, the important proteins have $\beta^* \approx \beta^{ML}$ and less important features $\beta^- \approx 0$ (130).

In general, the posterior distribution was defined as;

$$p(\beta, \alpha_j | y) \propto p(y | \beta) p(\beta) p(\alpha_j) \quad (5.8)$$

where $p(y|\beta, \alpha_j)$ was the likelihood of $y \sim \text{Bernouli}(\theta)$, $p(\beta)$ and $p(\alpha_j)$ the priors and $p(\beta, \alpha_j | y)$ the joint posterior distribution (187).

The joint posterior distribution has a non-standard form and thus cannot be sampled directly, therefore we sampled from the posterior probability distribution using the Markov chain Monte Carlo (MCMC) estimation stochastic process implemented in brms package (189).

5.2.4 Network Models

Human beings' genes and proteins interact with each another through complex biological networks. Malaria is one of the diseases with a complex biological structure and transmission pattern (190). Visualizing how different proteins interact within different cohorts was very important for this multicentre study. We developed a directed network combining both the sites and top features selected in each site. The edge represented the relationship between the site and the feature, and the vertex was either the site or a feature. The magnitude of the feature vertex was determined by the number of times the feature was top in each site. Each vertex was uniquely identified and labelled as shown in the results section. The three essential components of our vertex were the site identifier (tail), the protein or the feature which was the (head) and the attributable information. Ours was a directed network, since the edge connects the elements of the site list with those of the feature list, but not vice versa (184,190).

To create the network, we had an adjacency matrix: a square matrix whose i, j th cell contains the information of the edge from i – *feature* to and an incidence matrix: a two-

dimensional array whose row elements represent vertices, and whose column elements represent edges (184).

5.2.5 Feature selection using probability outcome.

To further reduce the misclassification bias, we propose another approach of using the probability of having fever as the outcome. However, a training class is needed to define the probability thus we only used Junju, Kilifi data that had consistent training set available. This is done as a sub-analysis to show the utilization of the probabilities as defined in Chapter 3 to perform feature selection. The probabilities estimated from the Bayesian latent class model had a probability mass at zero since most of the individuals had no fever. We therefore applied zero-inflated beta Bayesian Lasso model to select the features and compared this with the zero-inflated binomial Bayesian Lasso model (191).

5.2.5.1 Zero-inflated beta Bayesian Lasso

The beta distribution is a flexible approach to model data measured in a continuous scale on the open interval (0,1) due to its flexibility of having different shapes. This distribution the best fit for our data set since the values ranges between zero and one with a high mass at zero due to the non-febrile participants.

In general, the beta distribution is defined as (166,192);

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad y \in (0,1) \quad (5.9)$$

Where $\Gamma(\cdot)$ is the gamma function. In practical datasets like our case, practical data include zeros the data is modelled as a mixture of two distributions mixture of two

distributions: a beta distribution and a degenerate distribution in a known value c , where $c = 0$ for this dataset. Then the distribution function is defined as (166,192)

$$bi_c(y; \alpha, \mu, \phi) = \begin{cases} \nu & \text{if } y=c \\ (1-\nu) f(y; \mu, \phi) & \text{if } y \in (0,1) \end{cases} \quad (5.10)$$

Where $f(y; \mu, \phi)$ is the beta density and $\nu = P(y=0)$ is the probability mass at zero.

There is the likelihood for $\theta = (\nu, \mu, \phi)$ given the outcome variable $(y_1, \dots, y_n)$ is defined as (166,192)

$$p(y | \theta) = bi_c(y_i; \nu, \mu, \phi) = p(\nu) p(\mu, \phi) \quad (5.11)$$

Letting $\beta = \{\beta_1, \beta_2, \dots, \beta_p\}$ be the parameters from the model with a horseshoe and random effects priors defined as shown in Equation 5.6 and Equation 5.7 respectively

In general, the posterior distribution was defined as;

$$p(\beta, \alpha_j | y) \propto p(y | \beta) p(\beta) p(\alpha_j) \quad (5.12)$$

where $p(y | \beta, \alpha_j)$ was the likelihood of $y \sim bi_c(y; \theta)$, $p(\beta)$ and $p(\alpha_j)$ the priors for the Beta and random effects respectively and $p(\beta, \alpha_j | y)$ the joint posterior distribution (166). The joint distribution $p(\beta, \alpha_j | y)$ is not available in a closed form and we apply Markov chain Monte Carlo (MCMC) method to draw posterior samples on the parameters leveraging on the available package brms build on JAGS (189,193).

5.3 Results

5.3.1 Binary Outcome Results

5.3.1.1 Modified Poisson Regression:

The estimated risk-ratios (RR) from the modified Poisson regression were ranked from the lowest to the highest as shown Figure 5-5.

The estimated RR ranged between ~ 0.7 and ~ 1.12 with the robust standard errors as shown in the Figure 5-6. The top 20 proteins picked from this approach were, PF113, MSP3, MSP2CH15019DBL2, MSP11, PTEX150, PF3D7_0525800(IMC1g), PF3D7_0629500_SEG2, MSP2CHO, MSRP4, MSP1Block2K1flank, MSP5, RAMA, MSPDBL2C-terminal, AMA1, MSP2, RH1, SERA7, PF3D7_1136200 and SERA3.

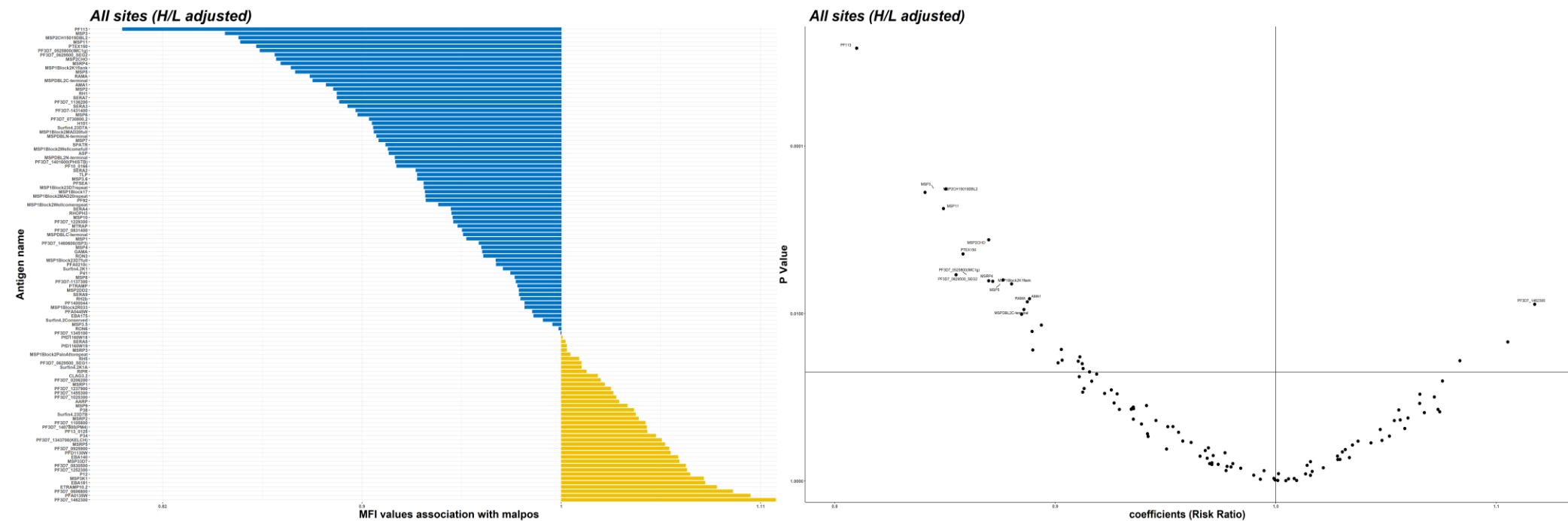


Figure 5-5: Modified Poisson regression RR estimates. The bar graph shows the estimated RR, and the scatter plot shows the distribution of the RR and the corresponding p-value. The p-values are not adjusted for multiple testing.

5.3.1.2 Random Forest

The Random Forest model was fit with an mtry of 37 and 10000 trees which were selected as the best parameters using a 10-fold cross validation. To handle the class imbalance which was observed to bias the OOB error in the malaria positive group, SMOTE technique was used and reduced the error significantly as shown in Figure 5-7

The top 20 antigens from the RF approach were MSP4, MSP2, PF113, PfD1160W18, MSP10, PfD1160W19, MSP1, MSRP1, MSP5, SERA2, SPATR, PF3D7_1401600_PHISTB_, MSRP5, PF3D7_0629500_SEG2, MSRP2, SERA5, PF10_0166, PF3D7_0730800.2 and PFA0445W.

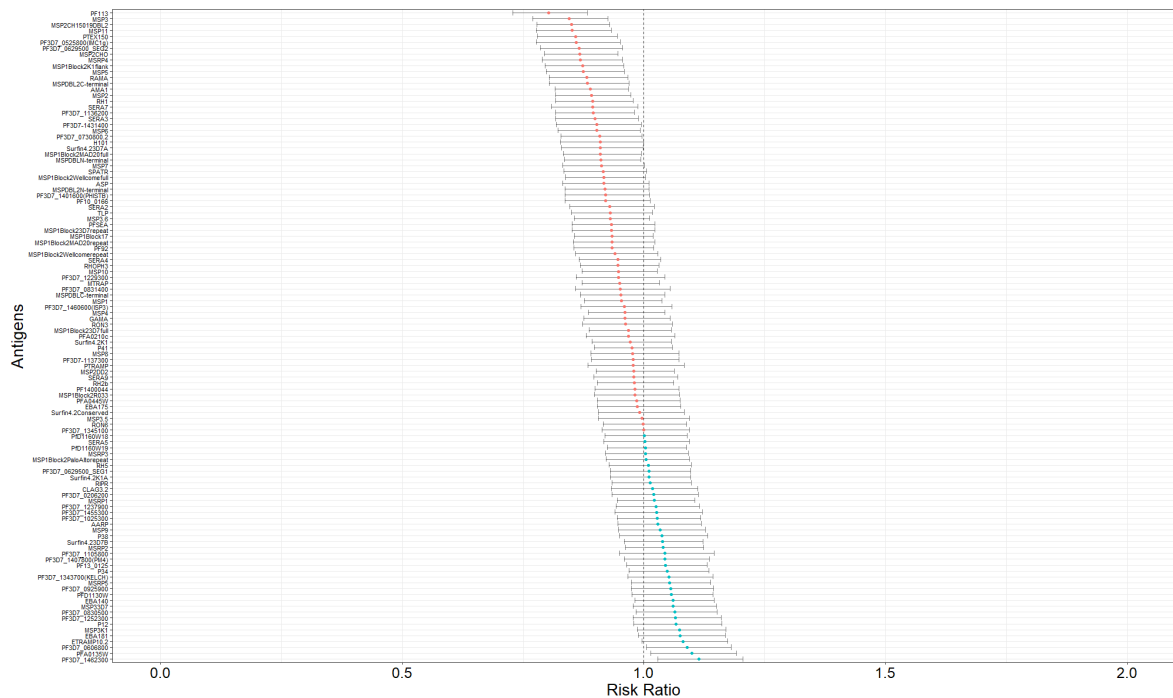


Figure 5-6: Estimated risk ratios and the Confidence intervals from the MP model

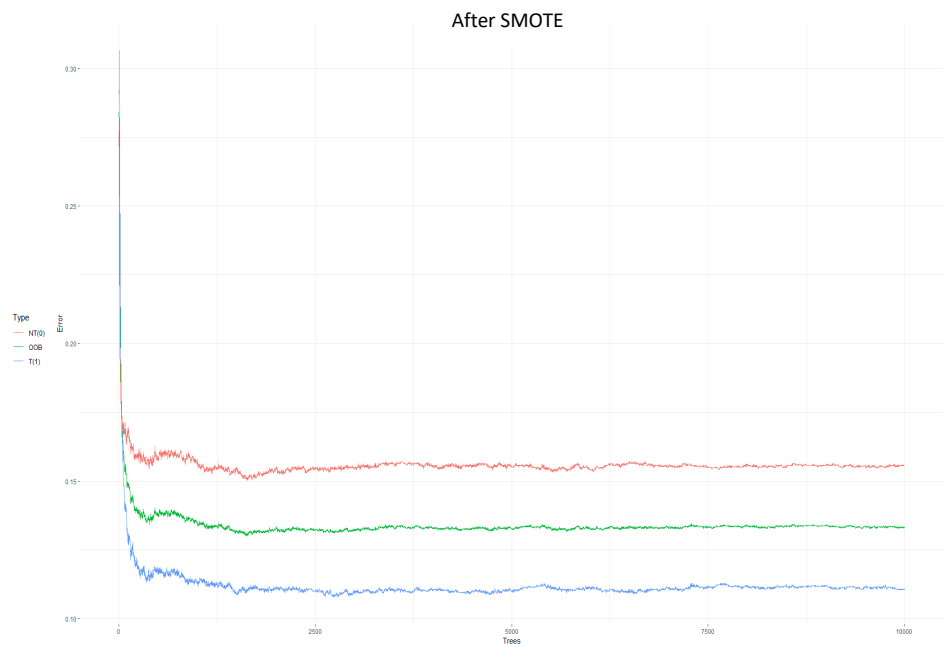
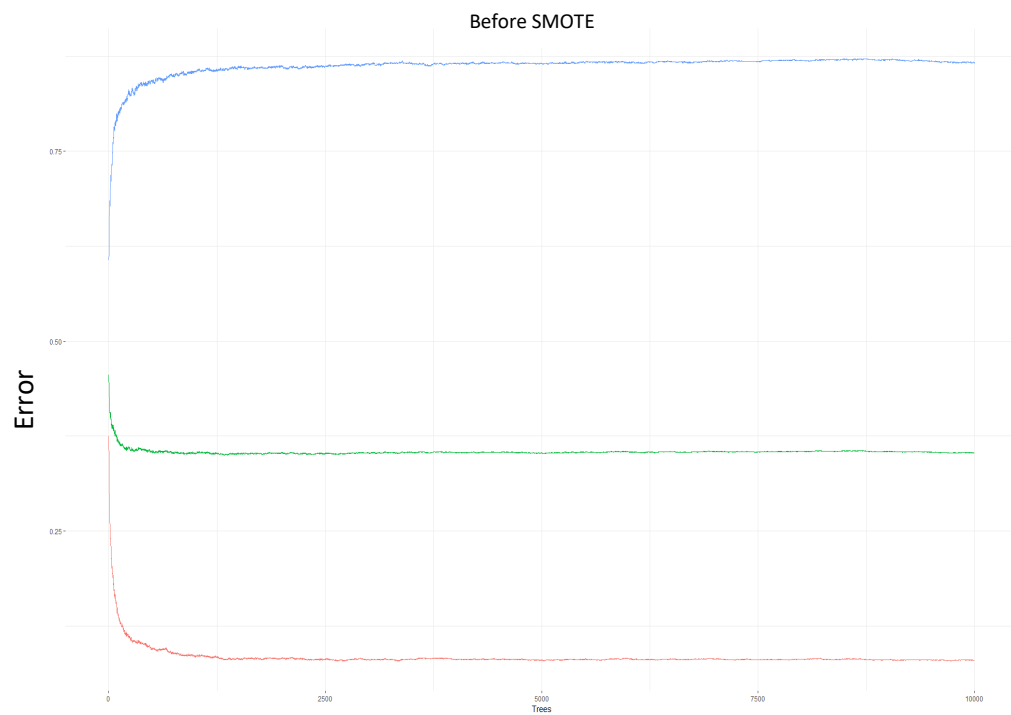


Figure 5-7: The OOB error before SMOTE and after SMOTE with “mtry” of 57 and 10000 trees

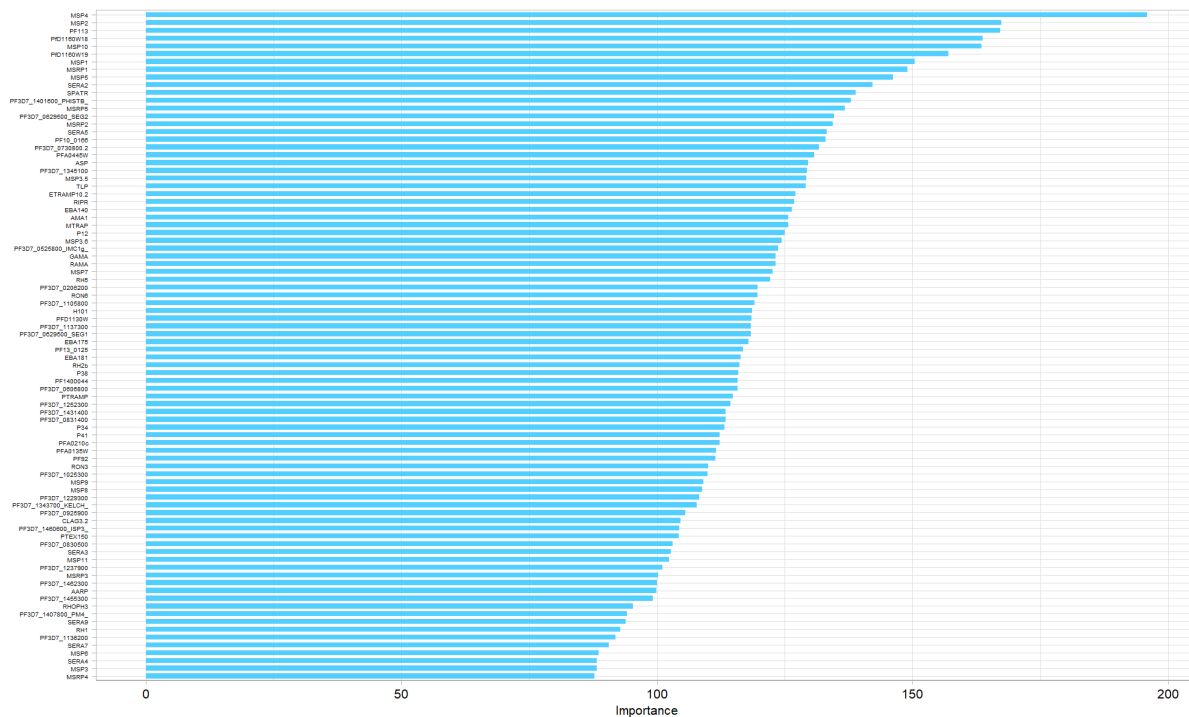


Figure 5-8: Rank importance using the Random Forest approach

5.3.1.3 Boosting Trees (GPBoost)

To fit the boosting trees model, we used the “grid_search_tune_parameters” and “gpb.cv” functions in Python and R respectively with 20-fold cross validation. The grids used to search for the best ‘learning_rate’ and ‘min_data_in_leaf’ were [0.5,0.1,0.05,0.01] and [5,10,20,50,100,200,300,500] respectively. From R we got a 27 nrounds, learning_rate of 0.5 and 200 for the min_data_in_leaf and Python gave 307 nrounds,a learning_rate of 0.05 and 100 for the min_data_in_leaf. We used both estimates to fit the model and they gave the same order of importance. The top 20 antigens picked were PF113, MSP3, MSRP1, EBA181, MSP10, PF3D7_1136200, SERA7, PF3D7_1401600(PHISTB), MSRP2, MSP9, MSP8, MSP3.6, MSP11, RAMA, MSP3.5,

PF3D7_0629500_SEG2, SERA3, MSP3.6, PF3D7_0629500_SEG1, RH1, MSP6, SERA7 and TLP. Notably **PF113**, **PF3D7_0525800(IMC1g)**, **AMA1**, **MSP11**, **PF3D7_1136200** **had** their 95% credible intervals (95 C.I) not crossing zero as shown in Figure 5-10. The model diagnostics shows the model was a good fit since the observed and predicted values were similar, Figure 5-10 C.

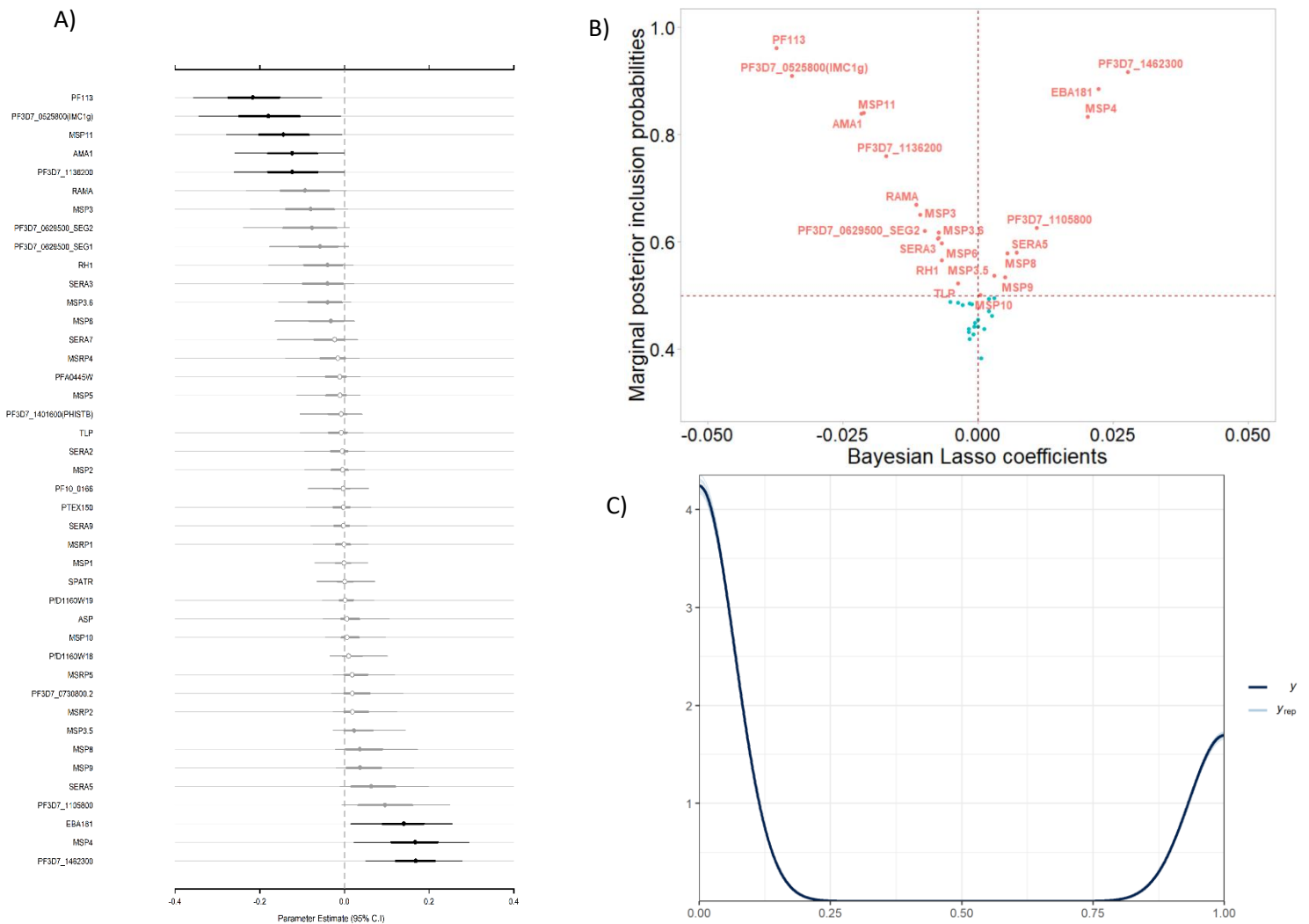


Figure 5-10: Bayesian Lasso estimates, A) the beta coefficients with the 95% credible intervals, B) the marginal posterior inclusion probabilities, the labelled ones (red in colour) had a probability >0.5, C) Model diagnostic plot for the observed and predicted values.

5.3.1.5 Network Model.

The results from the network model showed PF113, MSP3, MSP11, PF3D7_1401600(PHISTB) and MSP3.5 were identified as the most prominent antigens among the sites. The east African sites were observed to cluster together with PF113 exhibiting as the most interconnected feature. Other prominent interconnected features included PF3D7_1136200, MSP7, MSRP4, MSP3.6, MSP4 , AMA1 , MSP5, MSP1, RH1, PF3D7_0629500_SEG2, H101, EBA175, SERA3 and PF3D7_0525800(IMC1g). While less prominent connections emerged which SERA5, PfD1160W18, PF1400044, P34, PF3D7_1237900, PF3D7_1343700(KELCH) and PF3D7_1252300. The nodes without a connection were excluded in the figure below.

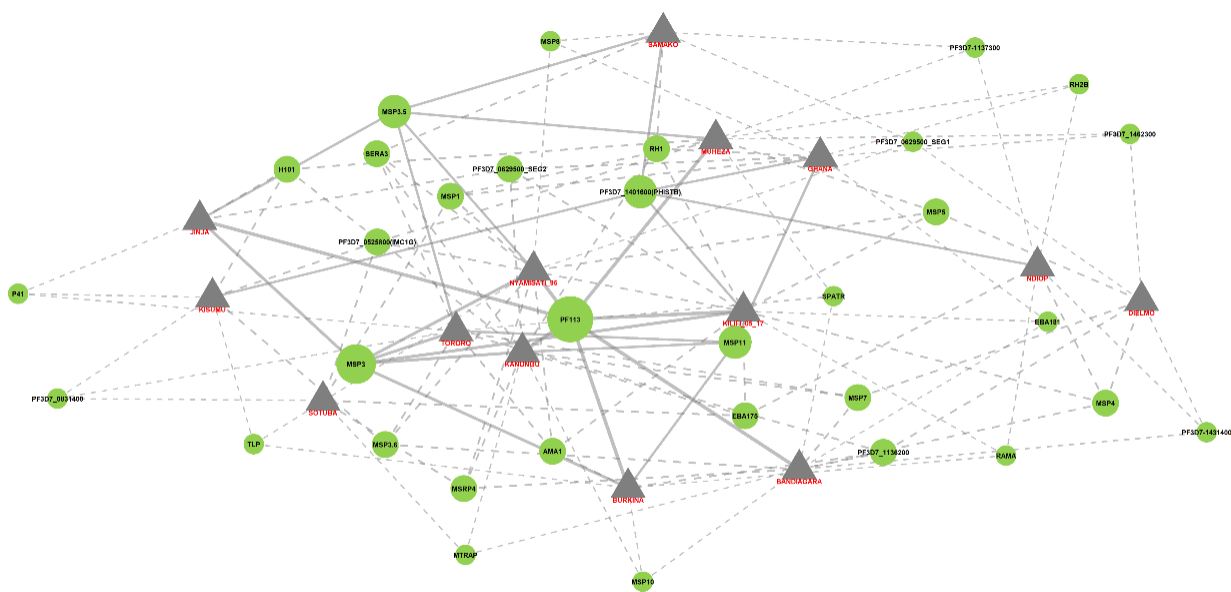


Figure 5-11: A network diagram showing the relationship of the top antigens per site. This analysis was done for each site and the network model created with the top features

5.3.2 Probability Outcome Results

We used 3,607 data from Junju, Kilifi collected between 2005 and 2017 as described in detail in Chapter 3, study design section. This was a subset of the multi-centre data since Kilifi samples had a training set available to use the Bayesian latent class models to estimate the probability of having clinical malaria. The probability values have a mass at zero as shown in the figure below. A zero-inflated Bayesian model with a horse-shoe prior on the β and global scaling parameter was fitted using 20000 iterations, a warmup of 2000, and a thinning of 1. The model took approximately 6.7hrs on a Core i7-8650U 1.90Ghz 4 cores 32GB RAM, computer. This was compared with a Bayesian binomial

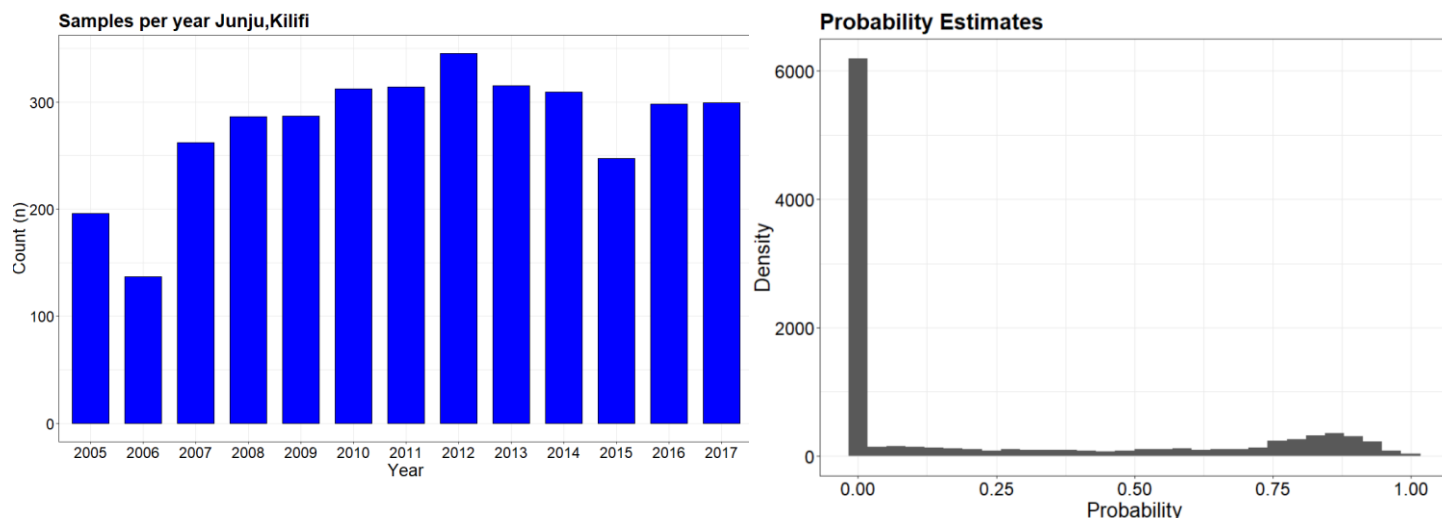


Figure 5-12: Sample size and probability distribution of the Junju cohort

model with a horse-shoe prior on the β with the same conditions which took approximately 2.5hrs.

The model diagnostics using trace and density plots to assess convergence and quality of inference was done using the posterior estimates included in the supplementary for this Chapter. All plots suggested convergence to a stationarity distribution according to the Gelman-Rubin convergence checks (194). Furthermore, posterior predictive checks on all models in these analyses showed the models fit the data well; there were no systematic departures from the model prediction as shown in the Supplementary Figure 5-1.

The zero-inflated model selected MSP6, MSP10, MSP11, SERA3, AMA1, MSP2, PF113, PF3D7_1136200 (95% credible interval not crossing zero) and MSP1, MSP3, MSP4, PF3D7_0629500_SEG2, PF3S7_0730800.2 were also among the top antigens (50% credible interval not crossing zero). The Binomial model selected PF3D7_0525800(IMC1G), MSP11, PF113, SERA2, MSP5, MSP6, RAMA, PFD1160W19,

RH1 and AMA1. There was an overlap of 8 antigens that were selected by the two approaches. The zero inflated model though had better WAIC of -1118.4 and the binomial model had a WAIC of 3965.5, thus the former model has a better statistical fit than the latter.

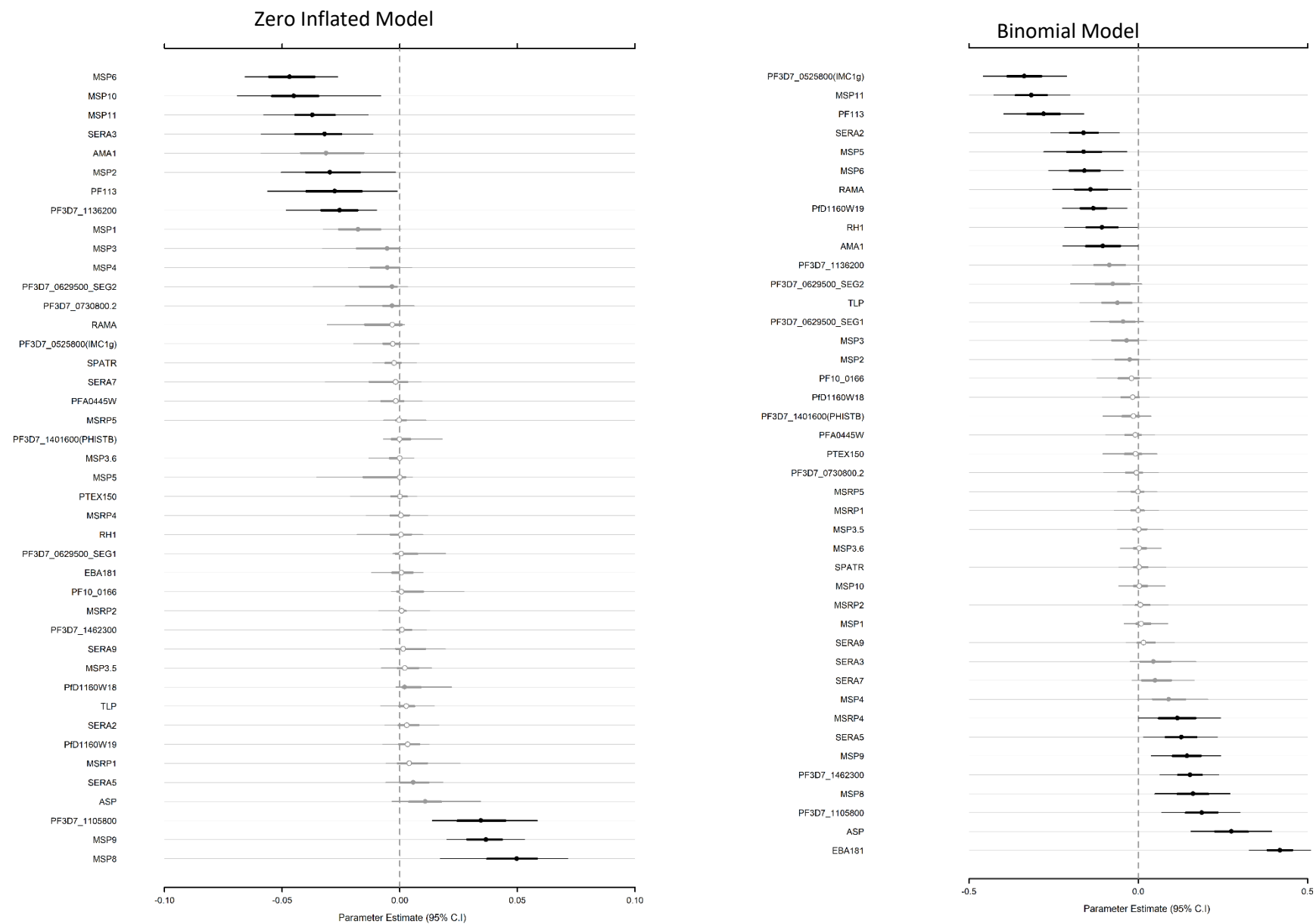


Figure 5-13:Top antigens picked by zero inflated and binomial mixed effect models for Junju, Kilifi dataset.

5.4 Discussion:

The selection of important features in regression problems and immunological studies has occupied the minds of many statisticians and immunologists respectively. Several machine learning feature selection methods have been developed. Specifically in malaria immune-epidemiology studies antigens expressed on or associated with the surface of *P. falciparum* merozoites are thought to be key targets of blood stage immunity against malaria. Utilizing robust and non-biased statistical approaches to comprehensively identify the best immune targets from the merozoite stages of *P. falciparum* is of importance to inform vaccine development. In this study we utilised techniques that have been used previously in malaria research combined with machine learning approaches that could capture the complexity of interactions between the malaria antibodies.

The data used here was from a micro-array chip designed with over 100 *P. Falciparum* proteins and the methods not only identified features reported in previous correlates analyses but also other novel candidates. Systematic and non-systematic sources of bias are common in high throughput protein microarray experiments. One of the biases is background noise. The localization and distribution of background signal intensities guide the selection of the appropriate correction strategy (87,104). In this experimental design with multiple mini-arrays and slides, we did not visualize any significant background pattern within the mini-arrays or block. For this reason, we selected the local background correction method. In protein array studies, local background subtraction debatably gives in principle an unbiased estimator of the true signal since it is arguably an impartial estimate of the local non-specific intensity(105). The protein micro array chip used here

had three technical replicates which is important to test within-sample variability. As we observed in Chapter 4, caution in interpreting the CV is recommended as small differences in signal intensities between replicates at the lower end of the detection threshold can yield misleadingly high CVs. Therefore, for this study we set our low detection at 1000 and for those with high CV's we selected the best 2 replicates.

Among the methods discussed in Chapter 4 for normalisation, VSN technique was the optimal approach to handle the mean variance dependence. The key design issues to consider for the best normalisation technique are the extent of biological sample variation anticipated and experimental design. Adopting the wrong method increases the probability of detecting false negative, and false positives. Batch effects create an additional unwelcome source of variation that could further reduce statistical power. These must be considered when data are processed at different times, by different users and on different instruments, among others (99). An important consideration that particularly applies to large cohort studies is the random processing of samples to ensure that batch effects do not exaggerate pre-existing genuine biological variation. The Combat approach was the best fit batch correction technique to use for our data before proceeding to feature selection.

Feature selection methods.

We combined the features selected from modified Poisson, Random Forest, Tree-Boosting and Bayesian with spike and slab modelling techniques and then applied the Horseshoe prior which is a cousin of the lasso penalty to select the important features from the set

of selected features. Since the data was based on a multicentre study, we developed a network model to visualize the interplay of the top features among the different cohorts. Modified Poisson combined with breadth scores have been used previously in single cohort studies to select features associated with protection (16,43). Modified Poisson models provide relative risk as the parameter of primary interest and assist in convergence problems that sometimes arise with binomial regression models (43). The modified Poisson is based on MLE approach and maximum likelihood approaches is slowed down by large covariate set. Also, a single complicated and complex model which might have a high variance leading to overfitting or underfitting (195). Therefore, we utilised univariate model on each antibody adjusting for the individual's overall response to reduce the complexity of the model. However, this approach does not adjust for the potential interplay between different covariates (47).

To handle this limitation of modified Poisson regression, we utilised a random forest approach to pick the top features. This approach has been utilised in another malaria study to identify distinct immune signatures that are predictive for individual level disease protection. They utilized data from three prospective cohort studies conducted in Kenya, Kenya/Tanzania and Mali but analysed them independently (47). The bootstrap aggregation (Bagging) technique is applied in random forest. This technique aggregates many trees handling the limitation of overfitting and error due to bias (136). Random forest rarely suffers from multicollinearity but have been reported to have a bias towards the correlated variables (196,197). The measure of the importance of the variable is measured with respect to the prediction of the outcome variable and reduced OOB error

(136). However, very few random forest methods have been extended to handle mixed effect approaches (198). One of the approaches developed specifically for Gaussian response variable is the Mixed-effects random forest for clustered data (MERF) and they are not appropriate to solve classification problems (199).

The context of a non-Gaussian response variable with mixed effects was present in our data with site as the random effect. Handling mixed effect structure present in a dataset improves model prediction performance and statistical inference underlying the complex structure. To model this, we utilised the Boosting approach which has been reported to handle complex non-linearities, discontinuities, and high-order interactions, it is robust to outliers and multicollinearity (138). An improved prediction accuracy was observed with this technique and an overlap of the top features with Modified Poisson and Random Forest approaches. This approach took longer time to process on the R software but a lot faster using the Python interpreter. This could have been associated with the different cross-validation parameters picked by both approaches, the latter had lower number of '*min_data_in_leaf*', '*max_depth*' and '*max_bin*'. Both the random forest and boosting techniques are more advantageous for predictive accuracy and the top features can either be associated with protection or exposure.

Using the binary outcome which was used for all multicenter data, the Modified Poisson, Random Forest and Tree boosting approaches together selected top 40 features. From this set of 40 features, the main interest was to identify the top candidates associated with protection against clinical malaria. A methodology that applies the lasso properties and handles multicollinearity was the best choice to pick the most important features.

This technique has been applied in HIV studies to gain knowledge on key features contributing to HIV profile differences and identifying the correlates of antibody mediated protection. In the HIV study an MLE lasso-based approach was used combined with s-PLSDA (68). Here we use the Bayesian inference technique with lasso properties. The lasso based on MLE has been shown to drop most of the correlated variables even if they are discriminants and randomly selects one representative among a set of correlated features(200). A horseshoe prior a cousin of the lasso penalty was used on the Bayesian β parameters for both the binary outcome and probability outcome. This prior has been shown to have a good computational performance compared to the other penalty-based priors(125). Specifically for the probability outcome a zero-inflated model was used due to the probability mass at zero for the non-febrile participants(191). Importantly, the Bayesian technique helps in identifying the features associated with the risk of having clinical disease and the ones associated with exposure. Five top features associated with reduced risk of clinical malaria were selected, they had a point estimate of less than zero and the credible interval not crossing zero.

Network models help to visualize complex associations and linkages among different features in research. Network analysis as a distinct scientific enterprise has been used to build a rapid linkage disequilibrium (LD) network connection of SNP pairs in genome prediction of traits with different genetic architecture and also visualize the complex network of close pedigree genetic diversity within the *domesticated grapes*(201,202),. Also, network models were used to visualize top biomarkers identified as correlates of Ab-mediated protection in a HIV study from multiple vaccine cohorts (68). In our study,

we applied the same technique to visualize the interplay of the top features selected in different cohorts and how they interlink. PF113 exhibited the most interconnected feature comprising of 7 sites followed by MSP3 and MSP11. The top candidates selected from the overall model were observed to have a large representation in the network model which encoded the complex relational structures of the top candidates within the different cohorts in Africa (184).

Misclassification bias

To help in further reducing the misclassification bias in malaria studies, we proposed the use of a probabilities of fever as an outcome. These estimated probabilities of fever have been used in determining risk of developing clinical episodes in malaria vaccine trials. In the vaccine studies, the probabilities estimated from the Bayesian latent class model were proposed as a better approach to compare the placebo and control groups (88,160). Comparatively, in our study, the model with the probability outcome had a better statistical fit compared to the binary approach. This shows that the utilization of probability explains more variation compared to the binary outcome. The binary approach was also observed to have wider credible intervals. The global scaling parameter on the horse-shoe priors was observed to affect the number of features selected and the prediction accuracy of the model. One of the major limitations of utilizing the probability outcome, is that a training class is needed to compute the probability values since as discussed in Chapter 3. A higher computational cost was also observed while utilizing the probability outcome due to the complexity of the posterior distribution and modelling the mixed effects for the zero-inflation part (166,192) .

The probability outcome approach had a better statistical fit due to its robustness compared to the binary outcome however this was only done for the Junju, Kilifi site due to lack of training samples in the other sites. It would be important to compare the performance of the probability outcome using multiple sites in feature selection.

5.4.1 Selected feature and biological importance:

Some of the identified antigens have been studied in other cohorts to understand their role in naturally acquired immunity. The *P. falciparum* protein 113 (Pf113) and apical membrane antigen 1 (AMA1) were shown to provide protection in South-eastern Gabon (203). The PF3D7_0525800 inner membrane complex protein 1g, putative (IMC1g) is termed to be essential and having a moderate mutant fitness score in the PlasmoDB (204) and shown evidence of immunogenicity in adults by a study done by Kamuyu et al in Nyamisati, Tanzania (205). MSP11 also known as the surface localisation of PF3D7_1036000 and PF3D7_1136200 have been proposed to be good candidates for a multicomponent blood-stage malaria vaccine (41). A few of the identified features are being tested in vaccine trials and others have shown immunogenicity in other cohorts. With this multicentre study, the identification of these proteins validated the overall approach, and added confidence that the untested features can be evaluated further using functional and diversity studies to understand their polymorphism and use animal models to broadly understand immunogenicity (205).

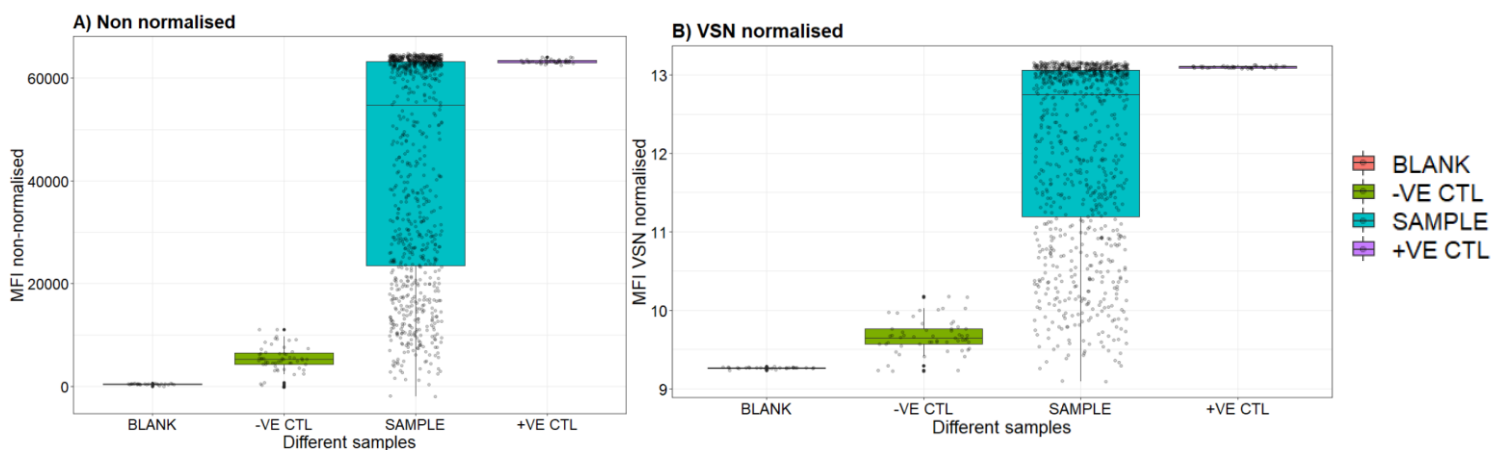
5.4.2 Conclusion:

The probability outcome-based model had a better statistical fit compared to the binary model helping in reducing the misclassification bias. Nevertheless, in our study, there was a limitation of lack of training samples for most of the multicentre cohorts thus we did not test the probability model in the overall data. The network model helped to visualize complex associations and linkages among different sites and the identified features were similar to what was observed in the overall model. A multicentre network analysis incorporating other immunological parameters of malaria protection, such as, functional assays and soluble parameters would be of importance to further understand the interplay of malaria protection studies from different sites. From the overall analysis, PF113, PF3D7_0525800(IMC1g), AMA1, MSP11, PF3D7_1136200 antigens were selected as the top features associated with protection. However, these and other antigens with a protective effect would be recommended for further functional assays to inform vaccine development. The probability outcome-based model had a better statistical fit compared to the binary model helping in reducing the misclassification bias. Nevertheless, in our study, there was a limitation of lack of training samples for most of the multicentre cohorts thus we did this for Junju, Kilifi cohort only.

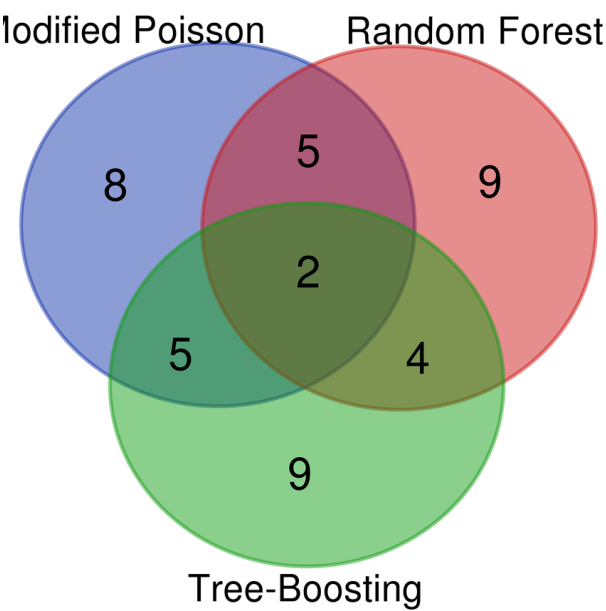
5.5 Supplementary



Supplementary Figure 5-7: TAG antigens that had high MFI values. GST and MBP had very high MFI values compared to CD4TAG, thus the antigens tagged with MBP, and GST were excluded

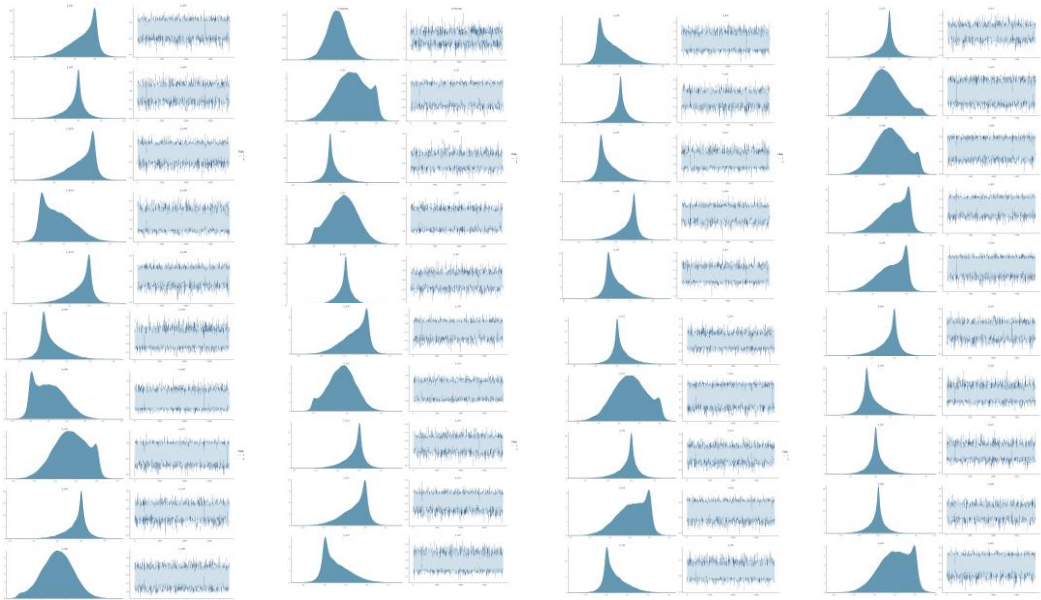


Supplementary Figure 5-8: Performance of VSN normalisation on blank which is the sample buffer labelled as BLANK , negative (malaria naïve sera) controls labelled as -VE CTL, serum samples labelled as SAMPLE and positive (malaria immune sera) controls labelled as +VE CTL.

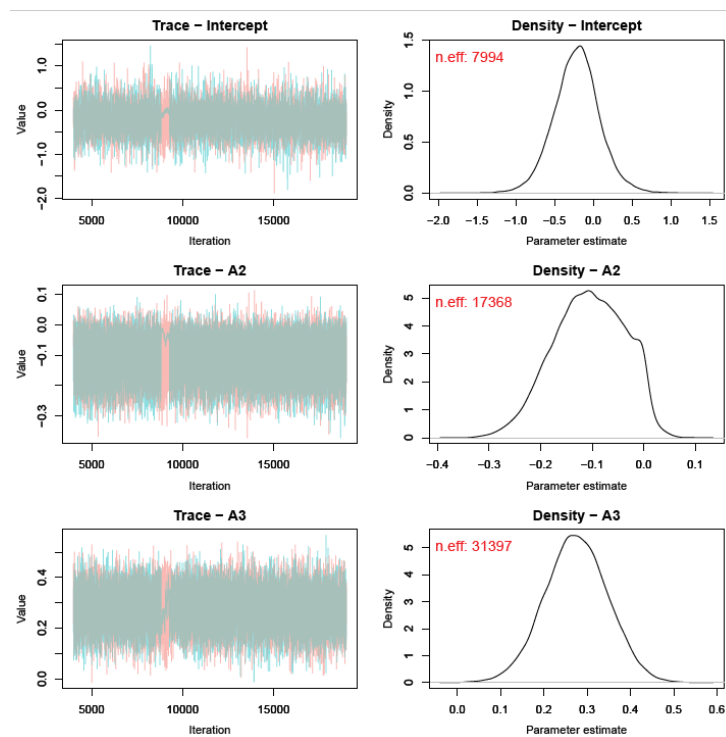
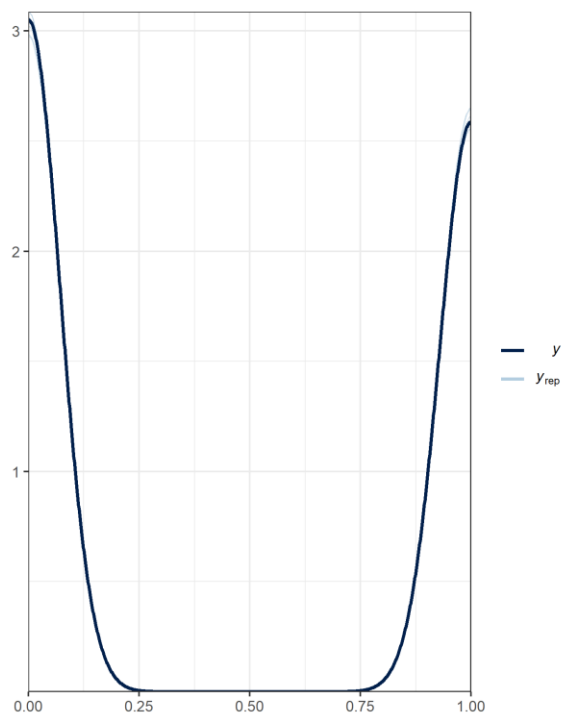


Models	Total Shared Features
Modified Poisson Random Forest Tree-Boosting	2 PF3D7_1401600(PHISTB) PF113
Modified Poisson Random Forest	5 PF3D7_0629500_SEG2 MSP2 PF10_0166 SERA2 MSP5
Modified Poisson Tree-Boosting	5 RAMA PF3D7_1136200 SERA7 MSP3 MSP11
Random Forest Tree-Boosting	4 MSP4 MSRP2 MSRP1 MSP10
Modified Poisson	RH1 PTEX150 8PF3D7_0525800(IMC1g) TLP AMA1 MSRP4 SERA3 MSP6
Random Forest	PfD1160W19 SERA5 SPATR ASP 9MSRP5 PFA0445W PfD1160W18 MSP1 PF3D7_0730800.2
Tree-Boosting	9 PF3D7_0629500_SEG1 MSP9 MSP3.5 PF3D7_1462300 MSP8 MSP3.6 EBA181 PF3D7_1105800 SERA9

Supplementary Figure 5-9: A total 42 antibodies as shown used in Bayesian Lasso that were picked by MP, RF and Boosting approaches.



Supplementary Figure 5-10: Bayesian Lasso model diagnostics for the overall data analysis. The figure shows the 2 chains mixed well



Supplementary Figure 5-11: Model Diagnostic zero-inflated and Binomial model

CHAPTER 6 : DISCUSSION

The process of identifying important biological features is challenging especially for outcomes prone to differential misclassification and significant biological complexity. Consequently, researchers across the globe, are facing challenges in making data-driven decisions and recommendations. This is particularly relevant for studies on natural immunity to malaria. A strong rationale for developing and improving malaria vaccines comes from epidemiological studies in malaria-endemic countries. These studies show that individuals continuously exposed to malaria gradually develop “natural immunity” that initially prevents death from severe malaria, and subsequently recurrent clinical episodes of mild malaria. Therefore, identifying the targets underlying naturally acquired immunity is of importance to further inform continuing vaccine development with a view to achieving the vision of a world free of malaria.

This thesis explored the utilization of various statistical techniques to identify correlates of malaria protection from multiple cohort’s data. We proposed a workflow of ensuring the disease and the correlates have been measured without bias and applied machine learning approaches to select the important candidates. Chapter 3 focused on case definition approaches to reduce misclassification bias in malaria studies. Chapter 4 focused on developing a statistical tool to handle systematic and non-systematic bias that comes with protein micro-array data. Chapter 5 focused on identifying top correlates of disease protection using classical and machine learning approaches. This chapter joins together the findings from prior chapters, as the investigation work is concluded.

6.1 Identifying correlates of disease protection in malaria

Cohort studies inform the development of malaria vaccines through investigating factors associated with natural immunity (16,41,76). In general, following recruitment, immunological responses are analysed in relation to the participants' subsequent clinical experience of malaria. Many of the cohorts established in the past 3 decades focused on either a single, or small number of proteins would be sufficient for protective immunity (9,16,41,76,83,85,110). The methodological differences in those studies make results for a given antigen difficult to compare. To the best of our knowledge this is the largest malaria multicentre study to apply similar methodological approaches to identify important correlates of clinical malaria protection and propose the utilization of probabilities as opposed to binary outcome in feature selection where a training data set is available. We propose a work flow of ensuring disease is defined without bias, handle systematic and non-systematic laboratory data bias and perform feature selection.

6.2 Case definition

In areas with high malaria transmission, asymptomatic carriage of parasites is highly prevalent, and it is therefore likely that an individual with a non-malarial fever has coincidental parasitemia. This immunotolerance can be because of naturally acquired protection against malaria (2). Thus, parasitemia accompanied by fever may not be adequate to define an episode of clinical malaria and may lead to over-estimation of the malaria burden leading to differential misclassification (18,21). The logistic power model is the widely used technique for case definition (19,24,25) and rarely the Bayesian latent class model (17,91) to handle this differential misclassification.

In this study we compared the performance of the Bayesian and Logistic case definition approaches and investigated the utilisation of probabilities over binary outcome variables in feature selection. Cases corresponded to fever episodes detected by active case detection and control slides were those collected during the cross-sectional bleed. The choice of cases and controls may estimate varying lambda estimates. For example, if cases are recruited in health facilities (incident cases) where the episodes might be more severe compared to household active follow-ups (prevalent cases) may give different attributable fractions since the latter cases are more severe.

For the logistic power model, we tested the power function at different values to select the parameter that best maximised the log-likelihood. To reduce the impact of serial autocorrelation among parasite densities observations of the same individual, if there were consecutive parasite densities within 2 weeks, only the first measurement was considered. However, using the repeated values has negligible impact on the lambda estimates as reported by Smith et al (21).

The Bayesian latent class model performed well in the estimation of the probabilities compared to the logistic power model and this was suspected due to the properties of the logit (sigmoid function) used in logistic models. Therefore, we used the probabilities estimated through the Bayesian latent class approach to compare with the binary outcome. The binary outcome was created after selecting the best parasite density threshold cut-off using the logistic power model. This study shows that the utilization of probabilities in modelling association with clinical disease protection had a better statistical fit compared to the binary base models. Other studies have shown, continuous

variables, like the probabilities we use here, to have more variation information and statistical power and are sometimes preferred over the categorization of data (93). Several articles (92–94) have also pointed out problems associated with the categorization of data. However, a training class is needed; a sample that comes from the component of the mixture corresponding to children without fever but who may have parasites to use the Bayesian latent class models (17).

6.3 Protein microarray data processing

Protein microarray technology is increasingly utilised for vaccine candidate discovery in the 'omic era' with hundreds to thousands of antigen-specific antibodies analysed simultaneously (37,54,172). In malaria this technology has been used to study association with sporozoite-induced protective immunity (206), chloroquine chemoprophylaxis (CPS) induced immunity in Controlled Human Malaria Infection Studies (207), gametocyte proteins association with reduced malaria transmission (208) among others. However, the studies used different data management strategies to handle the systematic and non-systematic biases associated with this protein micro-array were used. Additionally, tools to guide the rigorous processing of protein microarray data are limited, require investigation by trial and error, are time-consuming to optimize and not centralized (62). We developed protGear (87) with various statistical and computational functionalities to address background bias (105) within-sample variation (38), mean-variance dependence (100) and batch differences (98).

6.3.1 Background Correction

The localization and distribution of background signal intensities guide the selection of the appropriate correction strategy (104–106). Although local background subtraction from the foreground resulting is considered to give unbiased estimates (105), negative values are often generated, and the consistency of the results across different mini-arrays and slides needs to be assured. Additional methods including the novel half moving minimum we developed that overcome these limitations have been included (57) in the tool. For the protein microarray data used in feature selection, local background subtraction was used since we did observe any significant background pattern within the mini-arrays or blocks.

6.3.2 Within sample variation

The inclusion of two or more technical replicates is vital for the robust analysis of within-sample variability (55). The variability is estimated coefficient of variation (CV) estimation. The lower the CV, the lower signal variation and, typically, the better the quality of the replicates(62). Although advisable, a higher CV value could not always translate to bad quality replicates especially for small MFI values. Replicates with low values tend to give very high CV values even with a small difference. As a result, we implemented a function within the package to allow users calculate the CV's, define the lowest MFI cut-off to flag the replicates and select the best replicates. Since the range of the MFI values for the samples used here ranged between 0-65000, we used a 1000 MFI cut-off to flag the CV's and the replicates with a "bad" CV were repeated.

6.3.3 Normalisation

The choice of normalization technique to handle mean-variance dependence is essential, and mainly depends on the design of the experiment and the assumptions of the normalization technique (180). As for example, some normalization approaches, such as LOESS and Quantile assume that majority of the probe intensities are not differentially expressed (37,38,106,113). As observed, this cannot readily be made for focused antibody microarrays, such as ours. Consequently, if either of these two approaches is selected as the best approach, the assumptions are violated and this can increase the probability of including false negatives, and potentially false-positive signals. The RLM normalisation approach that was specifically developed for protein microarrays (38) requires that the secondary control normalisation proteins yield constant signals across and within slides. These control proteins are expected to have a low CV and can be used for both normalizations and evaluating the slide variability (38). Although the MVD was significantly reduced with all methods tested, the VSN method was optimal for this analysis. Our tool, protGear provides a novel powerful single platform that empowers users to sequentially interrogate all these options to determine the optimal solution for the data in question.

6.3.4 Batch Correction

Batch effects create an additional unwelcome source of variation that could further reduce statistical power. These must be considered when data are processed at different times, by different users and on different instruments, among others (99). An important consideration that particularly applies to large cohort studies like ours where the random

processing of samples is to be considered to ensure that batch effects do not exaggerate pre-existing genuine biological variation. A limited number of batch correction approaches have been proposed and these typically accommodate experimental designs with a minimum of 25 batches. We adopted the ComBat batch correction from “sva” package in R has been adopted since it has been reported to be robust to outliers in small batch sizes (98).

6.3.5 Protein micro-array data processing suites

To the best of our knowledge, protGear is the only protein micro-array processing tool and interactive dashboard that provides a choice of functions to address each major source of bias, outlines the theoretical background necessary to guide selection and generates custom graphical and tabular outputs to support data interrogation and interpretation. protGear provides a, state-of-the-art, one-stop adaptable workflow for protein microarray data pre-processing. It can be coupled to software such as Sweave or knitr (181) for report generation and sharing along with the raw data to promote data reproducibility. The user-friendly, interactive, web-based, and graphical interface requires limited R-experience and will enhance broad uptake in the infectious disease and vaccine development community.

However, protGear does not automatically select the best strategy for data processing and the user is guided by the structure of the data. Additionally, the tool cannot process protein micro array data with more than 3 laboratory replicates. This will be considered in the next development of the package.

6.4 Feature selection of Merozoite targets of protective immunity

Machine learning approaches can complement classical analytical methods for identifying correlates of protective immunity. While Modified Poisson regression involves univariate evaluation of the different features measure, machine learning approaches can in addition identify interplay and relationships between antibody features that are predictive of protection. It is worth indicating that all examined methods employ the same feature set on the training dataset with k-fold cross-validation.

6.4.1 Modified Poisson

The main advantage of using the classical regression especially modified Poisson is that it gives you coefficients that shows the direction in which a certain feature associated with. However, the modified Poisson is based on MLE approach and maximum likelihood approaches is slowed down by large covariate set. Also, fitting a single complicated and complex model with a big dataset which might have a high variance leading to overfitting or underfitting (120,123) and having very small significant effect sizes. In our study we observed the estimated risk ratios were close to 1 (Modified Poisson Results). Using L1 regularized logistic regression has been shown to be effective even with many irrelevant features however it affected by multicollinearity (123). The modified Poisson models a binary outcome adjusting for known confounders followed by ranking the important features using the risk ratio estimated (43), therefore we did not use this approach for the probability based feature selection.

6.4.2 Random Forest

Random forest is an ensemble supervised learner based on many randomly fit decision trees and provides a feature important measure (69). As a classifier, random forest performs an absolute feature selection, by considering all possible combinations of measured antibodies then extracting the most discriminating features for the classification problem (136). Random forest compared to modified Poisson showed to be robust against single noisy features with a large amount of non-discriminatory variance. However, class imbalance was observed to affect the OOB error, and this was handled using the SMOTE technique (73). The random forest in our data did not perform well with categorised predictors therefore we use the continuous normalised MFI values as the covariates. There was an overlap of 7 features selected between the Random Forest and the modified Poisson approaches. To note for the random forest, it picks the top antigens that discriminate between the susceptible and cases, therefore, it has some features associated with exposure. Very few random forest-based methods have been extended to handle mixed effects and use zero-inflated outcome to solve classification problems (198)

6.4.3 Boosting trees

Tree boosting is also a supervised machine learning technique that improves the prediction performance and reduces the risk of overfitting and can be used for both classification and regression problems. In this classification study, the *gpbboost* technique was employed to combine the advantages of tree-boosting with mixed effects structure. Some of the advantages, are handling complex non-linearities, discontinuities, and high-

order interactions, also, it is robust to outliers and multicollinearity (Sigrist 2020). In our analysis we used Python and R to implement the models, but it was computationally effective in the former platform even with a wider cross-validation grid search. An improved prediction accuracy was observed with a wider cross-validation and an overlap of the top features with Modified Poisson and Random Forest approaches. Both the random forest and boosting techniques are more advantageous for predictive accuracy and the top features can either be associated with protection or exposure. Most of the zero-inflated distributions available for tree boosting are for count data but none was identified to handle the zero-inflated beta for our probability outcome (209).

6.4.4 Bayesian Lasso

The above techniques produced a set features we applied the Bayesian lasso technique to help remove redundancy and noise and then remain with an optimal antigen subset (121,123). On top of that, this Bayesian inference-based technique assisted in comparing the binary and probability outcomes performance in feature selection (166). Compared to the MLE penalization techniques though they have been reported to be computationally faster they pose the risks of non-convergence, large bias, sensitivity with initial values for both correlated and zero-inflated data (125,189). Also, they perform better in a data set with multicollinearity, and offer additional advantages such as readily available uncertainty estimates, automatic estimation of the penalty parameter (131,187). We used the horse-shoe prior as our penalization approach over the other Bayesian model space priors due to its reduced computational cost and ability to scale very high dimensional problems (125,133).

With the rising popularity of beta regression approaches in the last decade, significant advancement has been made in the methodological, computational, and theoretical development (210). This growth allowed us to compare the performance of probability of having fever and binary outcomes in feature selection problems. Probabilities of having fever have been utilized in vaccine trials to determine the risk of developing clinical episodes (88,160). Though the utilization of zero-inflated beta regression warranting regularization of the likelihood had a higher computational cost, it had a better statistical fit while benefiting on the data power, predictive performance and reduced risk of Type I error that comes with continuous data.

From the set of around 40 features in the overall data, the Bayesian lasso identified 4 important features with a credible interval not crossing zero and 15 features with an estimated posterior probability of inclusion of above 0.5. In the sub-analysis of comparing the probability and the binary outcome, the latter picked fewer features compared to the probability-based model and had a better statistical fit.

6.4.5 Network Models

We aimed to gain more understanding relationships between features contributing to profile differences and the different sites. This was done by adapting correlation network analysis tools frequently used in the transcriptomics field. The resulting network model depicted the features occurring in multiple sites as top candidates (Figure 5-11 big circles) was like the top candidates selected by the overall multisite model contributing protection against malaria. The top candidates selected from the overall model were observed to have a large representation in the network model which encoded the complex relational

structures of the top candidates within the different cohorts in Africa (184). Overall, these network model point to unique relationships between all features and sites among the different proteins tested. This may help in the selection site specific antibodies associated with protective immunity for further biological studies that may help identify mechanisms underlying correlates.

6.4.6 Misclassification bias and feature selection

In this study we utilised logistic power and Bayesian latent class modelling techniques to handle differential classification in malaria case definition. Both models showed similar patterns in estimation of attributable fractions though the former gave lower values than the latter. The Bayesian latent class and logistic regression estimate the probabilities of having fever however the logistic power model had high probability estimates of over 0.5 for most of the parasite densities. This was suspected to happen due to the logit (sigmoid function) function used to calculate the probability of the occurrence of a specific categorical event. Therefore, we used the probability estimates from the Bayesian latent class model for our beta regression-based models.

The top features picked by the binary and the probability-based models were similar, however the credible intervals from the binary model were wider. Therefore, the probability-based approach picked more “significant” features compared to the binary model. This approach also had a better statistical fit, it had lower BIC when it was used to model association with protection in Chapter 3 and a lower WAIC in the feature selection approach in Chapter 5. The standard errors from the probability models were lower compared to the binary regression approach.

6.4.7 Selected feature and biological importance

The machine learning approaches compliment the classical regression approaches aimed at selecting important correlates of protection. Thus, these machine learning tools provide an additional powerful method for comparing immune profiles of susceptible and sick individuals to provide greater systematic insights underlying the relationships of features that may contribute to naturally acquired immunity. As we have shown, antibodies are a major component of a host's natural acquired immunity against malaria. From the set of 86 humoral features studied, we identified the important correlates of protection, and a few are being tested in vaccine trials and others have shown immunogenicity in other cohorts. With this multicentre study, the identification of these proteins validated the overall approach, and added confidence that the untested features can be evaluated to understand their polymorphisms and maybe use animal models to evaluate immunogenicity and protective efficacy as potential vaccine candidates (205). We also show that diverse set of antigens would be a better consideration for vaccine development since more than one antibody was identified as important.

6.4.8 Strengths and limitations

The strengths and limitations of this study are discussed in epidemiological and methodological aspects. Show casing the different methodologies to handle bias associated with malaria case definition and protein micro array are the main methodological strengths of the study. In case definition, utilization of probabilities over categorical variables provided a better statistical fit. Also, we develop an interactive open-source suite to handle protein micro array data processing.

This study utilizes rigorous immuno-epidemiological analyses to comprehensively identify and prioritize the best immune targets from the merozoite stages of *P. falciparum*. We analyse antibody responses to at least 80 unique proteins to identify those that are consistently associated with a lower risk of malaria across multiple endemic sites in Kenya, Uganda, Tanzania, Ghana, Mali, Burkina Faso and Senegal. We identify potential candidates for consideration in further lab studies to inform the development of the second-generation malaria vaccines.

The study had a few limitations, largely due to the nature of the data that was received from the different sites and also due to the constraints of getting more lab data due to the pandemic constraints. A major limitation was lack of a training sample data from most of the cohorts' studies to implement the Bayesian latent class models. Therefore, we could not compare the performance of using probability of fever as the outcome in the overall multicentre data and did this only for the Junju cohort, Kilifi.

The other limitation was a longer processing time of the Bayesian lasso posteriors on a personal computer, and we would recommend running the models in a High-Performance Computer with Clusters to reduce the processing time. Specifically for the probability models, calculating the marginal probability of having clinical malaria during a follow-up was a challenge and this would be important especially for longitudinal studies investigating immunogenicity over many years.

The protGear package implements a Shiny dashboard but we could not get an open access platform to host the dashboard and users must run this from their personal computer.

6.4.9 Recommendations.

Our recommendations are guided by the different procedures and techniques we followed while answering our 3 research questions.

Regarding case definition, we recommend.

- 1) Utilization of active follow-up data for case definition to avoid under estimation or over estimation of the attributable fractions.
- 2) Use Logistic power model to estimate the threshold and Bayesian latent class for probability estimation

While processing the protein microarray data,

- 1) We recommend creating an electronic database of the plate maps for big studies to fasten the process of merging epidemiological and laboratory data.
- 2) We recommend the randomization of samples during the design of the study to ensure there is minimal variation between the arrays.
- 3) We recommend inspection of the scanned images before data extraction for any obvious spatial artefacts.
- 4) Compare the data from various PMT/Power settings to select the optimum settings for data that will be analysed together.
- 5) We recommend utilization of the median MFI values as it is relatively insensitive to outliers, commonly observed in protein micro array data.
- 6) Caution in interpreting the CV is recommended as small differences in signal intensities between replicates at the lower end of the detection threshold can yield misleadingly high CVs.

- 7) Normalisation is guided by the purpose of the data. For monoclonal based studies avoid using VSN and Loess based techniques.

Finally, from the feature selection objective,

- 1) Check for correlation in the extracted data to ensure you utilize models that can handle this. In our case, we utilized Bayesian lasso technique which has this capability.
- 2) For large datasets, we recommend running the Bayesian models with spike and slab prior in a High-Performance Computer with Clusters to reduce the processing time.

Unanswered question.

The data for other immunological parameters such as functional assays and soluble parameters was not generated for the samples due to the pandemic constraints. This limited my ability to include these parameters in the network model to test the interplay across various malaria immunological parameters and sites.

6.4.10 Summary and Conclusions

The general objective of this study was to develop analytical approaches that are sensitive to the challenge that when the outcome or predictor is prone to differential misclassification and to identify the correlates of protection against disease protection using machine learning approaches. To handle differential misclassification in malaria studies we propose the utilization of propensity scores as the outcome where a training sample is available rather than dichotomizing the data. We developed an R package to

handle the systematic and non-systematic biases associated with protein micro-array data. Specifically, we developed various approaches to handle background noise depending on the distribution of background values and normalisation depending on the pattern of mean variance dependence. Finally, we utilised machine learning approaches to identify important correlates of protection and observed that utilizing the probability outcome had a better statistical fit compared to using a dichotomous outcome, however, it was computationally and time intensive.

References

1. Bejon P, Warimwe G, Mackintosh CL, Mackinnon MJ, Kinyanjui SM, Musyoki JN, et al. Analysis of immunity to febrile malaria in children that distinguishes immunity from lack of exposure. *Infect Immun*. 2009;77(5):1917–23.
2. Marsh K, Kinyanjui S. Immune effector mechanisms in malaria. *Parasite Immunol*. 2006;28(1–2):51–60.
3. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. 1999;86(4):843–55.
4. Speybroeck N, Praet N, Claes F, Van Hong N, Torres K, Mao S, et al. True versus apparent malaria infection prevalence: the contribution of a Bayesian approach. *PLoS One*. 2011;6(2):e16705.
5. WHO WHO, others. World malaria report 2020: 20 years of global progress and challenges [Internet]. World Health Organization; 2020 [cited 2021 Nov 23]. Available from: <https://www.who.int/publications/i/item/9789240015791>
6. Lover AA, Harvard KE, Lindawson AE, Smith Gueye C, Shretta R, Gosling R, et al. Regional initiatives for malaria elimination: building and maintaining partnerships. *PLoS Med*. 2017;14(10):e1002401.
7. Moonasar D, Maharaj R, Kunene S, Candrinho B, Saute F, Ntshalintshali N, et al. Towards malaria elimination in the MOSASWA (Mozambique, South Africa and Swaziland) region. *Malar J*. 2016;15(1):1–5.
8. Secretariat EAC. Great Lakes Cross Border Malaria Initiative Meeting Report. 2019;
9. Jagannathan P, Muhindo MK, Kakuru A, Arinaitwe E, Greenhouse B, Tappero J, et al. Increasing incidence of malaria in children despite insecticide-treated bed nets and prompt anti-malarial therapy in Tororo, Uganda. *Malar J*. 2012;11(1):435.

10. Noor AM, Amin AA, Akhwale WS, Snow RW. Increasing coverage and decreasing inequity in insecticide-treated bed net use among rural Kenyan children. *PLoS Med.* 2007;4(8):e255.
11. Knox TB, Juma EO, Ochomo EO, Jamet HP, Ndungo L, Chege P, et al. An online tool for mapping insecticide resistance in major *Anopheles* vectors of human malaria parasites and review of resistance status for the Afrotropical region. *Parasit Vectors.* 2014;7(1):76.
12. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, et al. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasites & vectors.* 2010;3(1):117.
13. WHO WHO, Others. Estimating functional survival of long-lasting insecticidal nets from field data. Vector Control Technical Expert Group Report to MPAC [Internet]. 2016 [cited 2021 Nov 23]. Available from: http://www.who.int/malaria/mpac/mpac_sep13_vcteg_llin_survival_report.pdf
14. Alonso PL, Brown G, Arevalo-Herrera M, Binka F, Chitnis C, Collins F, et al. A research agenda to underpin malaria eradication. *PLoS Med.* 2011;8(1):e1000406.
15. Tanner M, Greenwood B, Whitty CJM, Ansah EK, Price RN, Dondorp AM, et al. Malaria eradication and elimination: views on how to translate a vision into reality. *BMC Med.* 2015;13(1):167.
16. Osier FHA, Fegan G, Polley SD, Murungi L, Verra F, Tetteh KKA, et al. Breadth and magnitude of antibody responses to multiple *Plasmodium falciparum* merozoite antigens are associated with protection from clinical malaria. *Infect Immun.* 2008;76(5):2240–8.
17. Vounatsou P, Smith T, Smith AFM. Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *J R Stat Soc Ser*

- C (Applied Stat. 1998;47(4):575–87.
18. Smith T, Vounatsou P. Logistic regression and latent class models for estimating positivities in diagnostic assays with poor resolution. *Commun Stat Methods*. 1997;26(7):1677–700.
 19. Mwangi TW, Ross A, Snow RW, Marsh K. Case Definitions of Clinical Malaria under Different Transmission Conditions in Kilifi District, Kenya. 2005;
 20. Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C. Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nat Med*. 1999;5(3):340–3.
 21. Schellenberg JRMA, Smith T, Alonso PL, Hayes RJ. What is clinical malaria? Finding case definitions for field research in highly endemic areas. *Parasitol Today*. 1994;10(11):439–42.
 22. Rogers WO, Atuguba F, Oduro AR, Hodgson A, Koram KA. Clinical case definitions and malaria vaccine efficacy. *J Infect Dis*. 2006;193(3):467–73.
 23. Afrane YA, Zhou G, Githeko AK, Yan G. Clinical malaria case definition and malaria attributable fraction in the highlands of western Kenya. *Malar J*. 2014;13(1):405.
 24. Dicko A, Mantel C, Kouriba B, Sagara I, Thera MA, Doumbia S, et al. Season, fever prevalence and pyrogenic threshold for malaria disease definition in an endemic area of Mali. *Trop Med & Int Heal*. 2005;10(6):550–6.
 25. Atieli HE, Zhou G, Lee M-C, Kweka EJ, Afrane Y, Mwanzo I, et al. Topography as a modifier of breeding habitats and concurrent vulnerability to malaria risk in the western Kenya highlands. *Parasites & vectors*. 2011;4(1):241.
 26. Noor AM, Kinyoki DK, Mundia CW, Kabaria CW, Mutua JW, Alegana VA, et al. The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000–10: a spatial and temporal analysis of transmission intensity. *Lancet*. 2014;383(9930):1739–47.
 27. Depinay N, Franetich JF, Grüner AC, Mauduit M, Chavatte J-M, Luty AJF, et al.

- Inhibitory effect of TNF- α on malaria pre-erythrocytic stage development: influence of host hepatocyte/parasite combinations. *PLoS One*. 2011;6(3):e17464.
28. Ndungu FM, Mwacharo J, Kimani D, Kai O, Moris P, Jongert E, et al. A statistical interaction between circumsporozoite protein-specific T cell and antibody responses and risk of clinical malaria episodes following vaccination with RTS, S/AS01E. *PLoS One*. 2012;7(12):e52870.
 29. Ansong D, Asante KP, Vekemans J, Owusu SK, Owusu R, Brobbey NAW, et al. T cell responses to the RTS, S/AS01 E and RTS, S/AS02 D malaria candidate vaccines administered according to different schedules to Ghanaian children. *PLoS One*. 2011;6(4):e18891.
 30. Agnandji ST, Fendel R, Mestré M, Janssens M, Vekemans J, others. Induction of *Plasmodium falciparum*-Specific CD4⁺ T Cells and Memory B Cells in. 2011;
 31. Osier FHA, Feng G, Boyle MJ, Langer C, Zhou J, Richards JS, et al. Opsonic phagocytosis of *Plasmodium falciparum* merozoites: Mechanism in human immunity and a correlate of protection against malaria. *BMC Med*. 2014;12(1):1–15.
 32. Cutts JC, Powell R, Agius PA, Beeson JG, Simpson JA, Fowkes FJI. Immunological markers of *Plasmodium vivax* exposure and immunity: a systematic review and meta-analysis. *BMC Med*. 2014;12(1):1–20.
 33. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 2002;419(6906):498–511.
 34. Mitchell P. A perspective on protein microarrays. *Nat Biotechnol*. 2002;20(3):225–9.
 35. Chen B, Liu Y, Shen Y, Xia Z, He G, Liang S. Protein microarrays in proteome-wide applications. *J Proteomics & Bioinforma*. 2014;(S12):1.

36. Zong Y, Zhang S, Chen H-T, Zong Y, Shi Y. Forward-phase and reverse-phase protein microarray. In: Microarrays. Springer; 2007. p. 363–73.
37. Sundaresh S, Doolan DL, Hirst S, Mu Y, Unal B, Davies DH, et al. Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics*. 2006;22(14):1760–6.
38. Sboner A, Karpikov A, Chen G, Smith M, Dawn M, Freeman-Cook L, et al. Robust-linear-model normalization to reduce technical variability in functional protein microarrays. *J Proteome Res*. 2009;8(12):5451–64.
39. Corran PH, O'Donnell RA, Todd J, Uthapibull C, Holder AA, Crabb BS, et al. The fine specificity, but not the invasion inhibitory activity, of 19-kilodalton merozoite surface protein 1-specific antibodies is associated with resistance to malarial parasitemia in a cross-sectional survey in The Gambia. *Infect Immun*. 2004;72(10):6185–9.
40. Roussilhon C, Oeuvray C, Müller-Graf C, Tall A, Rogier C, Trape J-F, et al. Long-term clinical protection from falciparum malaria is strongly associated with IgG3 antibodies to merozoite surface protein 3. *PLoS Med*. 2007;4(11):e320.
41. Osier FH, Mackinnon MJ, Crosnier C, Fegan G, Kamuyu G, Wanaguru M, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Sci Transl Med*. 2014;6(247):247ra102--247ra102.
42. Stanisic DI, Richards JS, McCallum FJ, Michon P, King CL, Schoepflin S, et al. Immunoglobulin G subclass-specific responses against *Plasmodium falciparum* merozoite antigens are associated with control of parasitemia and protection from symptomatic illness. *Infect Immun*. 2009;77(3):1165–74.
43. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–6.
44. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Vol. 398. John Wiley & Sons; 2013.

45. Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*. 2012;28(13):1766–74.
46. Helb DA, Tetteh KKA, Felgner PL, Skinner J, Hubbard A, Arinaitwe E, et al. Novel serologic biomarkers provide accurate estimates of recent *Plasmodium falciparum* exposure for individuals and communities. *Proc Natl Acad Sci*. 2015;112(32):E4438--E4447.
47. Valletta JJ, Recker M. Identification of immune signatures predictive of clinical protection from malaria. *PLoS Comput Biol*. 2017;13(10):e1005812.
48. Rappuoli R, Aderem A. A 2020 vision for vaccines against HIV, tuberculosis and malaria. *Nature*. 2011;473(7348):463–9.
49. Kinyanjui SM, Bejon P, Osier FH, Bull PC, Marsh K. What you see is not what you get: implications of the brevity of antibody responses to malaria antigens and transmission heterogeneity in longitudinal studies of malaria immunity. *Malar J*. 2009;8(1):1–8.
50. Bustamante LY, Powell GT, Lin Y-C, Macklin MD, Cross N, Kemp A, et al. Synergistic malaria vaccine combinations identified by systematic antigen screening. *Proc Natl Acad Sci*. 2017;114(45):12045–50.
51. Mourão LC, Morais CG, Bueno LL, Jimenez MC, Soares IS, Fontes CJ, et al. Naturally acquired antibodies to *Plasmodium vivax* blood-stage vaccine candidates (PvMSP-119 and PvMSP-359-798) and their relationship with hematological features in malaria patients from the Brazilian Amazon. *Microbes Infect*. 2012;14(9):730–9.
52. Polley SD, Mwangi T, Kocken CHM, Thomas AW, Dutta S, Lanar DE, et al. Human antibodies to recombinant protein constructs of *Plasmodium falciparum* Apical Membrane Antigen 1 (AMA1) and their associations with protection from malaria. *Vaccine*. 2004;23(5):718–28.

53. Smith T, Schellenberg JA, Hayes R. Attributable fraction estimates and case definitions for malaria in endemic. *Stat Med*. 1994;13(22):2345–58.
54. Kamuyu G, Tuju J, Kimathi R, Mwai K, Mburu J, Kibinge N, et al. KILchip v1. 0: a novel *Plasmodium falciparum* merozoite protein microarray to facilitate malaria vaccine candidate prioritization. *Front Immunol*. 2018;9:2866.
55. Duarte JG, Blackburn JM. Advances in the development of human protein microarrays. *Expert Rev Proteomics*. 2017;14(7):627–41.
56. De Assis RR, Jain A, Nakajima R, Jasinskas A, Felgner J, Obiero JM, et al. Analysis of SARS-CoV-2 antibodies in COVID-19 convalescent blood using a coronavirus antigen microarray. *Nat Commun*. 2021;12(1):1–9.
57. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47--e47.
58. Mannsperger HA, Gade S, Henjes F, Beissbarth T, Korf U. RPPanalyzer: Analysis of reverse-phase protein array data. *Bioinformatics*. 2010;26(17):2202–3.
59. Gupta S, Manubhai KP, Kulkarni V, Srivastava S. An overview of innovations and industrial solutions in Protein Microarray Technology. *Proteomics*. 2016;16(8):1297–308.
60. Guo T, Ma H, Zhou Y. Bioinformatics analysis of microarray data to identify the candidate biomarkers of lung adenocarcinoma. *PeerJ*. 2019;7:e7313.
61. Turewicz M, Ahrens M, May C, Marcus K, Eisenacher M. PAA: an R/bioconductor package for biomarker discovery with protein microarrays. *Bioinformatics*. 2016;32(10):1577–9.
62. Duarte JDG, Goosen RW, Lawry PJ, Blackburn JM. PMA: Protein Microarray Analyser, a user-friendly tool for data processing and normalization. *BMC Res Notes*. 2018;11(1):1–6.

63. Xu Z, Huang L, Zhang H, Li Y, Guo S, Wang N, et al. PMD: A resource for archiving and analyzing protein microarray data. *Sci Rep*. 2016;6(1):1–5.
64. Fishman D, Kuzmin I, Adler P, Vilo J, Peterson H. PAWER: protein array web exploreR. *BMC Bioinformatics*. 2020;21(1):1–8.
65. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res*. 2015;43(W1):W566--W570.
66. Finney OC, Danziger SA, Molina DM, Vignali M, Takagi A, Ji M, et al. Predicting antidisease immunity using proteome arrays and sera from children naturally exposed to malaria. *Mol & Cell Proteomics*. 2014;13(10):2646–60.
67. Dubey A. Machine learning approaches in drug development of HIV/AIDS. *Int J Mol Biol Open Access*. 2018;3(1):23–5.
68. Chung AW, Kumar MP, Arnold KB, Yu WH, Schoen MK, Dunphy LJ, et al. Dissecting polyclonal vaccine-induced humoral immunity against HIV using systems serology. *Cell*. 2015;163(4):988–98.
69. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
70. Liaw A, Wiener M, others. Classification and regression by randomForest. *R news*. 2002;2(3):18–22.
71. Team RC, others. R: A language and environment for statistical computing. Vienna, Austria; 2013.
72. Tolo\csi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011;27(14):1986–94.
73. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
74. Ogutu BR, Apollo OJ, McKinney D, Okoth W, Siangla J, Dubovsky F, et al. Blood stage malaria vaccine eliciting high antigen-specific antibody concentrations

- confers no protection to young children in Western Kenya. PLoS One. 2009;4(3):e4708.
75. Mutabingwa TK, Bolla MC, Li J-L, Domingo GJ, Li X, Fried M, et al. Maternal malaria and gravidity interact to modify infant susceptibility to malaria. PLoS Med. 2005;2(12):e407.
 76. Bejon P, Mwacharo J, Kai O, Mwangi T, Milligan P, Todryk S, et al. A phase 2b randomised trial of the candidate malaria vaccines FP9 ME-TRAP and MVA ME-TRAP among children in Kenya. PLOS Clin Trial. 2006;1(6):e29.
 77. Thera MA, Doumbo OK, Coulibaly D, Diallo DA, Sagara I, Dicko A, et al. Safety and allele-specific immunogenicity of a malaria vaccine in Malian adults: results of a phase I randomized trial. PLoS Hub Clin Trials. 2006;1(7):e34.
 78. Dicko A, Sagara I, Diemert D, Sogoba M, Niambele MB, Dao A, et al. Year-to-year variation in the age-specific incidence of clinical malaria in two potential vaccine testing sites in Mali with different levels of malaria transmission intensity. Am J Trop Med Hyg. 2007;77(6):1028–33.
 79. Sagara I, Rulisa S, Mbacham W, Adam I, Sissoko K, Maiga H, et al. Efficacy and safety of a fixed dose artesunate-sulphamethoxypyrazine-pyrimethamine compared to artemether-lumefantrine for the treatment of uncomplicated falciparum malaria across Africa: a randomized multi-centre trial. Malar J. 2009;8(1):1–10.
 80. Trape J-F, Tall A, Sokhna C, Ly AB, Diagne N, Ndiath O, et al. The rise and fall of malaria in a west African rural community, Dielmo, Senegal, from 1990 to 2012: a 22 year longitudinal study. Lancet Infect Dis. 2014;14(6):476–88.
 81. Zwetyenga J, Rogier C, Tall A, Fontenille D, Snounou G, Trape J-F, et al. No influence of age on infection complexity and allelic distribution in Plasmodium falciparum infections in Ndiop, a Senegalese village with seasonal, mesoendemic malaria. Am J Trop Med Hyg. 1998;59(5):726–35.

82. Trape J-F, Rogier C. Combating malaria morbidity and mortality by reducing transmission. *Parasitol today*. 1996;12(6):236–40.
83. Asante KP, Owusu-Agyei S, Cairns M, Dodoo D, Boamah EA, Gyasi R, et al. Placental malaria and the risk of malaria in infants in a high malaria transmission area in Ghana: a prospective cohort study. *J Infect Dis*. 2013;208(9):1504–13.
84. Owusu-Agyei S, Asante KP, Adjuik M, Adjei G, Awini E, Adams M, et al. Epidemiology of malaria in the forest-savanna transitional zone of Ghana. *Malar J*. 2009;8(1):1–10.
85. Ouédraogo A, Tiono AB, Diarra A, Sanon S, Yaro JB, Ouedraogo E, et al. Malaria morbidity in high and seasonal malaria transmission area of Burkina Faso. *PLoS One*. 2013;8(1):e50036.
86. Kilama M, Smith DL, Hutchinson R, Kigozi R, Yeka A, Lavoy G, et al. Estimating the annual entomological inoculation rate for *Plasmodium falciparum* transmitted by *Anopheles gambiae* sl using three sampling methods in three sites in Uganda. *Malar J*. 2014;13(1):1–13.
87. Mwai K, Kibinge N, Tuju J, Kamuyu G, Kimathi R, Mburu J, et al. protGear: A protein microarray data pre-processing suite. *Comput Struct Biotechnol J*. 2021;19:2518–25.
88. Smith TA. Measures of clinical malaria in field trials of interventions against *Plasmodium falciparum*. *Malar J*. 2007;6(1):1–7.
89. Wu CFJ. On the convergence properties of the EM algorithm. *Ann Stat*. 1983;95–103.
90. Gonçalves L, Subtil A, de Oliveira MR, do Rosário V, Lee P-W, Shaio M-F. Bayesian latent class models in malaria diagnosis. *PLoS One*. 2012;7(7):e40633.
91. Vounatsou P, Smith T, Kitua AY, Alonso PL, Tanner M. Apparent tolerance of *Plasmodium falciparum* in infants in a highly endemic area. *Parasitology*.

- 2000;120(1):1–9.
92. Collins GS, Ogundimu EO, Cook JA, Manach Y Le, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med*. 2016;35(23):4124–35.
 93. Altman DG, Royston P. The cost of dichotomising continuous variables. *Bmj*. 2006;332(7549):1080.
 94. Frøslie KF, Røislien J, Laake P, Henriksen T, Qvigstad E, Veierød MB. Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) Study. *BMC Med Res Methodol*. 2010;10(1):1–5.
 95. Weinberg CR. How bad is categorization? *Epidemiology*. 1995;345–7.
 96. Thoresen M. Spurious interaction as a result of categorization. *BMC Med Res Methodol*. 2019;19(1):1–8.
 97. Feng F, Ataca ST, Ran M, Wang Y, Breen M, Kepler TB. Gain-Scanning for Protein Microarray Assays. *J Proteome Res*. 2020;19(7):2664–75.
 98. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
 99. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17(1):29–39.
 100. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* [Internet]. 2002;18(Suppl 1):S96–104. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/18.suppl_1.S96
 101. Zhu X, Gerstein M, Snyder M. ProCAT: a data analysis approach for protein

- microarrays. *Genome Biol.* 2006;7(11):R110.
102. Gu Q, Sivanandam TM. Optimizing scan parameters for antibody microarray experiments: accelerating robust systems diagnostics for life sciences. *Omi a J Integr Biol.* 2014;18(6):385–98.
 103. Pickett S, Carriedo S, Wang C. Determining the Signal-to-Noise Ratio and Optimal Photomultiplier gain setting in the GenePix 4000B. Axon Instruments, Inc; 2001.
 104. Silver JD, Ritchie ME, Smyth GK. Microarray background correction: maximum likelihood estimation for the normal--exponential convolution. *Biostatistics.* 2009;10(2):352–63.
 105. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, et al. A comparison of background correction methods for two-colour microarrays. *Bioinformatics.* 2007;23(20):2700–7.
 106. Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics.* 2003;19(7):825–33.
 107. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4(2):249–64.
 108. McGee M, Chen Z. Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data. *Stat Appl Genet Mol Biol.* 2006;5(1).
 109. Brown CE. Applied multivariate statistics in geohydrology and related sciences. Springer Science \& Business Media; 2012.
 110. Delfani P, Dexlin Mellby L, Nordström M, Holmér A, Ohlsson M, Borrebaeck CAK, et al. Technical advances of the recombinant antibody microarray technology platform for clinical immunoproteomics. *PLoS One.* 2016;11(7):e0159138.
 111. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing

- transformation for gene-expression microarray data. *Bioinformatics*. 2002;18(suppl_1):S105--S110.
112. Dazard J-E, Rao JS. Joint adaptive mean--variance regularization and variance stabilization of high dimensional data. *Comput Stat \& data Anal*. 2012;56(7):2317--33.
 113. Ballman K V, Grill DE, Oberg AL, Therneau TM. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*. 2004;20(16):2778--86.
 114. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform*. 2018;19(1):1--11.
 115. Yuan J, Hegde PS, Clynes R, Foukas PG, Harari A, Kleen TO, et al. Novel technologies and emerging biomarkers for personalized cancer immunotherapy. *J Immunother cancer*. 2016;4(1):3.
 116. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*. 2011;6(2):e17238.
 117. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882--3.
 118. Friedman J, Hastie T, Tibshirani R, others. The elements of statistical learning. Vol. 1. Springer series in statistics New York; 2001.
 119. Liu H, Motoda H. Computational methods of feature selection. CRC Press; 2007.
 120. Gareth J, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.
 121. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC; 2019.
 122. Bruce P, Bruce A, Gedeck P. Practical Statistics for Data Scientists: 50+ Essential

Concepts Using R and Python. O'Reilly Media; 2020.

123. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Statistical Methodol [Internet]*. 2005 Apr 1 [cited 2020 Nov 6];67(2):301–20. Available from:
<https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2005.00503.x>4010.1111/%28ISSN%291467-9868.TOP_SERIES_B_RESEARCH
124. Rajaratnam B, Roberts S, Sparks D, Dalal O. Lasso regression: estimation and shrinkage via the limit of Gibbs sampling. *J R Stat Soc Ser B Stat Methodol*. 2016;153–74.
125. Sutton M. Bayesian Variable Selection. In: *Case Studies in Applied Bayesian Data Science*. Springer; 2020. p. 121–35.
126. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin*. 2010;20(1):101.
127. Ročková V, George EI. The spike-and-slab lasso. *J Am Stat Assoc*. 2018;113(521):431–44.
128. Castillo I, Schmidt-Hieber J, der Vaart A, others. Bayesian linear regression with sparse priors. *Ann Stat*. 2015;43(5):1986–2018.
129. Brown PJ, Griffin JE. Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal*. 2010;5(1):171–88.
130. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika*. 2010;97(2):465–80.
131. Armagan A, Dunson DB, Lee J. Generalized double Pareto shrinkage. *Stat Sin*. 2013;23(1):119.
132. Bhattacharya A, Pati D, Pillai NS, Dunson DB. Dirichlet--Laplace priors for optimal shrinkage. *J Am Stat Assoc*. 2015;110(512):1479–90.

133. Bhadra A, Datta J, Polson NG, Willard B, others. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* 2017;12(4):1105–31.
134. Bai R, Ghosh M. High-dimensional multivariate posterior consistency under global-local shrinkage priors. *J Multivar Anal.* 2018;167:157–70.
135. Ghosh P, Tang X, Ghosh M, Chakrabarti A, others. Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Anal.* 2016;11(3):753–96.
136. Genuer R, Poggi J-M. *Random Forests with R*. Springer; 2020.
137. Janitza S, Hornung R. On the overestimation of random forest's out-of-bag error. *PLoS One.* 2018;13(8):e0201904.
138. Sigrist F. Gaussian Process Boosting. *arXiv Prepr arXiv200402653*. 2020;
139. Sigrist F. gpboost: Combining Tree-Boosting with Gaussian Process and Mixed Effects Models [Internet]. 2021. Available from: <https://cran.r-project.org/package=gpboost>
140. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. *Stat Appl Genet Mol Biol.* 2010;9(1).
141. Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol.* 2008;7(1).
142. Barker M, Rayens W. Partial least squares for discrimination. *J Chemom A J Chemom Soc.* 2003;17(3):166–73.
143. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics.* 2011;12(1):1–17.
144. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol.* 2017;13(11):e1005752.

145. Hornung R, Wright MN. Block Forests: Random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics*. 2019;20(1):1–17.
146. Zierer J, Menni C, Kastenmüller G, Spector TD. Integration of “omics” data in aging research: from biomarkers to systems biology. *Aging Cell*. 2015;14(6):933–44.
147. Hornung R, Wright MN. blockForest package: Block Forests: Random Forests for Blocks of Clinical and Omics Covariate Data [Internet]. 2019. Available from: <https://cran.r-project.org/package=blockForest>
148. Marsh K, Kinyanjui S. Immune effector mechanisms in malaria. *Parasite Immunol*. 2006;28(1–2):51–60.
149. Smith T, Schellenberg JA, Hayes R. Attributable fraction estimates and case definitions for malaria in endemic. *Stat Med*. 1994 Nov;13(22):2345–58.
150. Mwangi TW, Ross A, Snow RW, Marsh K. Case Definitions of Clinical Malaria under Different Transmission Conditions in Kilifi District, Kenya. 2005;
151. Smith T, Charlwood JD, Kihonda J, Mwankusye S, Billingsley P, Meuwissen J, et al. Absence of seasonal variation in malaria parasitaemia in an area of intense seasonal transmission. *Acta Trop*. 1993;54(1):55–72.
152. Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics*. 1993;865–72.
153. Olotu A, Fegan G, Williams TN, Sasi P, Ogada E, Bauni E, et al. Defining clinical malaria: the specificity and incidence of endpoints from active and passive surveillance of children in rural Kenya. *PLoS One*. 2010;5(12):e15569.
154. Olotu A, Fegan G, Wambua J, Nyangweso G, Ogada E, Drakeley C, et al. Estimating Individual Exposure to Malaria Using Local Prevalence of Malaria Infection in the Field. *PLoS One* [Internet]. 2012 [cited 2021 Jun 4];7(3):e32929. Available from: www.plosone.org

155. Mogeni P, Williams TN, Fegan G, Nyundo C, Bauni E, Mwai K, et al. Age, Spatial, and Temporal Variations in Hospital Admissions with Malaria in Kilifi County, Kenya: A 25-Year Longitudinal Observational Study. Grais RF, editor. PLOS Med. 2016 Jun;13(6):e1002047.
156. Muthui MK, Mogeni P, Mwai K, Nyundo C, Macharia A, Williams TN, et al. Gametocyte carriage in an era of changing malaria epidemiology: A 19-year analysis of a malaria longitudinal cohort. Wellcome open Res. 2019;4.
157. O'Meara WP, Bejon P, Mwangi TW, Okiro EA, Peshu N, Snow RW, et al. Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. Lancet. 2008/11/06. 2008;372(9649):1555–62.
158. Snow RW, Kibuchi E, Karuri SW, Sang G, Gitonga CW, Mwandawiro C, et al. Changing malaria prevalence on the Kenyan coast since 1974: climate, drugs and vector control. PLoS One. 2015;10(6):e0128792.
159. Vounatsou P, Smith T, Kitua AY, Alonso PL, Tanner M. Apparent tolerance of Plasmodium falciparum in infants in a highly endemic area. Parasitology. 2000;120(1):1–9.
160. Small DS, Cheng J, Ten Have TR. Evaluating the efficacy of a malaria vaccine. Int J Biostat. 2010;6(2).
161. Scott JAG, Bauni E, Moisi JC, Ojal J, Gatakaa H, Nyundo C, et al. Profile: The Kilifi health and demographic surveillance system (KHDSS). Int J Epidemiol. 2012;41(April):650–7.
162. Murungi LM, Kamuyu G, Lowe B, Bejon P, Theisen M, Kinyanjui SM, et al. A threshold concentration of anti-merozoite antibodies is required for protection from clinical episodes of malaria. Vaccine. 2013;31(37):3936–42.
163. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2021. Available from: <https://www.r-project.org/>

164. Spiegelhalter D, Thomas A, Best N, Lunn D. OpenBUGS user manual. Version. 2007;3(2):2007.
165. Buis M. ZOIB: Stata module to fit a zero-one inflated beta distribution by maximum likelihood. 2012;
166. Liu F, Kong Y. zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. *R J.* 2015;7(2):34.
167. Nkumama IN, O'Meara WP, Osier FHA. Changes in malaria epidemiology in Africa and new challenges for elimination. *Trends Parasitol.* 2017;33(2):128–40.
168. Polley SD, Mwangi T, Kocken CHM, Thomas AW, Dutta S, Lanar DE, et al. Human antibodies to recombinant protein constructs of *Plasmodium falciparum* Apical Membrane Antigen 1 (AMA1) and their associations with protection from malaria. *Vaccine.* 2004;23(5):718–28.
169. Snow RW, Omumbo JA, Lowe B, Molyneux CS, Obiero J-O, Palmer A, et al. Relation between severe malaria morbidity in children and level of *Plasmodium falciparum* transmission in Africa. *Lancet.* 1997;349(9066):1650–4.
170. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *Am J Epidemiol.* 2004;159(7):702–6.
171. Sundaresh S, Doolan DL, Hirst S, Mu Y, Unal B, Davies DH, et al. Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics.* 2006;22(14):1760–6.
172. Doolan DL, Mu Y, Unal B, Sundaresh S, Hirst S, Valdez C, et al. Profiling humoral immune responses to *P. falciparum* infection with protein microarrays. *Proteomics.* 2008;8(22):4680–94.
173. Mordmüller B, Surat G, Lagler H, Chakravarty S, Ishizuka AS, Lalremruata A, et al. Sterile protection against human malaria by chemoattenuated PfSPZ vaccine. *Nature.* 2017 Feb 23;542(7642):445–9.

174. Barry AE, Trieu A, Fowkes FJI, Pablo J, Kalantari-Dehaghi M, Jasinskas A, et al. The stability and complexity of antibody responses to the major surface antigen of *Plasmodium falciparum* are associated with age in a malaria endemic area. *Mol Cell Proteomics*. 2011;10(11).
175. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*. 2002;18(Suppl 1):S96–104.
176. GenePix. GenePix Pro 4.0 User Guide Rev. G. 2002 [Internet]. GenePix; 2002. Available from:
https://ipmb.sinica.edu.tw/microarray/index.files/GenePix_Pro_4.1_Manual_RevG.pdf
177. Sundaresh S, Randall A, Unal B, Petersen JM, Belisle JT, Hartley MG, et al. From protein microarrays to diagnostic antigen discovery: A study of the pathogen *Francisella tularensis*. *Bioinformatics*. 2007;23(13).
178. Barbacioru CC, Wang Y, Canales RD, Sun YA, Keys DN, Chan F, et al. Effect of various normalization methods on Applied Biosystems expression array system data. *BMC Bioinformatics*. 2006;7:1–14.
179. Do JH, Choi D-K. Normalization of Microarray Data: Single-labeled and Dual-labeled Arrays. Vol. 22, *Mol. Cells*.
180. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: From disarray to consolidation and consensus. *Nat Rev Genet*. 2006;7(1):55–65.
181. Xie Y. knitr: a comprehensive tool for reproducible research in R. *Implement Reprod Comput Res*. 2014;3–32.
182. Rts S. Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet*. 2015;386(9988):31–45.

183. Rossum G Van. Python Software Foundation. Python Language Reference, Version 2.7. 1995.
184. Butts CT. network: a Package for Managing Relational Data in R. *J Stat Softw.* 2008;24(1):1–36.
185. Kuhn M. The caret package. *J Stat Softw.* 2009;28(5).
186. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol.* 2008;77(4):802–13.
187. Yang M, Wang M, Dong G. Bayesian variable selection for mixed effects model with shrinkage prior. *Comput Stat.* 2020;35(1):227–43.
188. Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, et al. Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. *Bayesian Stat.* 2011;9:165–85.
189. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw.* 2017;80(1):1–28.
190. Luke DA. A user's guide to network analysis in R. Springer; 2015.
191. Peng X, Li G, Liu Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J Comput Biol.* 2016;23(2):102–10.
192. Ospina R, Ferrari SLP. Inflated beta distributions. *Stat Pap.* 2010;51(1):111–26.
193. Plummer M. JAGS: Just another Gibbs sampler. URL <http://mcmc-jags.sourceforge.net>. 2013;
194. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat.* 1998;7(4):434–55.
195. Choi I, Chung AW, Suscovich TJ, Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, et al. Machine learning methods enable predictive modeling of antibody feature: function relationships in RV144 vaccinees. *PLoS Comput Biol.*

2015;11(4):e1004185.

196. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Stat Comput.* 2017;27(3):659–78.
197. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics.* 2008;9(1):1–11.
198. Pellagatti M, Masci C, Ieva F, Paganoni AM. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat Anal Data Min ASA Data Sci J.* 2021;14(3):241–57.
199. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. *J Stat Comput Simul.* 2014;84(6):1313–28.
200. Bühlmann P, Rütimann P, van de Geer S, Zhang C-H. Correlated variables in regression: clustering and sparse estimation. *J Stat Plan Inference.* 2013;143(11):1835–58.
201. Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, et al. Genetic structure and domestication history of the grape. *Proc Natl Acad Sci.* 2011;108(9):3530–5.
202. Wimmer V, Lehermeier C, Albrecht T, Auinger H-J, Wang Y, Schön C-C. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics.* 2013;195(2):573–87.
203. Pegha-Moukandja I, Imboumy-Limoukou R-K, Tchitoula-Makaya N, Mouinga-Ondeme A-G, Biteghe-Bi-Essone JC, Mba DN, et al. High level of specific anti-plasmodium falciparum merozoite IgG1 antibodies in rural asymptomatic individuals of dienga, south-eastern Gabon. *Eur J Microbiol Immunol.* 2017;7(4):247–60.
204. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids*

- Res. 2009;37(suppl\1):D539--D543.
205. Gathoni K. Identifying the Merozoite Targets of Protective Immunity to *Plasmodium Falciparum* Malaria. 2017.
 206. Trieu A, Kayala MA, Burk C, Molina DM, Freilich DA, Richie TL, et al. Sterile protective immunity to malaria is associated with a panel of novel *P. falciparum* antigens. *Mol \& Cell Proteomics*. 2011;10(9).
 207. Obiero JM, Campo JJ, Scholzen A, Randall A, Bijker EM, Roestenberg M, et al. Antibody biomarkers associated with sterile protection induced by controlled human malaria infection under chloroquine prophylaxis. *Msphere*. 2019;4(1):e00027--19.
 208. Stone WJR, Campo JJ, Ouédraogo AL, Meerstein-Kessel L, Morlais I, Da D, et al. Unravelling the immune signature of *Plasmodium falciparum* transmission-reducing immunity. *Nat Commun*. 2018;9(1):1–14.
 209. Borisov A, Runger G, Tuv E, Lurponglukana-Strand N. Zero-inflated boosted ensembles for rare event counts. In: *International Symposium on Intelligent Data Analysis*. 2009. p. 225–36.
 210. Liu F, Eugenio EC. A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Stat Methods Med Res*. 2018;27(4):1024–44.

Appendix

Ethics Approval



R14/49 Mr Kennedy Mwai Wambui

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M190121

NAME: Mr Kennedy Mwai Wambui
(Principal Investigator)
DEPARTMENT: Epidemiology and Biostatistics

PROJECT TITLE: Analytical frameworks for studies to identify correlates of disease protection in the presence of misclassification bias

DATE CONSIDERED: 25/01/2019

DECISION: Approved unconditionally

CONDITIONS:

SUPERVISOR: Ass. Prof E. Musenge, Prof F. Osier and Dr S. Kinyanjui

APPROVED BY: 
Dr CB Penny, Chairperson, HREC (Medical)

DATE OF APPROVAL: 19/07/2019

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **January** and will therefore be due in the month of **January** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES