Predicting malaria in an highly endemic country

using clinical and environmental data

Kate Zinszer, M.Sc.
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University
Montreal, Canada

April 2014

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies in
partial fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D.)

© Kate Zinszer 2014
# Table of Contents

ABSTRACT .................................................................................................................. I

RESUMÉ ..................................................................................................................... III

ACKNOWLEDGEMENTS ............................................................................................... V

STATEMENT OF FINANCIAL SUPPORT ................................................................... VII

STATEMENT OF ORIGINALITY ............................................................................... VIII

CONTRIBUTION OF AUTHORS ............................................................................... IX

LIST OF ABBREVIATIONS ........................................................................................ XI

LIST OF TABLES ........................................................................................................ XIII

LIST OF FIGURES ....................................................................................................... XIV

CHAPTER 1: INTRODUCTION ..................................................................................... 1
  1.1 DISSERTATION ORGANIZATION ........................................................................ 2

CHAPTER 2: BACKGROUND ....................................................................................... 3
  2.1 UGANDA ............................................................................................................. 3
  2.2 MALARIA TRANSMISSION AND SYMPTOMS ................................................. 4
  2.3 RISK FACTORS FOR INFECTION ..................................................................... 4
  2.4 CONTROL AND PREVENTION MEASURES .................................................. 5
    2.4.1 Diagnostics ................................................................................................. 5
    2.4.2 Treatment ................................................................................................... 5
    2.4.3 Insecticide-treated nets ............................................................................ 6
    2.4.4 Indoor residual spraying ......................................................................... 7
    2.4.5 Intermittent preventive treatment ............................................................ 8
    2.4.6 Malaria surveillance ................................................................................. 8

CHAPTER 3: RESEARCH OBJECTIVES ....................................................................... 11

CHAPTER 4: SCOPING REVIEW OF MALARIA FORECASTING ............................... 12
  4.1 PREFACE ............................................................................................................ 12
  4.2 MANUSCRIPT 1: A SCOPING REVIEW OF MALARIA FORECASTING: PAST WORK AND FUTURE DIRECTIONS ................................................................. 13

CHAPTER 5: DEFINING HEALTH FACILITY CATCHMENT AREAS .................... 37
  5.1 PREFACE ............................................................................................................ 37
  5.2 MANUSCRIPT 2: DETERMINING HEALTH-CARE FACILITY CATCHMENT AREAS IN UGANDA USING DATA ON MALARIA-RELATED VISITS ........................................... 38

CHAPTER 6: FORECASTING MALARIA USING ENVIRONMENTAL AND CLINICAL PREDICTORS ................................................................. 61
  6.1 PREFACE ............................................................................................................ 61
  6.2 MANUSCRIPT 3: FORECASTING MALARIA IN A HIGHLY ENDEMIC COUNTRY USING ENVIRONMENTAL AND CLINICAL PREDICTORS ......................................................... 62
  6.3 APPENDIX A: DESCRIPTION OF FINAL MODELS ....................................... 82
  6.4 SUPPLEMENT 1: ACCURACY MEASURE COMPARISONS ............................ 83
ABSTRACT

Malaria is a public health crisis, with between 154 and 289 million cases worldwide in 2011, 80% of which are in sub-Saharan Africa. International agencies have prioritized reduction of the malaria burden, investing an estimated $1.84 billion in 2012 alone in malaria control and prevention programs in malaria endemic countries. There has been a drastic increase in resources dedicated to malaria prevention and control efforts over the past decade. Malaria thrives in poor tropical and subtropical countries where local resources are limited. Accurate disease predictions and early warnings of increased disease burden can provide public health and clinical health services with information critical for targeting malaria control and prevention measures in an efficient manner. There have been numerous studies that have developed malaria forecasting models although the limitations of several of the studies include narrowly focusing on environmental predictors and the use of scale-dependent measures. Common, scale-free accuracy measures are essential, as they will facilitate the comparison of findings between studies and between methods and likely lead to improvements in the field of malaria forecasting. The aim of my thesis work was to develop and evaluate statistical models that integrate environmental and clinical data to forecast malaria across different settings in a highly endemic country.

Specifically, the first objective was to systematically examine and summarize the literature on malaria forecasting models. A scoping review was conducted and the findings of this review have informed the methods and predictors included in the forecasting models used in this thesis. The second objective was to evaluate different methods of defining the catchment areas of health facilities and this allowed us to estimate the geographic regions served by the health facilities and their study populations. The third and final objective was to identify significant predictors of malaria across different settings and forecast horizons. Two forecasting models were developed for each of
the six Uganda Malaria Surveillance Project (UMSP) sites, short-term (4 weeks) and long-term (52 weeks) models for a total of 12 models. Remote sensing data were obtained for the environmental predictors and the UMSP clinical data were obtained for clinic- and patient-level predictors. Models were evaluated in terms of forecast error on data that were not used for model development. Most of the models with the lowest forecast error included both environmental and clinical predictors, and the parameters of the models often varied within a site and across sites. Generally, the short-term models were able to predict variations in malaria counts whereas the intermediate and long-term models were more useful in predicting cumulative cases (e.g., number of cases within 30 weeks).

The collective work of this thesis should advance the field of public health surveillance and more specifically malaria forecasting in a number of ways: in providing methodological guidelines for future forecasting studies, in providing a simple method for catchment definition that can be applied to define the geographic limits of a forecasting model, in identifying the importance of clinical predictors for forecasting malaria, in demonstrating the development of forecasting models with high spatial and temporal resolutions, and in raising important points of consideration for future forecasting work.
RESUMÉ

Avec ses 154 à 289 millions de victimes en 2011, dont 80% provenaient de la partie subsaharienne d’Afrique, la malaria est un problème majeur de santé publique. Les agences internationales ont privilégié le contrôle de la malaria, en investissant un montant estimé à 1,84$ milliard en 2012 dans les programmes de prévention et de lutte contre la malaria et dans les pays où la malaria est endémique. Il y a eu une forte augmentation des ressources consacrées à la prévention et au contrôle de la malaria au cours de la dernière décennie. La malaria se développe dans les pays tropicaux et subtropicaux pauvres où les ressources sont limitées. Des prévisions précises de l’occurrence de la malaria et des alertes précoce permettant de détecter son augmentation peuvent fournir des outils aux cliniques de santé publique et de l'information essentielle pour cibler efficacement son contrôle et certaines mesures de prévention. Plusieurs études ont développé des modèles de prévision de la malaria mais la majorité de ces études comprennent des limites telles que leur focalisation sur des facteurs prédictifs de l'environnement seulement et sur des mesures de précision qui sont influencées par l'échelle utilisée. Des mesures de précision qui sont indépendantes de l'échelle et des mesures communes sont indispensables dans ce contexte, car elles faciliteront la comparaison des résultats entre les études et entre les méthodes et permettront d'améliorer les prévisions de la malaria. Le but de ce travail de thèse était de développer et d'évaluer des modèles statistiques qui intègrent des données cliniques et environnementales afin de prévoir la malaria dans les différents contextes d’un pays où la maladie est fortement endémique.

Plus précisément, le premier objectif était d'examiner systématiquement et de résumer la littérature scientifique sur les modèles de prévision de la malaria. Une revue exploratoire de la littérature a été menée et les résultats de cette étude ont guidé le choix des méthodes et des prédicteurs inclus dans les modèles de prévision. Le deuxième objectif était d'évaluer la manière de définir les circonscriptions des services de santé. Cette recherche nous a permis d'estimer les régions
géographiques desservies par les services de santé et les populations étudiées. Le troisième et dernier objectif était d'identifier les prédicteurs significatifs de la malaria dans les différents contextes et dans les différents horizons de prévisions. Deux modèles de prévision, un à court terme (4 semaines) et un à long terme (52 semaines), ont été élaborés pour chacun des six sites du projet de surveillance de la malaria en Ouganda (PSMO) totalisant 12 modèles. Des données de télédétection ont été obtenues pour les prédicteurs environnementaux et des données cliniques de PSMO ont été obtenus pour les prédicteurs sanitaires (dispensaires et patients). Les modèles ont été évalués en fonction de l'erreur de prévision sur des données qui n'ont pas été utilisées pour l'élaboration du modèle. La plupart des modèles avec l'erreur de prévision la plus faible incluaient à la fois des prédicteurs environnementaux et non environnementaux, et les paramètres des modèles variaient souvent au sein d'un site et entre les sites. En général, les modèles à court terme ont été en mesure de prédire les variations des cas de malaria, alors que les modèles intermédiaires et à long terme étaient plus utiles pour prédire les cas cumulatifs (par exemple, le nombre de cas dans les 30 semaines).

Les résultats de cette thèse devraient faire progresser le domaine de la prévision de la malaria de plusieurs façons: en fournissant des lignes directrices méthodologiques pour de futures études prévisionnelles, en fournissant une méthode simple pour définir les circonscriptions des services de santé applicable pour définir les limites géographiques d'un modèle de prévision, en démontrant l'importance des prédicteurs cliniques pour la prévision de la malaria, en fournissant un exemple de modèle de prévision avec une résolution spatiale et temporelle, et finalement, en soulevant des points importants à prendre en considération pour de futurs travaux de prévision.
ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisor, David Buckeridge. I am grateful for his support and encouragement in getting this project off the ground and for his guidance, technical and professional, throughout the thesis. Thank you for being a wonderful mentor and for providing a number of opportunities during my doctorate work. I also thank my committee members, Timothy Brewer and John Brownstein, for their insight, support, and enthusiasm during this project.

Ruth Kigozi has been instrumental in this project and I greatly appreciate her time and support. I am indebted to Sarah Staedke and Heidi Hopkins for supporting and facilitating the implementation of this project along with Grant Dorsey and Moses Kamya. I would also like to thank the rest of the Uganda Malaria Surveillance Project and Infectious Disease Research Collaboration teams including Humphrey Wanzira, Arthur Mpimbaza, and Asadu Sswerwanga for their assistance, support, and humour. I would also like to thank Peter Okiu and Connie Balayo from the National Malaria Control Programme, along with Sussann Nasr from President’s Malaria Initiative for their assistance. I thank Will McInnes and Shanley Thompon for their help with the remote sensing data. I encountered numerous individuals from a variety of agencies and institutions that assisted, directly or indirectly, with this project and I am grateful for their help.

I would especially like to thank my labmates within the Surveillance Lab, with special mention of Katia Charland, Aman Verma, Luc De Montigny, Katherine Morrison, and Luke Mondor. Katia has tremendous statistical knowledge and skill, and I have learned a great deal from her and truly appreciate the time that she has dedicated to helping this project. Aman is a patient teacher and I am grateful for his assistance with R and other analytic issues. Luc and Katherine both provided assistance with different spatial analyses while Luke provided excellent editorial comments on my work. Thank you. My experience as a doctorate student was greatly enhanced because of my labmates.
I would also like to thank Robyn Tamblyn, Tina Wolfson, and Abby Lippman for their support at different times during my PhD. My dear friends from the program: Carmen Messerlian, Serene Joseph, Lena Shah, Nisha Almeida, Britt McKinnon, Martine Elbejani – what a journey, I cherish your support and thank you for always listening. I sincerely thank the staff, especially André Yves, Katherine Hayden, Rosalba Pupo, Patricia Plouffe, faculty, and students within the department for their support and for making my experience in the department enjoyable. I would like to thank my friends and family for their encouragement and for not asking (too often) when I will be finishing. David Doucet, thank you for everything and especially for nodding even when you perhaps did not understand nor agree, and for the wonderful wine.
STATEMENT OF FINANCIAL SUPPORT

I would like to acknowledge the financial support that I received for my doctoral studies. I received doctoral awards from the Fonds de recherche santé, McGill University Health Centre Research Institute, the International Development Research Centre, and the Department of Epidemiology, Biostatistics, and Occupational Health (McGill University). I received travel awards for conferences, data collection, and for dissemination activities from the Faculty of Medicine (McGill University), Department of Epidemiology, Biostatistics, and Occupational Health (McGill University), Canadian Institutes of Health Research, and L’Axe en santé mondiale du Réseau de recherche en santé des populations de Québec.
STATEMENT OF ORIGINALITY

The work presented in this thesis constitutes original scholarship and advances the knowledge in the field of malaria epidemiology, malaria forecasting, and health facility catchment areas. In the first manuscript, I present a scoping review on this body of literature. The objective of the review was to identify and assess forecasting methods used to forecast malaria, and is intended to serve as a resource to inform future forecasting studies. This has been the first systematic review of the malaria forecasting literature. The second manuscript evaluates three different methods of defining health facility catchment areas. In this manuscript, a new method of catchment definition is presented and this method will be useful in resource-limited settings where information is often limited. The third manuscript presents catchment-specific forecasting models that contained environmental and clinical predictors. Short-term and long-term forecasting models were developed, and each model was evaluated for its forecasting accuracy. This is the first study to assess clinical predictors such as treatment in malaria forecasting.

While I have received guidance from my committee members and co-authors on substantive, statistical, and methodological aspects of this thesis, I declare that the conception, execution, and drafting of the work in this thesis were my own.
CONTRIBUTION OF AUTHORS

Manuscript 1: A scoping review of malaria forecasting: past work and future directions

I developed the initial concept for this manuscript, which evolved through numerous revisions with the co-authors. I conducted the literature searches and reviewed the literature along with Aman Verma and Zhuoyu Sun. Aman Verma also assisted with the data extraction. I created the tables and figures, and wrote the first draft of the manuscript. David Buckeridge, Timothy Brewer, and John Brownstein were central in suggesting material to remove or ideas to develop further. Katia Charland and David Buckeridge both played crucial roles with their methodological expertise. All authors critically reviewed and edited the final version of the manuscript.

Manuscript 2: Determining health-care facility catchment areas in Uganda using data on malaria-related visits

I developed the initial concept for this manuscript, which evolved through numerous revisions with the co-authors. I performed all analysis, created the tables and figures, and wrote the first draft of the manuscript. The methodology was developed in collaboration with Katia Charland. Ruth Kigozi, Grant Dorsey, and Moses Kamya provided the dataset and Grant Dorsey suggested material to remove or ideas to develop further. David Buckeridge was actively involved in the methodological discussions as well as the interpretation of results and editing all versions of the manuscript. All authors critically reviewed and edited the final version of the manuscript.
Manuscript 3: Forecasting malaria in a highly endemic country using environmental and clinical predictors

Zinszer K, Kigozi R, Charland C, Dorsey G, Brewer TF, Brownstein JS, Kamya MR, Buckeridge DL. [To be submitted to Emerging Infectious Diseases]

The research objectives for this manuscript were developed in collaboration with David Buckeridge. I performed all data management and cleaning, analyses, created the tables and figures, and wrote the first draft of the manuscript. Ruth Kigozi, Grant Dorsey, and Moses Kamya provided the UMSP dataset. Ruth Kigozi and Grant Dorsey also assisted with understanding the practical implications of this work and assisted with the interpretation of the results and suggestions for improvement. Timothy Brewer and John Brownstein provided meaningful suggestions for material to remove or ideas to develop further. Katia Charland and David Buckeridge assisted with the methodological aspects of the project and were actively involved in the discussions of results and interpretations as well as editing all versions of the manuscript. All authors critically reviewed and edited the final version of the manuscript.
LIST OF ABBREVIATIONS

ACTs: artemisinin-based combination therapy
ACF: autocorrelation function
AE: absolute error
AIC: Akaike information criterion
AL: artemether-lumefantrine
AR: autoregressive
ARIMA: auto-regressive integrated moving average
ARIMAX: auto-regressive integrated moving average with exogenous input
CI: confidence interval
CCR: cumulative case rate
CQ: chloroquine
EVI: enhanced vegetation index
GLMs: generalized linear models
GM: Grey model
HAC: heteroskedasticity and autocorrelation consistent estimators
HIV: Human immunodeficiency virus
IDRC: Infectious Disease Research Collaboration
IPT: intermittent preventive treatment
IRS: indoor residual spraying
ITNs: insecticide-treated nets
km: kilometer
LST: land surface temperature
MAE: mean absolute error
MAPE: mean absolute percent error
MASE: mean absolute scaled error
MIS: Malaria Indicator Survey
MODIS: moderate resolution imaging spectroradiometer
MOH: Ministry of Health
NDVI: normalized difference vegetation index
NMCP: National Malaria Control Programme
PACF: partial autocorrelation function
PMI: President’s Malaria Initiative
RDTs: rapid diagnostic tests
SA$_3$: seasonal average that incorporated deviations from the last three observations
SARIMA: seasonal auto-regressive integrated moving average
SIR: susceptible-infected-recovered model
SMAPE: symmetric mean absolute percentage error
SP: sulphadoxine-pyrimethamine
TRMM: Tropical Rainfall Measuring Mission
UMSP: Uganda Malaria Surveillance Project
UTM: Universal Transverse Mercator system
VSEIRS: vector-susceptible-exposed-infected-recovered-susceptible model
**LIST OF TABLES**

**Table 4.1** Characteristics of malaria forecasting studies included in review (n=29) ..........................................................20  
**Table 4.2** Summary of malaria forecasting methods (n=29) ..............................................................................................................25  
**Table 4.3** Time-varying predictors considered in malaria forecasting models ..................................................................................27  
**Table 4.4** Parameters included in the mathematical forecasting models ..............................................................................................28  
**Table 5.1** Cumulative rate of confirmed malaria cases at six health-care facilities in Uganda, by catchment area definition, 2012 ........................................................................................................46  
**Table 5.2** Parishes, population and malaria cases in health-care facility catchment areas, by catchment area definition, Uganda, 2012 ........................................................................................................47  
**Table 6.1** Potential predictors ..........................................................................................................................................................68  
**Table 6.2** UMSP site characteristics ..............................................................................................................................................71  
**Table 6.3** Predictors included in final models ..................................................................................................................................72  
**Table 6.4** Error for selected forecast horizons for the long-term forecast models, by UMSP site from June 1, 2012 to May 31, 2013 ..............................................................................................................75  
**Table 6.5** Percent error for cumulative case predictions, by UMSP site from June 1, 2012 to May 31, 2013 ...........................78  
**Table S1** Overall errors by short-term and long-term model using average forecasted value, by UMSP site from June 1, 2012 to May 31, 2013 ........................................................................................................83  
**Table S2-1** Predictors included in final models ...............................................................................................................................85  
**Table S2-2** SMAPE for forecast horizons 1 to 4 for short-term, intermediate-term, long-term models, by UMSP site from June 1, 2012 to May 31, 2013 ..................................................................................86  
**Table S3** SMAPE for short-term ARIMA models and 4-week smoothing models, and cumulative case error for long-term ARIMA models and 52-week smoothing models ........................................................................88
LIST OF FIGURES

Figure 2.1 Malaria endemicity levels in Uganda showing the approximate locations of the Uganda Malaria Surveillance Project (UMSP) health facilities ........................................................................................................................................ 9

Figure 4.1 Flow of literature searches and screening process ................................................................................................................................................................................................. 19

Figure 5.1 Cumulative rate of confirmed malaria cases in parishes containing patients who attended six health-care facilities, Uganda, 2012 ........................................................................................................................................ 45

Figure 5.3 The Nagongera health-care facility’s catchment area as determined using the three parameters: straight-line distance, road network distance and cumulative case ratio for malaria-related visits, Uganda, 2012 ........................................................................................................................................ 49

Figure A-1a Straight-line catchment of 5, 10, 20 and 30 km for the Aduku health facility .................................................................................................................................................... 53

Figure A-1b Road network catchment of 5, 10, 20 and 30 km for the Aduku health facility .................................................................................................................................................... 53

Figure A-1c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Aduku health facility ........................................................................................................................................ 54

Figure A-2a Straight-line catchment of 5, 10, 20 and 30 km for the Kamwezi health facility .................................................................................................................................................... 54

Figure A-2b Road network catchment of 5, 10, 20 and 30 km for the Kamwezi health facility .................................................................................................................................................... 55

Figure A-2c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Kamwezi health facility ........................................................................................................................................ 55

Figure A-3a Straight-line catchment of 5, 10, 20 and 30 km for the Kasambya health facility .................................................................................................................................................... 56

Figure A-3b Road network catchment of 5, 10, 20 and 30 km for the Kasambya health facility .................................................................................................................................................... 56

Figure A-3c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Kasambya health facility ........................................................................................................................................ 57

Figure A-4a Straight-line catchment of 5, 10, 20 and 30 km for the Kiihihi health facility .................................................................................................................................................... 57

Figure A-4b Road network catchment of 5, 10, 20 and 30 km for the Kiihihi health facility .................................................................................................................................................... 58

Figure A-4c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Kiihihi health facility ........................................................................................................................................ 58

Figure A-5a Straight-line catchment of 5, 10, 20 and 30 km for the Walukuba health facility .................................................................................................................................................... 59

Figure A-5b Road network catchment of 5, 10, 20 and 30 km for the Walukuba health facility .................................................................................................................................................... 59

Figure A-5c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Walukuba health facility ........................................................................................................................................ 60

Figure 6.1 Outpatient health facilities of the Uganda Malaria Surveillance Program (UMSP) ................................................................................................................................................................. 66

Figure 6.2 Forecast horizon 1 from the short-term models with observed counts of malaria by UMSP site from June 1, 2012 to May 31, 2013 ................................................................................................................................................................................................. 73

Figure 6.3 Forecast horizon 20 from the long-term models with observed counts of malaria by UMSP site from June 1, 2012 to May 31, 2013 ................................................................................................................................................................................................. 74

Figure 6.4 Error by forecast horizon for long-term models, by UMSP site from June 1, 2012 to May 31, 2013 ................................................................................................................................................................................................. 76

Figure 6.5 Weekly observed counts sized by SMAPE error for forecast horizon 1, by UMSP site from June 1, 2012 to May 31, 2013 ................................................................................................................................................................................................. 77
Figure 6.6 Cumulative case predictions and observed cumulative cases by UMSP site from June 1, 2012 to May 31, 2013..........................................................78

Figure S3 Short-term ARIMA model and Holt-Winters model for June 2012, by UMSP site.........................88

Figure C-1 Vegetation index interpolated and observed series by UMSP site .................................................95

Figure C-2 Daytime temperature (in Celisus) interpolated and observed series by UMSP site ..................96

Figure C-3 Nighttime temperature (in Celisus) interpolated and observed series by UMSP site ...............97

Figure F-1 Weekly observed counts sized by SMAPE error for forecast horizon 2 of the short-term models, by UMSP site from June 1, 2012 to May 31, 2013.................................................................100

Figure F-2 Weekly observed counts sized by SMAPE error for forecast horizon 3 of the short-term models, by UMSP site from June 1, 2012 to May 31, 2013.................................................................101

Figure F-3 Weekly observed counts sized by SMAPE error for forecast horizon 4 of the short-term models, by UMSP site from June 1, 2012 to May 31, 2013.................................................................102
CHAPTER 1: Introduction

The burden of malaria is a global public health crisis. In 2011, an estimated 154 to 289 million cases occurred with 80% of these cases located in sub-Saharan Africa.\(^1\) Malaria infections kill an estimated one million people each year, 75% of whom are children.\(^1,2\) Governments and international agencies have prioritized reduction of the malaria burden which has resulted in channeling enormous resources into prevention and control efforts, including mass distribution of antimalarial medications, insecticide-treated nets, insecticide spraying, and monitoring and evaluation. In 2012 alone, an estimated $1.84 billion was invested in programs in malaria-endemic countries.\(^1\)

Malaria prospers in poor tropical and subtropical countries where resources are limited.\(^3\) Accurate disease predictions and early warning signals of an increase in disease burden can provide public health and clinical health services with the information needed to implement targeted approaches for malaria control and prevention that make effective use of limited resources. Accordingly, epidemic preparedness and early warning systems were priority areas of malaria control and prevention identified at the Abuja Summit.\(^4,5\) Additionally, the World Health Organization (WHO) and Roll Back Malaria\(^6\) have advocated for early warning and forecasting systems to inform control measures and supply chain management.\(^5,7\)

Malaria forecasting models have been developed for many endemic countries.\(^8,12\) Typically, these models use data on environmental risk factors, such as weather conditions, to forecast malaria for a specific geographic area over a certain period of time. Malaria has been forecasted using an

\(^{1}\) African leaders from 44 malaria endemic countries met in Abuja, Nigeria on April 25, 2000, and signed the Abuja Declaration.\(^4\)

\(^{2}\) The Roll Back Malaria Partnership was established in 1998 by the World Health Organization, United Nations Children’s Fund, United Nations Development Programme, and the World Bank and now consists of more than 500 partners.\(^4\)
assortment of methods and significant malaria predictors have been identified in a variety of settings. Non-environmental predictors, such as antimalarial treatment, have not been explored in previous forecasting work. The variability in how forecast accuracy is calculated and the lack of common forecast accuracy measures does not readily allow for comparison across studies.

The first objective of this thesis was to systematically examine and summarize the existing body of evidence on malaria prediction models by identifying methods of prediction and evaluation, as well as current research gaps.

The second objective of this thesis was to evaluate different methods of defining health facility catchment areas, which would be used to define the geographic boundaries for forecasting models used in this project.

The third objective of this thesis was to determine the relevance of environmental and non-environmental predictors of malaria across different settings in Uganda.

1.1 Dissertation organization

The format of this thesis is manuscript-based and consists of a collection of three manuscripts as well as separate chapters to provide a cohesive body of work. Each manuscript corresponds to a chapter and also includes a preface that explains the purpose of the manuscript and its relation to the objectives of the thesis. The Background is presented in Chapter 2 followed by Manuscript 1 (Chapter 3), which is a scoping review of malaria forecasting studies. The Objectives of my thesis are summarized in Chapter 4 and Chapter 5 (Manuscript 2) presents a new method for defining catchment areas of health facilities. Chapter 6 (Manuscript 3) describes the forecasting work and Chapter 7 presents the Conclusions of the thesis. All publications cited in the body of the thesis and in each of the manuscripts are listed in the References section at the end of the thesis.
CHAPTER 2: Background

2.1 Uganda

Uganda has one of the highest incidence rates of malaria worldwide.\textsuperscript{1} In Uganda, malaria is endemic in greater than 95\% of the country and is the leading cause of morbidity and mortality. Each year, there are an estimated ten to twelve million clinically confirmed cases.\textsuperscript{13} A national household survey in 2009 (2009 MIS) estimated the malaria prevalence among children less than five years old was 42\%.\textsuperscript{14} Furthermore, approximately 30,000 children under the age of five die each year in Uganda\textsuperscript{15} and the WHO has ranked Uganda ninth in the world in terms of the number of malaria deaths.\textsuperscript{1}

Uganda is an East African landlocked country situated on a large plateau, which is bordered by mountains, valleys, and Lake Victoria.\textsuperscript{16} Uganda has a relatively high altitude, 1,300-1,500 meters above sea level, and a mean annual temperature that ranges from 16°C in the southwest, to 30°C in the northeast and 25°C in rest of the country.\textsuperscript{14} The vegetation is varied, with tropical rain forests in the south, wooded savanna in central Uganda and semi-desert conditions in the north.\textsuperscript{16} There are two rainy seasons in the south from March to May and from September to December, although the timing varies depending upon geographic area. In the north and northeast, there tends to be a single rainy season from April to October.\textsuperscript{14}

Uganda has an estimated population of 35 million composed of numerous tribes that were grouped together during the British colonial period.\textsuperscript{17} The predominate religion is Catholicism (84\%)\textsuperscript{17} and 73\% of the population 15 years and older are estimated to be literate.\textsuperscript{18} Agriculture is the principal industry and Uganda also has substantial natural resources such as copper, gold, and oil.\textsuperscript{17} The per capita gross domestic product is $1,310 with 25\% of the population living below the poverty line. Approximately 16\% of the population live in urban areas\textsuperscript{18} with Kampala being the largest urban center with a population of 1.5 million\textsuperscript{17}.
2.2 Malaria transmission and symptoms

Malaria infections are caused by five types of *Plasmodium* protozoa, although primarily by *P. falciparum* in Africa.\(^{19}\) The 2009 MIS showed that *P. falciparum* accounted for 99% of malaria infections in Uganda.\(^{14}\) Infected female *Anopheles* mosquitoes transmit malaria by transferring the protozoa through human blood meals.\(^{19}\) Symptoms of infection typically appear ten to fifteen days after transmission and can include fever, vomiting, and chills. More severe infection can lead to anemia, respiratory distress, encephalitis, renal failure, coma, and death. Approximately 2% of clinical attacks of malaria are severe with a 50% mortality rate for severe malaria.\(^{20}\) Malarial infection is not always symptomatic.\(^{21}\) In endemic regions of sub-Saharan Africa, two to five percent of children are estimated to have asymptomatic parasitaemia.\(^{22,23}\) Chronic effects of infection include anemia, neurologic, cognitive, and developmental complications.\(^{19}\) Symptom relapses do not occur with *P. falciparum* but recrudescence can happen, which is the recurrence of malaria (not a new infection) that is caused by the inadequate clearing of the parasites from the bloodstream.\(^{24}\) Recrudescence is caused by treatment failure, which can be due to resistance or treatment with too low of a dosage.\(^{25}\)

2.3 Risk factors for infection

Many aspects of host (humans), agent (*Plasmodium*), vector (*Anopheles*), and environment conditions influence the risk, spread, and reemergence of malaria in populations. Individual and population characteristics play an important role in malaria transmission. Malnourished individuals\(^{26}\), HIV-infected individuals\(^{27}\), young children\(^{19,28}\) and pregnant women\(^{19,29}\) have been shown to have an increased risk of infection. Housing materials\(^{30,31}\), poor sources of potable water\(^{32}\), low maternal education\(^{33}\), high population density\(^{34}\), and high migration\(^{35}\) have all been associated with increased risk of infection.

Several environmental conditions are necessary for malaria transmission. Precipitation results in breeding sites for the aquatic stages of the mosquito’s life cycle and temperatures of 25°C to 30°C.
provide optimum conditions for the parasite and also influences the longevity and feeding frequency of the mosquito. It was been found that 1,800 to 2,000 meters in altitude are the upper limits of transmission due to temperature decreases. Vegetation also plays an important role, providing resting places for adult mosquitoes.

2.4 Control and prevention measures

The goal of malaria control is to reduce malaria to a level that is no longer a public health problem and it also has been defined as a 75% reduction of malaria morbidity and mortality by 2015. Strategies used in malaria control rely on antimalarial drugs, insecticide-treated nets, vector control, and monitoring and evaluation. Malaria control and prevention in Uganda heavily rely on effective case management and treatment in addition to large-scale distribution of insecticide-treated nets, indoor residual spraying, intermittent preventive treatment of malaria during pregnancy, and surveillance.

2.4.1 Diagnostics

Limited access to diagnostic services in many malaria endemic countries has led to a reliance on symptom-based diagnosis, leading to overdiagnosing an estimated 47% to 75% of individuals presenting with fever at health clinics. Misdiagnosis of malaria leaves the underlying cause of fever untreated and the unnecessary antimalarial use contributes to parasite resistance. In addition, the low specificity of clinical-based diagnoses of malaria hampers surveillance and population-based malaria control strategies. For example, accurate estimates of disease prevalence are essential for planning purposes, such as in determining the demand for treatment or in identifying high priority areas or populations to target with interventions (e.g., insecticide spraying).

Malaria diagnostics are essential for appropriate case management and the Uganda National Malaria Control Policy recommends that parasitological diagnosis with either microscopy or a rapid diagnostic test should become the standard for malaria case management at all health facilities.
malaria diagnosis, microscopy is considered the gold standard due to its high sensitivity and specificity when used by well-trained staff.\textsuperscript{44} Compared to microscopy, rapid diagnostic tests (RDTs) are faster and require less training. In spite of this recommendation, a national survey in 2007 estimated that 75\% of the health facilities in Uganda lack diagnostic capacity for malaria.\textsuperscript{45} To address this gap, international and national partners are supporting the procurement and distribution of equipment for microscopy and rapid diagnostic tests in health facilities, in addition to the training of health personnel.\textsuperscript{39}

2.4.2 Treatment

There is widespread \textit{P. falciparum} resistance to monotherapies such as chloroquine and multidrug resistance is increasing, but effective drug therapy is a major focus of malaria control strategies.\textsuperscript{46} In Uganda, resistance to chloroquine (CQ) and sulphadoxine-pyrimethamine (SP) monotherapies progressed rapidly during the 1990s.\textsuperscript{47} Since 2004, the national policy in Uganda for malaria has been based upon artemisinin-based combination therapy (ACT). The first-line treatment for uncomplicated malaria is artemether-lumefantrine (AL) and intravenous artesunate is the recommended treatment for severe malaria.\textsuperscript{39} Maintaining the supply of ACTs has been a perpetual problem in Uganda although there have been recent improvements due to a local manufacturer having received approval for ACTs. In addition, Global Fund\textsuperscript{48} through the Affordable Medicines Facility Malaria Initiative has negotiated price reductions with drug manufacturers.\textsuperscript{49} Moreover, several partners have focused on strengthening case management of patients with malaria.

\textsuperscript{44}The Global Fund is an international organization consisting of governments, international development partners, private sector, communities, and other organizations.\textsuperscript{48} It was created in 2002 to increase the resources available for the control and prevention of AIDS, tuberculosis and malaria.
2.4.3 Insecticide-treated nets

There are large-scale distribution efforts of insecticide-treated nets (ITNs) in areas of stable and unstable malaria transmission.\(^1\) A systematic review determined that in areas with stable *P. falciparum* transmission, ITNs reduced the incidence of uncomplicated malarial episodes by 50% and by 62% in areas of unstable transmission (compared to no nets).\(^50\) ITNs were first introduced in Uganda in 1990s by non-governmental agencies covering small populations.\(^51\) Since this time, there has been rapid expansion in the number of nets distributed with the goal of achieving universal coverage (one net per two people) in Uganda.\(^39\)

The 2009 MIS showed that 47% of households in Uganda owned one or more ITNs and that 44% of pregnant women and 33% of children under five had slept under an ITN the night before the survey.\(^14\) ITNs are distributed through mass campaigns, antenatal care clinics, and immunization clinics.\(^39\) In 2010, seven million nets were distributed in campaigns targeting pregnant women and children under five.

2.4.4 Indoor residual spraying

Indoor residual spraying (IRS) is the spraying houses with insecticides to kill *Anopheles* mosquitoes and has been attributed to the elimination of malaria in different parts of world.\(^52\) In 2006, Uganda began using IRS in selected districts, initially focusing on epidemic prone regions in the southwest, but subsequently shifting to highly endemic areas in the north.\(^53\) IRS is conducted prior to the transmission peaks, where it is expected to have the greatest impact on malaria transmission.\(^39\) In 2012, approximately 810,000 households were sprayed twice as part of the national IRS strategy. Mosquito resistance to insecticides has been problematic, requiring changing of the class of pesticide used. Community sensitization via awareness programs, community meetings, and radio announcements are conducted prior, during, and after each spray round.
2.4.5 Intermittent preventive treatment

Intermittent preventive treatment (IPT) is the preventive use of antimalarials for targeted populations with the goal of preventing malaria infection. A full course of treatment is given regardless if an individual is parasitaemic. In Uganda, IPT is targeted to pregnant women, who receive the treatment (sulfadoxine-pyrimethamine) during their regular antenatal care visits. The 2009 MIS showed that 32% of women reported receiving IPT during their last pregnancy. Current efforts are focused on improving the low IPT uptake.

2.4.6 Malaria surveillance

Malaria surveillance is essential to guide program planning and management, and to inform governments and donors on progress towards malaria control targets. Surveillance data also inform the implementation strategies, evaluation, and resource allocation of control programs such as diagnostics, treatment, ITN, and IRS. The public health reporting structure for malaria in Uganda is based upon malaria reports collected routinely from official government health facilities by the Ugandan National Malaria Control Programme (NMCP). Most reports are based upon clinical diagnoses with no laboratory confirmation. Aware of biases in the routine public health data, the Uganda Ministry of Health and the Uganda Malaria Surveillance Project (UMSP) established a sentinel site surveillance program in 2006.

UMSP has adopted a sentinel site approach for monitoring the malaria burden in specific sites within the country. The sites include six different outpatient health facilities that provide care free-of-charge, including diagnostic testing and medications. The sentinel sites were selected to reflect the diversity in malaria transmission intensity, as determined by entomological surveys (Figure 2.1). This outpatient sentinel surveillance program was implemented in staggered fashion, starting in July 2006 with the final site opening in August 2008. All six sentinel sites have laboratory equipment and staff trained for malaria diagnosis, case management, and data collection. At all sites, information is recorded
for each patient that visits the outpatient clinic including sex, age, date of visit, residence, and diagnosis. All symptomatic individuals (i.e., presence of fever) have a laboratory confirmatory test for malaria administered. For each suspected malaria case, the following information is recorded: symptoms, age, sex, pregnancy status, residential location, comorbidities, date of test, type of laboratory test, test results, and type of antimalarial treatment.

Figure 2.1 Malaria endemicity levels in Uganda showing the approximate locations of the Uganda Malaria Surveillance Project (UMSP) health facilities (Modified image from Yeka et al.15)
Malaria confirmation is usually completed by microscopy (92%) although RDTs (8%) are used occasionally, depending upon test availability. The use of RDTs first became available in January 2009 with the majority of RDTs (63%) being used at a single site (Kamwezi). A study was conducted which assessed the sensitivity and specificity of microscopy and RDTs at each of the UMSP sentinel sites. For microscopy, the average sensitivity and specificity were 76% and 98%, respectively; for RDTs, 98% and 88%, respectively. There is continuous monitoring of the quality of the microscopy and RDTs testing performed at the UMSP facilities.

The data collection instruments include a single standardized case record form that is completed for each patient by the attending physician. The forms use tick boxes and lists of options to minimize transcription errors (Appendix A). A full-time data officer that is fully supported by UMSP at each site enters the data electronically, which is centralized at the UMSP headquarters in Kampala. Staff at each sentinel site (which includes one medical officer, two clinical officers, five nurses, five midwives, four nursing assistants, one dental officer, one lab technician, one lab assistant, one records officer, one health educator, one health assistant, and one data officer) underwent a six-day training course three to six month post-UMSP implementation. Following the initial training, each site is visited approximately every 1 to 2 months by members of the UMSP team for clinical, laboratory, and surveillance support.
CHAPTER 3: Research objectives

The goal of my thesis work was to develop evidence to guide the selection and use of malaria forecasting methods that use environmental and clinical predictors across different settings in a highly endemic country.

The specific objectives were:

1) To systematically examine and summarize the existing body of evidence on malaria prediction models by identifying methods of prediction and evaluation, as well as current research gaps.

2) To evaluate different methods of defining health facility catchment areas.

3) To identify significant predictors of malaria across different settings and forecast horizons.
CHAPTER 4: Scoping review of malaria forecasting

4.1 Preface

Malaria forecasting can be a valuable tool for public and clinical health services in providing predictions of disease burden and early signals warning of increased burden. There have been numerous studies that have developed malaria forecasting models and Manuscript 1 presents a review of this body of literature. The objective of the review was to identify and assess forecasting methods used to forecast malaria, and it is intended to serve as a resource to inform future forecasting studies. This scoping review provided the basis for my subsequent dissertation work; informing the forecasting methods, choice of predictors, and how to address current limitations in malaria forecasting models.
4.2 Manuscript 1: A scoping review of malaria forecasting: past work and future directions

Kate Zinszer$^{1,2}$, Aman D. Verma$^{1,2}$, Katia Charland$^{1,2,3,4}$, Timothy F. Brewer$^{1,2}$, John S. Brownstein$^{1,3,4,5}$, Zhuoyu Sun$^1$, David L. Buckeridge$^{1,2,6}$

Addresses & Affiliations:

1Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Canada

2Surveillance Lab, Clinical and Health Informatics Research Group, McGill University, Montreal, Canada

3Children’s Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Harvard University, Boston, USA

4Division of Emergency Medicine, Children’s Hospital Boston, Boston, USA

5Faculty of Medicine, McGill University, Montreal, Canada

6Agence de la Santé et des services Sociaux de Montréal, Direction de santé publique, 3725 Saint-Denis Street, Montréal, Québec, Canada H2L 1M3


Available at: http://bmjopen.bmj.com/content/2/6/e001992.long
Abstract

Objectives: There is a growing body of literature on malaria forecasting methods and the objective of our review is to identify and assess methods, including predictors, used to forecast malaria.

Design: Scoping review. Two independent reviewers searched information sources, assessed studies for inclusion and extracted data from each study.

Information sources: Search strategies were developed and the following databases were searched: CAB Abstracts, EMBASE, Global Health, MEDLINE, ProQuest Dissertations & Theses, and Web of Science. Key journals and websites were also manually searched.

Eligibility criteria for included studies: We included studies that forecasted incidence, prevalence, or epidemics of malaria over time. A description of the forecasting model and an assessment of the forecast accuracy of the model were requirements for inclusion. Studies were restricted to human populations and to autochthonous transmission settings.

Results: We identified 29 different studies that met our inclusion criteria for this review. The forecasting approaches included statistical modelling, mathematical modelling, and machine learning methods. Climate-related predictors were used consistently in forecasting models, with the most common predictors being rainfall, relative humidity, temperature, and the normalized difference vegetation index. Model evaluation was typically based upon a reserved portion of data and accuracy was measured in a variety of ways including mean squared error and correlation coefficients. We could not compare the forecast accuracy of models from the different studies as the evaluation measures differed across the studies.

Conclusions: Applying different forecasting methods to the same data, exploring the predictive ability of non-environmental variables, including transmission reducing interventions, and using common forecast accuracy measures will allow malaria researchers to compare and improve models and methods, which should improve the quality of malaria forecasting.
ARTICLE SUMMARY

Article focus

- Accurate predictions of malaria can provide public health and clinical health services with the information needed to strategically implement prevention and control measures.
- The diversity in forecasting accuracy measures and the use of scale-dependent measures limits the comparability of forecasting results, making it difficult to identify the optimal predictors and methods for malaria forecasting.
- The objective was to identify and assess methods, including predictors, used to forecast malaria.

Key messages

- When performing forecasting, it is important to understand the assumptions of each method as well as the associated advantages and disadvantages.
- Common accuracy measures are essential as they will facilitate the comparison of findings between studies and methods.
- Applying different forecasting methods to the same data and exploring the predictive ability of non-environmental variables, including transmission reducing interventions, are necessary next steps as they will help determine the optimal approach and predictors for malaria forecasting.

Strengths and limitations of this study

- The strength of this review is that it is the first review to systematically assess malaria forecasting methods and predictors, and the recommendations in the review, if followed, should lead to improvement in the quality of malaria forecasting.
- A limitation of a literature review is that unpublished methods, if any, are omitted from this review.
INTRODUCTION

In 1911, Christophers\textsuperscript{56} developed an early warning system for malaria epidemics in Punjab based upon rainfall, fever-related deaths, and wheat prices. Since that initial system, researchers and practitioners have continued to search for determinants of spatial and temporal variability of malaria to improve systems for forecasting disease burden. Malaria forecasting is now conducted in many countries and typically uses data on environmental risk factors, such as climatic conditions, to forecast incidence for a specific geographic area over a certain period of time.

Malaria can be forecasted using an assortment of methods and significant malaria predictors have been identified in a variety of settings. Our objective was to identify and assess methods, including predictors, used to forecast malaria. This review is intended to serve as a resource for malaria researchers and practitioners to inform future forecasting studies.

METHODS

We included in our scoping review studies that forecasted incidence, prevalence, or epidemics of malaria over time. Whereas a systematic review is guided by a highly focused research question, a scoping review covers a subject area comprehensively by examining the extent, range, and nature of research activity on a topic.\textsuperscript{57} The studies had to use models that included prior malaria incidence, prevalence, or epidemics as a predictor. A description of the forecasting model and an assessment of the forecast accuracy of the model were requirements for inclusion. Studies were restricted to human populations and to autochthonous transmission settings. We excluded studies that provided only spatial predictions, exploratory analysis (e.g., assessing temporal correlations), mortality predictions, and/or individual-level transmission modelling. Commentaries, descriptive reports, or studies that did not include original research were also excluded. Additionally, for studies that were related (e.g., same setting and same methods with different time periods), the study with the most comprehensive data was included in the review.
A review protocol was developed and electronic search strategies were guided by a librarian experienced in systematic and scoping reviews. Papers were identified using medical subject headings and key word combinations and truncations: [“forecast*” or “predictive model*” or “prediction model*” or “time serie*” or “time-serie*”; AND “malaria*”]. The searches were not restricted by year or language although our searches were restricted by the historical time periods of the databases. The citation searches began on April 18, 2011 and the final citation search was conducted on May 29, 2012. We searched the following databases: CAB Abstracts (1910-2012 Week 20), EMBASE (1947-2012 May 28), Global Health (1910-April 2012), MEDLINE (1948-May Week 3 2012), ProQuest Dissertations & Theses (1861-May 29, 2012), and Web of Science (1899-May 28, 2012). We performed manual searches of the Malaria Journal (2000-May 29, 2012) and the American Journal of Tropical Medicine and Hygiene (1921-May 2012). Grey literature was also searched using Google Scholar, based upon the same key words used to search the databases. Additionally, the websites of the World Health Organization and the United States Agency for International Development were also examined for any relevant literature. To ensure that all appropriate references were identified, hand searching of reference lists of all included studies was conducted and any potentially relevant references were incorporated into the review process.

The citations were imported into EndNote X5 (Thomas Reuters) for management. Two main reviewers (KZ, AV) examined all citations in the study selection process with the exception of articles in Chinese, which were reviewed by a third reviewer (ZS). The first stage of review involved each reviewer independently identifying potentially relevant studies based upon information provided in the title and abstract. If it was uncertain whether to include or exclude a study during the first stage of review, the citation was kept and included in the full article review.

The second stage of review involved each reviewer independently identifying potentially relevant studies based upon full article review; data abstraction occurred for those articles that met
the inclusion criteria. From each study, we abstracted the following: setting, outcome, covariates, data source(s), time frame of observed data, forecasting and model evaluation methodologies, final models and associated measures of prediction accuracy. Quality of the included studies was not assessed as the objective was to conduct a scoping review and not a systematic review. Any discordance among the reviewers regarding inclusion or exclusion of studies or with respect to the information abstracted from the included studies was resolved by consultation with another author (DB).

RESULTS

Our search identified 613 potentially relevant articles for the scoping review after duplicate citations were removed (Figure 4.1). We identified 29 different studies that met our inclusion criteria for this review; they are described briefly in Table 4.1. Malaria forecasting has been conducted in 13 different countries with China as the most frequent site of malaria forecasting. The size of the geographic region of study ranged from municipal level to larger administrative divisions such as country and provinces or districts. Almost all of the studies (97%) used health clinic records of malaria infections from the general population as their data source for malaria infections, with one study using cohort data. Eleven (38%) of the 29 studies used laboratory confirmation of malaria cases (microscopy and/or rapid diagnostic tests), seven (24%) used clinical confirmation, and two (7%) used a mixture of clinical and microscopic confirmation. Nine studies did not state whether they used clinical or microscopic confirmation of malaria.
Figure 4.1 Flow of literature searches and screening process
<table>
<thead>
<tr>
<th>Authors (reference no.)</th>
<th>Population and setting</th>
<th>Model specifics</th>
<th>Malaria outcome</th>
<th>No. of data points used for training/testing</th>
<th>Evaluation measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adimi et al.(^{58})</td>
<td>Community health post data from 2004-7 for 23 provinces in Afghanistan; clinical confirmation</td>
<td>23 linear regressions (1 for each province); included autoregressive, seasonal and trend parameters</td>
<td>Monthly cases</td>
<td>31/6 (varied between provinces but last 6 months used only for testing)</td>
<td>Root mean squared error &amp; absolute difference</td>
</tr>
<tr>
<td>Chatterjee and Sarkar(^{59})</td>
<td>Municipal data for 2002-5 for Chennai city, India; microscopic confirmation Health service data from 1997-2003 for Karuzi province, Burundi; clinical confirmation</td>
<td>Logistic regression; polynornial and autoregressive parameters</td>
<td>Monthly slide positivity rate</td>
<td>36/1</td>
<td>95% CI (for predicted value and compared to observed)</td>
</tr>
<tr>
<td>Gomez-Elipe et al.(^{10})</td>
<td>District health centre data from 1994-2001 for Kahnnooj District, Iran; microscopic confirmation</td>
<td>Linear regression; adjusted for population, lagged weather covariates, autoregressive and seasonal parameters</td>
<td>Monthly incidence</td>
<td>60/24; 1 month ahead forecasts</td>
<td>95% CI, correlation, p-value trend line of difference (between predicted and observed)</td>
</tr>
<tr>
<td>Haghdoost et al.(^{60})</td>
<td>District health centre data from 1994-2001 for Kahnnooj District, Iran; microscopic confirmation</td>
<td>Separate Poisson regressions for <em>P. vivax</em> and <em>P. falciparum</em>, population offset, lagged weather covariates, seasonality and trend parameters</td>
<td>10-day cases</td>
<td>213/73</td>
<td>Average percent error</td>
</tr>
<tr>
<td>Rahman et al.(^{61})</td>
<td>Hospital data from 1992-2001 for all divisions of Bangladesh; clinical confirmation</td>
<td>4 linear regressions (1 for each administrative division and 1 for all of Bangladesh);</td>
<td>Yearly cases</td>
<td>10, 1 year was removed from series at a time</td>
<td>Root mean squared error &amp; relative bias (observed-predicted)</td>
</tr>
<tr>
<td>Reference</td>
<td>Data Source and Time Period</td>
<td>Methods</td>
<td>Outcome Measures</td>
<td>Results/Findings</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>-------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| Roy et al. 62      | Municipal data for Chennai city (2002-4) and Mangalore city (2003-7), India; microscopic confirmation | Environmental covariate for weeks of highest correlation 2 linear regressions (1 for each city); adjusted for population, lagged weather covariates, autoregressive term, interaction terms, polynomial terms | Monthly SPR (Chennai), monthly cases (Mangalore) | 28/8 (Chennai), 48/12 (Mangalore); 1 month ahead 95% CI  
| Teklehaiianot et al. 63 | Health facility data from 1990-2000 for all districts in Ethiopia; microscopic confirmation | 10 Poisson regressions (1 for each district); lagged weather covariates, autoregressive term, time trend and indicator covariates for week of the year Poisson regression; lagged weather covariates, autoregressive term | Weekly cases | 572 (varied between districts, training & testing); 52 weeks (year) were removed from series at a time; 1-4 week ahead forecasts  
| Xiao et al. 64     | Medical and health unit data from 1995-2007 for Hainan province, China; microscopic confirmation | Monthly incidence | 144/12 | T-test (predictive value significantly different than actual)  
| Yacob and Swaroop 65 | Medical data from 1944-6 for all health districts in Punjab; clinical confirmation | 19 linear regressions (1 for each district); include coefficients of correlation between rainfall and epidemic figures from 1914 to 1943 | Seasonal epidemic figure* | Coefficient of correlation (between actual and predicted epidemic figure)  
| Yan et al. 66      | Municipal data from 1951-2001 for Chongquing city, China | Linear regression; logarithm curve | Yearly cases | 50/1  
|                    |                                                                                             |                                                                         |                                                       | Visual inspection of predicted within range of actual values  


<table>
<thead>
<tr>
<th>ARIMA forecasting studies</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abeku et al.</strong>&lt;sup&gt;9&lt;/sup&gt;</td>
<td>Health clinics data from 1986-99 for 20 areas in Ethiopia; mixture of microscopic and clinical confirmed</td>
<td>20 models (1 for each area) compared approaches: Overall average, seasonal average, seasonal adjustment, ARIMA</td>
<td>Monthly cases</td>
</tr>
<tr>
<td><strong>Briët et al.</strong>&lt;sup&gt;8&lt;/sup&gt;</td>
<td>Health facility data from 1972-2005 for all districts in Sri Lanka; microscopic confirmation</td>
<td>25 models (1 for each district) compared approaches: Holt-Winters, ARIMA (seasonality assessed with fixed effects or harmonics) and SARIMA; lagged weather covariates</td>
<td>Monthly cases of malaria slide positives</td>
</tr>
<tr>
<td><strong>Liu et al.</strong>&lt;sup&gt;67&lt;/sup&gt;</td>
<td>Data from 2004-10 for China</td>
<td>SARIMA</td>
<td>Monthly incidence</td>
</tr>
<tr>
<td><strong>Wangdi et al.</strong>&lt;sup&gt;68&lt;/sup&gt;</td>
<td>Health center data from 1994-2008 for 7 districts in Bhutan; microscopic and antigen confirmation</td>
<td>7 models (1 for each district): SARIMA and ARIMAX; lagged weather covariates</td>
<td>Monthly cases</td>
</tr>
<tr>
<td><strong>Wen et al.</strong>&lt;sup&gt;69&lt;/sup&gt;</td>
<td>Data from 1991-2002 for Wanning County, China</td>
<td>SARIMA</td>
<td>Monthly incidence</td>
</tr>
<tr>
<td><strong>Zhang et al.</strong>&lt;sup&gt;70&lt;/sup&gt;</td>
<td>CDC data from 1959-79 for Jinan city, China; clinical confirmation</td>
<td>SARIMA; lagged weather covariates</td>
<td>Monthly cases</td>
</tr>
<tr>
<td><strong>Zhou et al.</strong>&lt;sup&gt;71&lt;/sup&gt;</td>
<td>Data from 1996-2007 for Huaiyuan County, China; microscopic and clinical confirmation</td>
<td>SARIMA</td>
<td>Monthly incidence</td>
</tr>
<tr>
<td><strong>Zhu et al.</strong>&lt;sup&gt;72&lt;/sup&gt;</td>
<td>Data from 1998-2007 for Huaiyuan and Tongbai counties, China</td>
<td>SARIMA</td>
<td>Monthly incidence rates</td>
</tr>
<tr>
<td>Study</td>
<td>Data Source</td>
<td>Forecasting Method</td>
<td>Parameters</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td><strong>Mathematical forecasting studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaudart et al.</td>
<td>Data from cohort of children from 1996-2000 in Bancoumana (municipality), Mali from 1996-2006; microscopic confirmation</td>
<td>VSEIRS model</td>
<td>Monthly incidence rate</td>
</tr>
<tr>
<td></td>
<td>Health centre data (passive and active surveillance) for Kutch (1987-2007) and Balmer (1985-2005) Districts, India; microscopic confirmation</td>
<td>2 models (1 for each district); compared 2 types of VSEIRS model to linear and negative binomial regressions</td>
<td>Monthly incidence for parameter estimation; Seasonal totals (Sept-Dec) for epidemic forecasting</td>
</tr>
<tr>
<td>Laneri et al.</td>
<td>Health centre data (passive and active surveillance) for Kutch (1987-2007) and Balmer (1985-2005) Districts, India; microscopic confirmation</td>
<td>2 models (1 for each district); compared 2 types of VSEIRS model to linear and negative binomial regressions</td>
<td>Monthly incidence for parameter estimation; Seasonal totals (Sept-Dec) for epidemic forecasting</td>
</tr>
<tr>
<td><strong>Neural network forecast studies</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cunha et al.</td>
<td>Ministry of Health data from 2003-9 for Cornwall City, Brazil; microscopic confirmation</td>
<td>Compared neural network to linear regression</td>
<td>Monthly cases</td>
</tr>
<tr>
<td>Gao et al.</td>
<td>Data from 1994-9 for Honghe State, China Hospital and clinic data from 1994-2001 for 19 provinces, Thailand; microscopic confirmation</td>
<td>Neural network</td>
<td>Monthly incidence</td>
</tr>
<tr>
<td>Kiang et al.</td>
<td>Data from 1994-9 for Honghe State, China Hospital and clinic data from 1994-2001 for 19 provinces, Thailand; microscopic confirmation</td>
<td>19 neural networks (1 for each province); various architectures used (varied by province)</td>
<td>Monthly incidence</td>
</tr>
<tr>
<td><strong>Other forecasting methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fang et al.</td>
<td>Data from 1956-88 for Xuzhou City, China</td>
<td>Grey and Grey Verhulst models (1,1)</td>
<td>Yearly incidence</td>
</tr>
<tr>
<td>Gao et al.</td>
<td>Data from 1998-2005 for Longgang District, China</td>
<td>Grey model (1,1)</td>
<td>Yearly incidence</td>
</tr>
<tr>
<td>Study</td>
<td>Data Source</td>
<td>Model Used</td>
<td>Prediction Type</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Guo et al.</td>
<td>Data from 1988-2010 China Medical data from 1925-6 for health districts in Punjab; clinical confirmation</td>
<td>Grey model (1,1)</td>
<td>Yearly incidence</td>
</tr>
<tr>
<td>Gill</td>
<td>Medical data from 1925-6 for health districts in Punjab; clinical confirmation</td>
<td>Grey model</td>
<td>Yearly incidence</td>
</tr>
<tr>
<td>Medina et al.</td>
<td>Community health center data from 1996-2004 (14 centers) for Niono District, Mali; clinical confirmation</td>
<td>Multiplicative Holt-Winters model, age-specific rates (3 age groups); compared to seasonal adjustment method</td>
<td>Monthly malaria consultation rates</td>
</tr>
<tr>
<td>Xu and Jin</td>
<td>Data from 2000-5 for Jiangsu Province, China</td>
<td>Grey model</td>
<td>Yearly cases</td>
</tr>
</tbody>
</table>

CI, confidence interval; ARIMA, auto-regressive integrated moving average; SARIMA, seasonal auto-regressive integrated moving average; ARIMAX, auto-regressive integrated moving average with exogenous input; VSEIRS: vector-susceptible-exposed-infected-recovered-susceptible model

*Seasonal epidemic figure is the ratio of October incidence to mean spring incidence
†Epidemic potential is the coefficient of variability of fevers during the month of October for the periods of 1868-1921
Forecasting studies

The forecasting approaches included statistical modelling, mathematical modelling, and machine learning methods. The statistical methods included generalized linear models, Auto-Regressive Integrated Moving Average (ARIMA) models, and Holt-Winters models. The mathematical models were based upon extensions of the Ross-MacDonald susceptible-infected-recovered malaria transmission model. Other authors predicted malaria incidence using neural networks, a machine learning technique.

Table 4.2 Summary of malaria forecasting methods (n=29)

<table>
<thead>
<tr>
<th>Forecasting method</th>
<th>No. of studies (ref. no.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>1^10,58-66,74,75</td>
</tr>
<tr>
<td>ARIMA</td>
<td>7,8,9,67-73</td>
</tr>
<tr>
<td>Grey methods</td>
<td>4^77,79,82</td>
</tr>
<tr>
<td>Smoothing methods*</td>
<td>3^8,9,11</td>
</tr>
<tr>
<td>Neural networks</td>
<td>3^75,77,79</td>
</tr>
<tr>
<td>Mathematical models</td>
<td>2^73,74</td>
</tr>
<tr>
<td>Visual</td>
<td>1^81</td>
</tr>
</tbody>
</table>

References in bold indicate multiple comparisons. ARIMA, auto-regression integrated moving average; GLM, generalised linear model.

*Includes Holt Winters, seasonal average, seasonally adjusted average, and simple average.

Twelve studies (41%) included in the review used generalized linear models to forecast malaria counts, rates, or proportions through linear, Poisson, or logistic regression. All but one of the regression models included climate related covariates such as rainfall, temperature, vegetation, and/or relative humidity. Typically, the weather covariates were lagged, to account for the delayed effects of weather on malaria infections. Two studies explored the effects of including covariates as higher order polynomials. Several of the studies used a generalized linear model approach to time series analysis by including previous (lagged) malaria incidence as an autoregressive covariate in the
model. Some models included terms for season or year to account for seasonal and annual variations.

Seven studies (24%) used forecasting approaches based on ARIMA modelling with some including a seasonal component (SARIMA). While not explicitly stated, many studies used a transfer function model, also known as ARIMAX. Typically, these ARIMA based models incorporated various meteorological series as covariates although one study also included data on the malaria burden in neighboring districts.  

Four studies (14%) from China used the Grey method for malaria forecasting, none of which incorporated predictors other than malaria incidence.  

There were two studies (7%) that used mathematical models. Gaudart et al. included a vector component in a susceptible-infected-recovered (SIR) type model and used data from a cohort of children, remote sensing data, literature, and expert opinions of entomologists and parasitologists. The study by Laneri et al. used a vector-susceptible-exposed-infected-recovered-susceptible (VSEIRS) model although they incorporated two different pathways from recovery to susceptibility that were based upon different time scales (seasonal and inter-annual), mimicking different transmission intensities. They found that rainfall had a significant effect on the inter-annual variability of epidemic malaria and including rainfall as a predictor improved forecast accuracy. The parameters in their models were selected based upon the literature as well as laboratory findings.

We identified three studies (10%) that used neural networks in their analyses, and each study used different input data and a unique network structure. Two of the studies used weather variables to predict malaria incidence. Gao et al. also included evaporation and sunshine hours to predict malaria incidence, two variables that were not included in any other study.

As shown in Table 4.3, climate-related predictors were used consistently in forecasting models, with the most common predictors being rainfall, relative humidity, temperature, and
Table 4.3 Time-varying predictors considered in malaria forecasting models

<table>
<thead>
<tr>
<th>Predictor</th>
<th>No. of studies (ref. no.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rainfall</strong></td>
<td></td>
</tr>
<tr>
<td>Total rainfall</td>
<td>11,8,10,58,60,63,64,68,70,74,77</td>
</tr>
<tr>
<td>Average rainfall</td>
<td>262,76</td>
</tr>
<tr>
<td>Rainy day index*</td>
<td>18</td>
</tr>
<tr>
<td>Number of rainy days/month</td>
<td>176</td>
</tr>
<tr>
<td><strong>Humidity</strong></td>
<td></td>
</tr>
<tr>
<td>Average relative humidity</td>
<td>760,62,64,68,70,76,77</td>
</tr>
<tr>
<td>Minimum humidity</td>
<td>159</td>
</tr>
<tr>
<td>Maximum humidity</td>
<td>159</td>
</tr>
<tr>
<td><strong>Temperature</strong></td>
<td></td>
</tr>
<tr>
<td>Maximum air temperature</td>
<td>830,59,60,63,64,68,70,76</td>
</tr>
<tr>
<td>Minimum air temperature</td>
<td>710,59,63,64,68,70,76</td>
</tr>
<tr>
<td>Average air temperature</td>
<td>462,64,76,77</td>
</tr>
<tr>
<td>Average LST</td>
<td>258,77</td>
</tr>
<tr>
<td>Temperature condition index</td>
<td>161</td>
</tr>
<tr>
<td><strong>Vegetation</strong></td>
<td></td>
</tr>
<tr>
<td>Average NDVI</td>
<td>210,58</td>
</tr>
<tr>
<td>Maximum NDVI</td>
<td>273,77</td>
</tr>
<tr>
<td>Vegetation condition index</td>
<td>161</td>
</tr>
<tr>
<td><strong>Other environmental predictors</strong></td>
<td></td>
</tr>
<tr>
<td>Average air pressure</td>
<td>270,76</td>
</tr>
<tr>
<td>Average air evaporation</td>
<td>176</td>
</tr>
<tr>
<td>Sunshine hours</td>
<td>176</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
</tr>
<tr>
<td>Malaria in neighboring districts</td>
<td>18</td>
</tr>
<tr>
<td>Population</td>
<td>159</td>
</tr>
</tbody>
</table>

*Rainy day index: the number of days per month when rainfall was larger than zero divided by the number of days that a reading for rainfall was available; LST, land surface temperature; NDVI, normalized difference vegetation index.

One study accounted for the effect of malaria incidence in neighboring districts, but it was not a significant predictor and was excluded from the final model. The mathematical models included non-time varying parameters such as the reporting fraction of
cases (proportion of malaria cases in a population that is reported to public health), average life expectancy, and several vector characteristics, which are listed in Table 4.4.

Table 4.4 Parameters included in the mathematical forecasting models

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Reference no.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vector</strong></td>
<td></td>
</tr>
<tr>
<td>Mean developmental delay</td>
<td>74</td>
</tr>
<tr>
<td>Number of bites per night</td>
<td>73</td>
</tr>
<tr>
<td>Probability of a susceptible becoming infected after one single bite from a contagious human</td>
<td>73</td>
</tr>
<tr>
<td>Mortality per day</td>
<td>73</td>
</tr>
<tr>
<td>Density</td>
<td>73</td>
</tr>
<tr>
<td>Length of gonotrophic cycle</td>
<td>73</td>
</tr>
<tr>
<td>Time lag of NDVI influence</td>
<td>73</td>
</tr>
<tr>
<td>Lowest NDVI value to influence behaviour</td>
<td>73</td>
</tr>
<tr>
<td><strong>Humans</strong></td>
<td></td>
</tr>
<tr>
<td>Probability of a susceptible human becoming infected after one single infected bite</td>
<td>73</td>
</tr>
<tr>
<td>Probability of becoming susceptible after being resistant</td>
<td>73, 74</td>
</tr>
<tr>
<td>Probability of acquiring contagiousness</td>
<td>73, 74</td>
</tr>
<tr>
<td>Probability of losing contagiousness</td>
<td>73, 74</td>
</tr>
<tr>
<td>Average human life expectancy</td>
<td>74</td>
</tr>
<tr>
<td>Infectivity of quiescent cases relative to full-blown infections</td>
<td>74</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
</tr>
<tr>
<td>Reporting fraction*</td>
<td>74</td>
</tr>
</tbody>
</table>

*Reporting fraction is the fraction of malaria cases in the population that are reported to public health; NDVI, normalized difference vegetation index
Evaluation methods

Authors used different approaches to evaluate the accuracy of forecasting models. A typical approach was to segment the data into a model building or training portion with the other portion (the ‘holdout’ sample) used for model validation or assessing forecast accuracy. The cross-validation approach used by Rahman et al.\textsuperscript{61} and Teklehaimanot et al.\textsuperscript{63} excluded one year of data at a time, fit the model to the remaining data, computed forecast error (prediction residual) using data from the missing year, and then repeated the process for the subsequent years. The accuracy of the predictions was then estimated from the prediction residuals. Some of the studies used all the available data to fit a model and did not reserve data for assessing forecast accuracy.\textsuperscript{73,74}

Studies compared the forecasts to observed values using various measures: mean squared error, mean relative error, mean percentage error, correlation coefficients, paired t-tests (between predicted and observed values), 95% confidence intervals (of predicted values and determined if observed values fell within the interval), and visualizations (e.g., graphical representations of observed and predicted values).

Comparison of forecasting methods

We could not compare the forecast accuracy of models from different studies due to the lack of common measures and the lack of scale independent measures. However, we briefly discuss the findings from studies that compared different methods within a single study.

Abeku et al.\textsuperscript{9} found that their ARIMA models provided the least accurate forecasts when compared with variations of seasonal averages, and the most accurate forecasts were produced by the seasonal average that incorporated deviations from the last three observations (SA\textsubscript{3}). In contrast, Briet et al.\textsuperscript{8} found that the most accurate model varied by district and forecast horizon, but the SARIMA approach tended to provide the most accurate forecasts, followed by an ARIMA model with seasonality using a sine term, then Holt-Winters, with the SA\textsubscript{3} providing the least accurate
forecasts. They also considered independent time series, such as rainfall and malaria cases in neighboring districts, in the models. Medina et al.\textsuperscript{11} determined that their Holt-Winters method provided more accurate forecasts and the accuracy did not deteriorate as rapidly as with the SA\textsubscript{3} method. Cunha et al.\textsuperscript{75} found that their neural network provided more accurate predictions across all three forecast horizons (3, 6, and 12 months) when compared to a logistic regression model.

**DISCUSSION**

Malaria forecasting can be an invaluable tool for malaria control and elimination efforts. A public health practitioner developed a simple forecasting method, which led to the first early warning system of malaria.\textsuperscript{56} Forecasting methods for malaria have advanced since that early work, but the utility of more sophisticated models for clinical and public health decision-making is not always evident. The accuracy of forecasts is a critical factor in determining the practical value of a forecasting system. The variability in methods is a strength of malaria forecasting, as it allows for tailored approaches to specific settings and contexts. There should also be continued effort to develop new methods although common forecasting accuracy measures are essential as they will help determine the optimal approach with existing and future methods.

When performing forecasting, it is important to understand the assumptions of forecast models and understand the advantages and disadvantages of each. Forecast accuracy should always be measured on reserved data and common forecasting measures should be used to facilitate comparison between studies. One should explore non-climate predictors, including transmission reducing interventions, as well as different forecasting approaches based upon the same data.

**Differences between forecasting methods**

The regression approach to time series prediction attempts to model the serial autocorrelation in the data through the inclusion of autoregressive terms and/or sine and cosine functions for seasonality. Generalized linear regression models are used commonly and their main advantages are
their flexibility and the intuitive nature of this approach for many people relative to ARIMA models. For example, the temporal dynamics observed in time series plots can be feasibly managed in generalized linear models by including several cyclic factors, interaction terms, and numerous predictors. The main disadvantages are that generalized linear models do not naturally account for correlation in the errors and the models may need to be complex to capture all the dynamics of the relationship within a series and between two or more series. Failure to accurately model serial autocorrelation may bias the estimation of the effect of malaria-related variables as well as underestimate the standard errors. Crucially, the regression model residuals must be examined for autocorrelation and it was not always evident that this occurred in the studies we identified that used this method. Additionally, it was not apparent if any remedial measures were used to account for the effect of autocorrelation on estimates of variance, e.g. re-estimating standard errors using heteroskedasticity and autocorrelation consistent (HAC) estimators.

ARIMA models are designed to account for serial autocorrelation in time series; current values of a series can be explained as a function of past values and past shocks. With ARIMA models, once the series have been detrended through differencing, any remaining seasonality can be modelled as part of additional autoregressive or moving average parameters of a SARIMA model. A rule of thumb is that 50 observations are a minimal requirement for ARIMA models, whereas SARIMA models require longer time series. The transfer function model, ARIMAX, extends ARIMA by also including as predictors current and/or past values of an independent variable. An advantage of ARIMA models versus generalised linear models (GLMs) is that ARIMA models naturally represent features of temporal patterns, such as seasonality and autocorrelation. As with generalized linear regression models, the residuals of ARIMA models need to be examined for residual correlation. Also, when incorporating an input series into the model, pre-whitening should occur prior to the cross-correlation assessment for the transfer function models. Pre-whitening is
when the residuals from an ARIMA model for the input series are reduced to ‘white noise’ and the same ARIMA model is applied to the output series. Authors did not always report that they pre-whitened the series prior to assessing cross-correlations. The relationship between the two resulting residual series is then estimated by the cross-correlation function. Without pre-whitening, the estimated cross-correlation function may be distorted and misleading.

Four studies from China used the Grey method for malaria forecasting. This forecasting method is essentially a curve fitting technique based on a smoothed version of the observed data. The Grey model appears most useful in predicting malaria when using a very short time series and when there is a strong linear trend in the data. This is due to the nature of the GM(1,1) model which will always generate either exponentially increasing or decreasing series. Its value in malaria prediction beyond that of the simpler statistical modelling approaches is yet to be determined.

The approach to prediction differs between mathematical models and other approaches such as generalized linear models, ARIMA and Grey models. The Ross-Macdonald mathematical model divides the population under study into different compartments such as susceptible, infected, and recovered, and uses differential equations to model the transition over time of individuals from one group to another. By using differential equations, these models can represent explicitly the dynamics of malaria infection, mosquito populations and human susceptibility. The disadvantages of mathematical models include the difficulty in finding appropriate, setting-specific data for the parameters. Also, the computational complexity of these models increases with the number of parameters, resulting in the omission of relevant features of malaria dynamics in order for the model to be manageable.

A neural network is a machine learning method that connects a set of inputs (e.g., weather covariates) to outputs (e.g., malaria counts). The connection between inputs and outputs are made via ‘neurons’ and the number of links and corresponding weights are chosen to give the best
possible fit to the training data. Neural networks have been proven to be useful in their capacity to handle non-linear relationships as well as a large number of parameters, and also their ability to detect all possible interactions between predictor variables.\textsuperscript{96} Mathematical models and neural networks are able to capture thresholds or limits on malaria transmission, which cannot be readily captured by statistical approaches. For example, in generalized linear models, a small decrease in the temperature leads to a small decrease in malaria incidence. Neural networks and mathematical models can represent explicitly that there will be no malaria transmission below a certain temperature. The disadvantages of neural networks include difficulties in determining how the network is making its decision and its greater computational burden\textsuperscript{97}; both of which depend upon the number of input parameters included in the model. Additionally, neural networks have a greater susceptibility to overfitting\textsuperscript{96} and several thousand observations are typically required to fit a neural network with confidence.\textsuperscript{97} Malaria time series are unlikely to contain several thousands of observations, perhaps unless the observations are aggregated over time (e.g., monthly) and location (e.g., national level).

Researchers have examined many forecasting methods, but published articles tend to describe the application of a single method to a unique dataset. Direct comparison of methods would be easier if multiple malaria forecasting methods were applied to the same data. This approach would allow identification of the methods that provide the most accurate short-term, intermediate, and long-term forecasts, for a given setting and a set of predictors. It would also allow exploration of gains in forecast accuracy by using a weighted combination of forecasts from several models and/or methods.\textsuperscript{98}

**Malaria predictors**

It has been suggested that climate and meteorological predictors have greater predictive power when modelling malaria incidence in areas with unstable transmission as compared to areas with
stable endemicity.\textsuperscript{99} It is interesting to note that nearly all of the models focused narrowly on a small number of environmental predictors despite the importance of other predictors of malaria incidence, such as land use, bednets, indoor residual spraying, and antimalarial resistance. Forecast accuracy may be weakened if transmission-reducing interventions are not considered in the models.

**Forecast evaluation**

Model selection based upon model fitting criteria such as Akaike’s information criterion, Bayesian information criterion, or the coefficient of determination, are standard measures considered when choosing a regression model. Using such measures to guide forecast model selection may result in selecting models with a greater number of parameters and “over-fitting”, which tends to result in inaccurate forecasts.\textsuperscript{100} For the purposes of forecasting, visualizations of forecasts compared to observations and forecast accuracy measures, such as the mean absolute forecast error, provide more direct and intuitive model selection criteria. When choosing how much of the series to reserve for testing the model, it is recommended to reserve at least as much as the maximum forecast horizon.\textsuperscript{101} Cross-validation is a more efficient use of data than partitioning a data set into train and test segment, although it is more computational intensive. It is recommended in cross-validation that only prior observations be used for testing a future value.\textsuperscript{101}

Various direct measures were used to estimate forecasting error. Absolute measures, such as the mean absolute error (MAE), are relevant for measuring accuracy within a particular series but not across series because the magnitude of the mean absolute error depends on the scale of the data.\textsuperscript{102} Percent errors, such as mean absolute percent error (MAPE) are scale-independent but are not recommended when the data involves 0 counts as MAPE cannot be calculated with 0 values. Also, the MAPE places a heavier penalty on forecasts that exceed the observed compared to those that are less than the observed.\textsuperscript{103} In economics, a measure called mean absolute scaled error
(MASE) has been recommended as a accuracy measure for forecasting\textsuperscript{102}. We recommend incorporating MASE into malaria forecast evaluation as this evaluation measure will facilitate comparison between studies. We also recommend reporting MAE as it allows an intuitive interpretation of the errors. Additionally, MAPE should be reported and a constant such as 1, could replace the 0 values in the series, allowing the calculation of MAPE. An advantage of MAPE as that it considers scale variance. For example, if we observed 70 counts of malaria but predicted 60, MAPE would be 14.3, MAE 10, and MASE 0.7. If we observed 15 counts of malaria but predicted 5, MAPE would be 66.7, MAE 10, and MASE 0.7. MAPE and MASE could be used to compare findings across series and studies, and also compared to one another to understand if and how they differ in their ranking of forecast accuracy. The MAE, MAPE, and MASE should be provided as site-specific measures for each forecast horizon, as summary measures for each site, and finally as summary measures for each forecast horizon across all sites (within a study).

**Conclusion**

Accurate disease predictions and early warning signals of increased disease burden can provide public health and clinical health services with the information needed to strategically implement prevention and control measures. Potential barriers to their usefulness in public health settings include the spatial and temporal resolution of models and accuracy of prediction. Models that produce coarse forecasts may not provide the precision necessary to guide targeted intervention efforts. Additionally, technical skill and lack of readily available data may reduce the feasibility of model utility in practice, which should be considered in developing malaria forecasting models if the intent is to use these models in clinical or public health settings. Applying different forecasting methods to the same data, exploring the predictive ability of non-environmental variables, including transmission reducing interventions, and using common
forecast accuracy measures will allow malaria researchers to compare and improve models and methods, and lead to the improvement in the quality of malaria forecasting.

Acknowledgements

We would like to thank various authors for responding to our questions and also to gratefully acknowledge Lorie Kloda for her assistance in our literature search strategies. We would especially like to thank the reviewers for critically reading the manuscript and providing insightful suggestions.

Competing interests

None declared.

Funding

This work was supported by the Canadian Institutes of Health Research Interdisciplinary Capacity Enhanced Team grant number HOA-80072.
CHAPTER 5: Defining health facility catchment areas

5.1 Preface

Forecasting models are applicable to a specific population and when using clinical data, forecasts apply to the catchment population for the facility. We did not have information on non-UMSP health facilities, therefore required a catchment definition method that could use the information provided by the UMSP data. This led to the development of a new approach for health facility catchment definition. The following manuscript describes the cumulative case ratio method and evaluates this method along with two other common distance approaches for catchment definition. The cumulative case ratio approach should be particularly useful in settings with limited data. This approach allowed us to estimate the geographic regions served by the health facilities and their study populations, which provided the geographic boundaries for the subsequent forecasting work.
5.2 Manuscript 2: Determining health-care facility catchment areas in Uganda using data on malaria-related visits

Kate Zinszer\textsuperscript{a}, Katia Charland\textsuperscript{a}, Ruth Kigozi\textsuperscript{b}, Grant Dorsey\textsuperscript{c}, Moses R. Kamya\textsuperscript{d}, David L. Buckeridge\textsuperscript{a}

\textbf{Addresses & Affiliations:}

\textsuperscript{a}McGill University, Montreal, Canada

\textsuperscript{b}Uganda Malaria Surveillance Project, Kampala, Uganda

\textsuperscript{c}University of California, San Francisco, USA

\textsuperscript{d}Makerere University College of Health Sciences, Kampala, Uganda


Available at: \url{http://www.who.int/bulletin/volumes/92/3/13-125260.pdf}
Abstract

Objective: To illustrate the use of a new method for defining the catchment areas of health-care facilities based on their utilization.

Methods: The catchment areas of six health-care facilities in Uganda were determined using the cumulative case ratio: the ratio of the observed to expected utilization of a facility for a particular condition by patients from small administrative areas. The cumulative case ratio for malaria-related visits to these facilities was determined using data from the Uganda Malaria Surveillance Project. Catchment areas were also derived using various straight line and road network distances from the facility. Subsequently, the 1-year cumulative malaria case rate was calculated for each catchment area, as determined using the three methods.

Findings: The 1-year cumulative malaria case rate varied considerably with the method used to define the catchment areas. With the cumulative case ratio approach, the catchment area could include non-contiguous areas. With the distance approaches, the denominator increased substantially with distance, whereas the numerator increased only slightly. The largest cumulative case rate per 1000 population was for the Kamwezi facility: 234.9 (95% confidence interval, CI: 226.2–243.8) for a straight line distance of 5 km, 193.1 (95% CI: 186.8–199.6) for the cumulative case ratio approach and 156.1 (95% CI: 150.9–161.4) for a road network distance of 5 km.

Conclusion: Use of the cumulative case ratio for malaria-related visits to determine health-care facility catchment areas was feasible. Moreover, this approach took into account patients’ actual addresses, whereas using distance from the facility did not.
**Introduction**

Knowledge of a health-care facility’s catchment area is important for assessing health service utilization, for calculating population-based rates of disease and for performing other important analyses. Different approaches to defining catchment areas have been developed, mostly in the field of health services research.\(^{104-108}\) One simple way of establishing the boundaries of a catchment area is to use distance from the facility – either the straight line distance, the distance patients have to travel or the distance travelled by patients in a given time.\(^{109,110}\) Under this approach it is assumed that people will visit the closest facility, which implies that distance is the overriding factor influencing attendance. However, distance is only one of many factors that influence the choice of health-care facility; others are the services available and the perceived quality of care.\(^{107,109}\)

Another approach, termed the patient-flow method, is based on the proportion of patients visiting or admitted to a health-care facility who come from a particular administrative area, such as a census tract or a postal code area: if the proportion exceeds a set minimum, that administrative area is included in the facility’s catchment area.\(^{108,111}\) With this approach, the catchment area is not limited by the distance between a patient’s residence and the facility. However, an arbitrary threshold is usually imposed on the minimum proportion of patients who must come from a particular area for it to be included in the catchment area. For example, postal code areas that account for less than 1% of admissions to a facility may be excluded from the catchment area.\(^{112}\) Consequently, some individuals who live in an area not considered part of a facility’s catchment area may regularly attend the facility. The likelihood that these minority “users” would be regarded as living outside the catchment area increases with the size of the administrative area. Another limitation is that an area may be excluded from the catchment area even though a large proportion of its population, or even the entire population, uses the facility because the proportion of patients attending the facility from
that area does not exceed the minimum.\textsuperscript{106} In this case, the chance of exclusion increases as the area’s population decreases.

Here we propose a new method for defining the catchment area of a health-care facility that builds and improves on the patient flow approach: the catchment area is defined using a statistical measure – the cumulative case ratio, which is the ratio of the observed to the expected utilization of the health-care facility for a particular condition by patients in an administrative area. We illustrate our method by using data on the utilization of malaria-related services to define the catchment areas of six health-care facilities in Uganda. Then, for each facility, we compare the cumulative case rate of confirmed malaria cases in the catchment area derived using this approach with the rate in areas derived using the straight line or road network distance from the facility.

Methods

In this analysis we used data on outpatients attending health-care facilities for suspected malaria collected by the Uganda Malaria Surveillance Project, in which a sentinel-site approach to monitoring the malaria burden in the country was adopted. The surveillance programme was implemented in a staggered fashion: it started in July 2006 and the final site opened in August 2008. We selected six sites to represent the diversity of malaria epidemiology in Uganda. They were all government, level-IV health centres that provided care free of charge, including diagnostic testing and medications, as has been described previously,\textsuperscript{54} and all had the laboratory equipment and trained staff needed for malaria diagnosis, case management and data collection. The data collected for each patient presenting to outpatient clinics included the patients’ demographic characteristics and parish of residence, the results of malaria diagnostic tests, the diagnosis and the treatments prescribed. The parish is the second smallest administrative unit in Uganda and each parish contains 5000 to 6000 inhabitants. A standardized case report form was used and data were entered electronically at each site by a data officer, who was supported by the Uganda Malaria Surveillance Project.
We determined which parishes should be included in the catchment area of each facility using three different parameters: the straight line distance from the facility, the road network distance from the facility and the cumulative case ratio for malaria-related visits. To derive catchment areas based on straight-line distance, we used distances of 5, 10, 20 and 30 km. These distances were selected because the 2009 Uganda Malaria Indicator Survey\textsuperscript{14} found that 96% of respondents lived within 9 km of a health-care facility and because the 2009 Uganda Household Survey\textsuperscript{113} reported that the average distance of a household from a government hospital was 20 km. The catchment area included all parishes that fell within circles centred on the facility with radii of 5, 10, 20 and 30 km, respectively. To derive catchment areas based on the road network distance, we used road distances of 5, 10, 20 and 30 km along the road networks surrounding each facility. The catchment area included all parishes located within a road distance of 5, 10, 20 or 30 km, respectively, from the facility. In addition, parishes were included if they were located less than 2 km from the nearest road. A parish that did not lie entirely within the distance circle or within the road network distance was included in the catchment area only if over 50% of its surface area lay within the relevant limit. Otherwise, it was excluded.

The cumulative case ratio was defined as the ratio of the observed to the expected number of malaria-related visits to a facility from a parish. Malaria-related visits included all visits between 1 January 2010 and 31 December 2012 by patients who had suspected or confirmed malaria or who tested negative for the disease. We used malaria-related visits because we wanted to include all users of malaria-related services, not only confirmed cases. The expected number of malaria-related visits to a facility from a particular parish was calculated by multiplying the parish’s population by the cumulative case rate for that facility. A parish was included in the catchment area if the upper limit of the 95% confidence interval for the cumulative case ratio for that parish was 1 or greater because a ratio less than 1 indicated that the parish contributed significantly fewer malaria-related visits than expected for its population.
Catchment areas were derived for each of the six sentinel sites using the three parameters and the cumulative case rate for each catchment area, however derived, over a 1-year period was calculated. The numerator was the total number of malaria cases confirmed between 1 January 2010 and 31 December 2012 from all parishes included in the catchment area. The denominator was the total population of all parishes included in the catchment area, which was derived using population estimates from the 2002 Uganda Population and Housing Census. During this period, an average of 98% of all patients with malaria symptoms were tested for malaria: the proportion ranged from 97% to 100% over the six sites.

Catchment areas were plotted on mapping files obtained from the Uganda Bureau of Statistics and the geographical coordinates of all parishes were recorded using zone 35 north of the Universal Transverse Mercator coordinate system. All analyses were performed using R software v2.14.0 and ArcGIS 10 (esri, Redlands, USA).

Results

Figure 5.1 displays the cumulative case rate per 1000 population in each parish for all malaria cases confirmed during 2012 at one of the six Uganda Malaria Surveillance Project health-care facilities. The figure also shows the locations of the six facilities and the variation in disease burden and geographical spread. No catchment area definition was applied. It can be seen from the figure that the parishes with the highest cumulative case rates either contained a facility or was adjacent to one. Over 40% of parishes had a cumulative case rate of 1 per 1000 or less in 2012.

Figure 5.2 shows the catchment area of the Nagongera health-care facility, as determined using the three parameters: straight-line distance, road network distance and cumulative case ratio for malaria-related visits. The largest geographical area was obtained using a straight-line distance of 30 km, whereas the smallest was obtained using the cumulative case ratio. In addition, use of the cumulative case ratio led to the inclusion of noncontiguous parishes. Figures illustrating the
corresponding catchment areas for the other five health-care facilities are shown in Appendix A (also available at: http://surveillance.mcgill.ca/users/kzinszer/WHObulletin/index.php).

As shown in Table 5.1, the cumulative rate of confirmed malaria cases varied considerably for most sites according to the way in which the catchment area was defined. In particular, the rate decreased with increasing distance from the facility for both straight-line and road network distances and, generally, was highest when the catchment area was defined using a distance of 5 km. The largest rates were observed for the catchment area of the Kamwezi health-care facility: 234.9 per 1000 when defined using a straight-line distance of 5 km, 193.1 per 1000 when defined using the cumulative case ratio and 156.1 per 1000 when defined using a road network distance of 5 km. Although the denominator in the cumulative case rate calculation became much larger as distance increased, there was no corresponding increase in the numerator. Hence, generally, the cumulative case rate decreased as the size of the catchment area increased. Table 5.2 shows the number of parishes included in each catchment area. When catchment areas were defined using a straight-line distance of 30 km, they included an average of 105 parishes; in contrast, when defined using the cumulative case ratio, they included an average of 10 parishes. The catchment area of the Aduku health-care facility did not contain any parishes when defined using a road network distance of 5 km (i.e. less than 50% of each parish’s area lay within the defined distance).
Figure 5.1 Cumulative rate of confirmed malaria cases in parishes containing patients who attended six health-care facilities, Uganda, 2012; (Quintile categories of the CCR was used)

Table 5.1 Cumulative rate of confirmed malaria cases at six health-care facilities in Uganda, a by catchment area definition, 2012

<table>
<thead>
<tr>
<th>Catchment method</th>
<th>Aduku</th>
<th>Kamwezi</th>
<th>Kasambya</th>
<th>Kihhi</th>
<th>Nagongera</th>
<th>Walukuba</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCRb (95%CI)</td>
<td>CCR (95%CI)</td>
<td>CCR (95%CI)</td>
<td>CCR (95%CI)</td>
<td>CCR (95%CI)</td>
<td>CCR (95%CI)</td>
</tr>
<tr>
<td>Straight-line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 km</td>
<td>89.2 (82.0, 97.0)</td>
<td>234.9 (226.2, 243.8)</td>
<td>218.3 (211.6, 225.1)</td>
<td>121.9 (117.5, 126.4)</td>
<td>119.9 (115.7, 124.3)</td>
<td>20.3 (19.6, 20.1)</td>
</tr>
<tr>
<td>10 km</td>
<td>45.9 (43.7, 48.2)</td>
<td>71.8 (69.3, 74.3)</td>
<td>116.2 (112.9, 119.6)</td>
<td>59.2 (57.3, 61.1)</td>
<td>43.4 (42.1, 44.7)</td>
<td>14.8 (14.3, 15.3)</td>
</tr>
<tr>
<td>20 km</td>
<td>13.3 (12.8, 14.0)</td>
<td>22.0 (21.2, 22.8)</td>
<td>39.5 (38.4, 40.7)</td>
<td>21.3 (20.7, 22.0)</td>
<td>12.1 (11.7, 12.4)</td>
<td>8.2 (7.9, 8.5)</td>
</tr>
<tr>
<td>30 km</td>
<td>7.2 (6.8, 7.5)</td>
<td>8.2 (7.9, 8.5)</td>
<td>21.8 (21.2, 22.5)</td>
<td>11.0 (10.7, 11.4)</td>
<td>6.1 (5.6, 6.7)</td>
<td>4.8 (4.6, 5.0)</td>
</tr>
<tr>
<td>Road networkd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 km</td>
<td>0 (-)</td>
<td>156.1 (150.9, 161.4)</td>
<td>218.3 (211.6, 225.1)</td>
<td>128.9 (124.3, 133.6)</td>
<td>143.3 (137.8, 148.9)</td>
<td>5.5 (5.1, 6.1)</td>
</tr>
<tr>
<td>10 km</td>
<td>48.5 (46.0, 51.1)</td>
<td>130.0 (125.6, 134.5)</td>
<td>218.3 (211.6, 225.1)</td>
<td>81.1 (78.6, 83.8)</td>
<td>49.3 (47.8, 50.9)</td>
<td>22.9 (22.2, 23.7)</td>
</tr>
<tr>
<td>20 km</td>
<td>17.6 (16.8, 18.5)</td>
<td>45.8 (44.2, 47.4)</td>
<td>49.0 (47.7, 50.5)</td>
<td>41.6 (40.3, 43.0)</td>
<td>17.2 (16.7, 17.7)</td>
<td>11.9 (11.5, 12.3)</td>
</tr>
<tr>
<td>30 km</td>
<td>11.7 (11.2, 12.2)</td>
<td>32.0 (30.9, 33.2)</td>
<td>33.0 (32.0, 33.9)</td>
<td>22.8 (22.1, 23.5)</td>
<td>9.0 (8.7, 9.2)</td>
<td>7.5 (7.3, 7.8)</td>
</tr>
<tr>
<td>Cumulative case ratioe</td>
<td>38.1 (36.3, 40.0)</td>
<td>193.1 (186.8, 199.6)</td>
<td>67.2 (65.3, 69.1)</td>
<td>87.6 (84.0, 90.3)</td>
<td>38.4 (37.3, 39.6)</td>
<td>43.3 (41.9, 44.7)</td>
</tr>
</tbody>
</table>

CCR cumulative case rate; CI, confidence interval; NA, not applicable.

a Data on the health-care facilities were collected by the Uganda Malaria Surveillance Project.

b Per 1000 population.

c Catchment areas were defined as lying within a specified straight-line distance from the facility.

d Catchment areas were defined as lying within a specified road network distance from the facility.

e A parish was included in the catchment area if the upper limit of the 95% CI for the cumulative case ratio for the parish (i.e. the ratio of observed to expected malaria-related visits from the parish) was 1 or greater.
Table 5.2 Parishes, population and malaria cases in health-care facility catchment areas, by catchment area definition, Uganda, 2012

<table>
<thead>
<tr>
<th></th>
<th>Aduku</th>
<th>Kamwezi</th>
<th>Kasambya</th>
<th>Kihhi</th>
<th>Nagongera</th>
<th>Walukuba</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>parish*</td>
<td>n/d†</td>
<td>parish*</td>
<td>n/d</td>
<td>parish*</td>
<td>n/d</td>
</tr>
<tr>
<td><strong>Straight-line</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 km</td>
<td>1</td>
<td>5,684</td>
<td>2</td>
<td>9,089</td>
<td>2</td>
<td>14,486</td>
</tr>
<tr>
<td></td>
<td>1,496</td>
<td>2,907</td>
<td>4,078</td>
<td>3,522</td>
<td>4,037</td>
<td>3,628</td>
</tr>
<tr>
<td>10 km</td>
<td>6</td>
<td>32,598</td>
<td>12</td>
<td>40,508</td>
<td>5</td>
<td>35,092</td>
</tr>
<tr>
<td></td>
<td>1,868</td>
<td>2,907</td>
<td>4,505</td>
<td>3,845</td>
<td>4,324</td>
<td>3,653</td>
</tr>
<tr>
<td>20 km</td>
<td>27</td>
<td>139,962</td>
<td>41</td>
<td>132,247</td>
<td>18</td>
<td>113,946</td>
</tr>
<tr>
<td></td>
<td>1,907</td>
<td>2,907</td>
<td>4,517</td>
<td>3,855</td>
<td>4,326</td>
<td>3,718</td>
</tr>
<tr>
<td>30 km</td>
<td>72</td>
<td>266,648</td>
<td>99</td>
<td>356,499</td>
<td>44</td>
<td>206,739</td>
</tr>
<tr>
<td><strong>Road network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 km</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>18,568</td>
<td>2</td>
<td>14,486</td>
</tr>
<tr>
<td></td>
<td>1,353</td>
<td>2,864</td>
<td>3,162</td>
<td>3,465</td>
<td>3,732</td>
<td>3,595</td>
</tr>
<tr>
<td>10 km</td>
<td>5</td>
<td>27,888</td>
<td>6</td>
<td>22,034</td>
<td>2</td>
<td>14,486</td>
</tr>
<tr>
<td></td>
<td>1,811</td>
<td>2,907</td>
<td>4,475</td>
<td>3,541</td>
<td>4,263</td>
<td>3,648</td>
</tr>
<tr>
<td>20 km</td>
<td>19</td>
<td>102,761</td>
<td>18</td>
<td>63,488</td>
<td>13</td>
<td>91,252</td>
</tr>
<tr>
<td></td>
<td>1,856</td>
<td>2,907</td>
<td>4,516</td>
<td>3,826</td>
<td>4,273</td>
<td>3,655</td>
</tr>
<tr>
<td>30 km</td>
<td>38</td>
<td>158,889</td>
<td>26</td>
<td>90,743</td>
<td>24</td>
<td>136,991</td>
</tr>
<tr>
<td><strong>Cumulative case ratio</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1,557</td>
<td>2,855</td>
<td>4,487</td>
<td>3,758</td>
<td>4,242</td>
<td>3,449</td>
</tr>
</tbody>
</table>

*a Catchment areas were defined as lying within a specified straight-line distance from the facility.

*b Catchment areas were defined as lying within a specified road network distance from the facility.

*c A parish was included in the catchment area if the upper limit of the 95% confidence interval for the cumulative case ratio for the parish (i.e. the ratio of observed to expected malaria-related visits from the parish) was one or greater.
Figure 5.3 The Nagongera health-care facility’s catchment area as determined using the three parameters: straight-line distance, a road network distance b and cumulative case ratio for malaria-related visits, c Uganda, 2012

a Catchment areas were defined as lying within a straight-line distance from the facility of 5, 10, 20 and 30 km, respectively.
b Catchment areas were defined as lying within a road network distance from the facility of 5, 10, 20 and 30 km, respectively.
c A parish was included in the catchment area if the upper limit of the 95% confidence interval for the cumulative case ratio for the parish (i.e. the ratio of observed to expected malaria-related visits from the parish) was 1 or greater.

Discussion

The differences observed between different estimates of the cumulative rate of confirmed malaria cases in catchment areas generally occurred because the numerator and denominator in the case rate calculation increased differentially with the distance used to define the catchment area. For example, the catchment area and its population were largest when a straight-line distance of 30 km was used; consequently, the denominator was also large. When a straight-line distance of 5 km was used, the numerator was only slightly smaller but the denominator was much smaller. Clearly the distance between a patient’s residence and the health-care facility was important but doubling the distance did not double the number of cases. Defining a catchment area according to distance from
the facility has the advantage of simplicity but this approach does not take into account where patients actually live. Moreover, although distance is important, it is not the only factor influencing a patient’s choice. Use of the cumulative case ratio is not affected by distance since it uses patients’ actual addresses. As Figure 5.2 demonstrates, this can result in catchment areas made up of noncontiguous parishes.

The main limitation of our approach follows from the assumption that the reason the number of malaria-related visits from a particular parish was lower than expected was primarily because utilization of the health-care facility by the parish’s population was low. However, lower than expected utilization could have been due to a low incidence of symptoms characteristic of malaria in the parish. If the purpose of defining a catchment area is to estimate a facility’s utilization rate, this assumption may not be of concern since areas of low utilization are excluded. However, if the purpose is to obtain a population-based estimate of disease burden, the assumption may introduce an error and additional data on the background level of the disease in question would be required to assess the potential level of that error. There are also limitations inherent in using health-care utilization data to calculate population-based estimates of disease burden. Although attendance at a health-care facility is influenced by several factors associated with the individual patient, it is also affected by the characteristics of the facility, such as its capacity, which will limit the number of cases that can be seen. In addition, we had no information on “competing” facilities, whose presence may have influenced attendance at the six facilities we studied. Consequently, we may have underestimated the true cumulative case rate of malaria in the catchment areas since our analysis included only cases at Uganda Malaria Surveillance Project facilities.

Accurate identification of a facility’s catchment area is important for: understanding the population served; for planning and evaluating service delivery, including the accessibility of services; and for deriving population-based health indicators, such as disease burden. Conversely, an
erroneous view of the catchment area can lead to inefficient and inadequate services, misspecification of the catchment population and potentially flawed decision-making on other facilities, such as deciding where to locate a new facility. Despite its limitations, the cumulative case ratio approach is more likely to produce an accurate estimate of the true catchment area than an approach using the straight line or road network distance because it is based on where patients actually live. In contrast, the distance approaches can lead to the inclusion of parishes where there are no cases of malaria.

The first consideration in using the cumulative case ratio approach to estimate a facility’s catchment area is to choose the basic administrative unit. It is best to use the smallest unit possible. The geographical location of patients must be recorded at the facility level and population data with the same geographical resolution must be obtained. Another consideration is the time period during which the catchment area is defined since catchment areas are affected by factors that vary over time, such as changes in the capacity of or the services provided by the facility or the opening or closing of other health-care facilities nearby. After exploring this issue, our findings (not reported) suggest that data from the most recent 2 or 3 years are sufficient for establishing current catchment areas. The use of a longer period would provide an insight into the stability of the catchment area over time. Catchment areas should be reassessed periodically. A final consideration is whether the catchment area should be based on all admissions to the facility or on the utilization of a particular service.

Our analysis demonstrates how population-based measures of disease burden, such as the malaria case rate, are dependent on the method used to define the catchment areas of health-care facilities. The cumulative case ratio approach to defining catchment areas we propose identified administrative units in which the utilization of a health-care facility was substantially lower than expected, thus enabling those units to be excluded from the facility’s catchment area. Our approach
is simple and reproducible and is based on using a statistical measure to decide which administrative units should be included in catchment areas.

**Funding:** This study was supported by doctoral awards from the International Development Research Centre in Ottawa, Canada and by the Fonds de recherche Santé, Montréal, Canada.
5.3 Appendix A: Maps of UMSP health facilities with different catchment definitions

Figure A-1a Straight-line catchment of 5, 10, 20 and 30 km for the Aduku health facility

Figure A-1b Road network catchment of 5, 10, 20 and 30 km for the Aduku health facility
Figure A-1c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Aduku health facility

Figure A-2a Straight-line catchment of 5, 10, 20 and 30 km for the Kamwezi health facility
Figure A-2b Road network catchment of 5, 10, 20 and 30 km for the Kamwezi health facility

Figure A-2c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Kamwezi health facility
Figure A-3a Straight-line catchment of 5, 10, 20 and 30 km for the Kasambya health facility

Figure A-3b Road network catchment of 5, 10, 20 and 30 km for the Kasambya health facility
Figure A-3c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Kasambya health facility

Figure A-4a Straight-line catchment of 5, 10, 20 and 30 km for the Kiihihi health facility
Figure A-4b Road network catchment of 5, 10, 20 and 30 km for the Kiihihi health facility

Figure A-4c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Kiihihi health facility
Figure A-5a Straight-line catchment of 5, 10, 20 and 30 km for the Walukuba health facility

Figure A-5b Road network catchment of 5, 10, 20 and 30 km for the Walukuba health facility
Figure A-5c Cumulative case ratio catchment with one-year parish-level cumulative case rates for the Walukuba health facility
CHAPTER 6: Forecasting malaria using environmental and clinical predictors

6.1 Preface

The aim of my thesis work was to develop and evaluate statistical models that integrated environmental and clinical data to forecast malaria across different settings in Uganda and Chapter 6 presents this work.

This chapter also includes supplemental materials that were not presented in the main body of the manuscript due to length considerations. For each of the UMSP catchment areas, a weekly time series model was built for short-term (4 weeks) and long-term (52 weeks) predictions of malaria. The models developed in this work address some of the limitations of previous published research by incorporating clinical predictors (e.g., antimalarial treatment) together with environmental predictors (e.g., rainfall), and by using a scale-independent measure of forecast accuracy. Furthermore, the forecast accuracy of the models was compared to other forecasting techniques, which is included as supplemental material (S1). This comparison was performed to understand how simpler forecasting methods would compare in forecast accuracy to the time series models. Other supplemental materials include the intermediate forecasting models (26 weeks) that were also built (S2), which were developed to understand how intermediate forecasting models differed in terms of predictors and performance relative to the short-term and long-term models. Additionally, three different forecasting accuracy measures were compared (S3): mean absolute percent error (MAPE), symmetric mean absolute percent error (SMAPE), and mean absolute scaled error (MASE). This was analysis was conducted to understand how the measures compared across different forecasting series.
6.2 Manuscript 3: Forecasting malaria in a highly endemic country using environmental and clinical predictors

Kate Zinszer\textsuperscript{a}, Ruth Kigozi\textsuperscript{b}, Katia Charland\textsuperscript{a}, Grant Dorsey\textsuperscript{c}, Timothy F. Brewer\textsuperscript{d}, John S. Brownstein\textsuperscript{e}, Moses R. Kamya\textsuperscript{f}, David L. Buckeridge\textsuperscript{a}

\textbf{Addresses & Affiliations:}

\textsuperscript{a}McGill University, 1140 Pine Ave West, Montreal, Quebec, Canada H3A 1A3

\textsuperscript{b}Uganda Malaria Surveillance Project, P.O. Box 7475, Kampala, Uganda

\textsuperscript{c}University of California San Francisco School of Medicine, 1001 Potrero Avenue, San Francisco, California, USA 94143

\textsuperscript{d}University of California Los Angeles, 2147 Murphy Hall, Los Angeles, California, USA 90095-1405

\textsuperscript{e}Boston Children’s Hospital, 300 Longwood Ave, Boston, Massachusetts, USA 02115

\textsuperscript{f}Makerere University College of Health Sciences, P.O. Box 7072 Kampala, Uganda

To be submitted to: \textit{Emerging Infectious Diseases}
ABSTRACT

Background
Malaria thrives in poor tropical and subtropical countries where local resources are limited. Accurate disease forecasts can provide public and clinical health services with the information needed to implement targeted approaches for malaria control that make effective use of limited resources.

Methods
Malaria forecasting models were developed using health facility data collected by the Uganda Malaria Surveillance Project and satellite-derived rainfall, temperature, and vegetation estimates. Short-term and long-term facility-specific forecasting models of confirmed malaria were developed using multivariate transfer function models.

Findings
The model with the most accurate forecasts varied by site and by forecast horizon. Clinical predictors were retained in the most accurate models across all facility sites and forecasting horizons, with the exception of one model. The average short-term error ranged from 20% to 96% over the forecasting period. The long-term models performed best for predicting the cumulative cases during the forecasting period (52 weeks) with error ranging from 2% to 22%.

Interpretation
Incorporating clinical predictors such as antimalarial treatment, improved the forecasting accuracy of several of the models. These results demonstrate the utility of using clinical predictors in conjunction with environmental predictors to forecast malaria. Despite our gains, variation in the temporal dynamics of malaria remains unexplained. With the mounting cost of the global fight against malaria and the drive towards elimination in many countries, accurate forecasts of malaria remain essential.
INTRODUCTION

Malaria forecasting methods have become more sophisticated since Christophers’ early work on forecasting malaria epidemics using rainfall, fever-related deaths, and wheat prices although the intent has remained unchanged: to inform malaria control and prevention by predicting burden or early warning of increasing burden. With the mounting cost of the global fight against malaria and the drive towards elimination in many countries, accurate forecasts of malaria remain essential. Accurate estimates of disease burden can inform the procurement and distribution of antimalarial medications and guide community-based prevention measures.

Malaria forecasting models have been developed in many endemic countries. Typically, these models use data on environmental risk factors, such as weather conditions, to forecast malaria for a specific geographic area over a certain interval of time. Clinical predictors, such as antimalarial treatment practices, have not been explored in previous forecasting work. Health facility data, however, are becoming increasingly available with the adoption of electronic medical records, allowing the inclusion of routinely collected clinical data into forecasting models. The objective of this study was to determine the relevance of environmental and clinical predictors of malaria across different settings in a highly endemic country. Uganda experiences one of the highest burdens of malaria in the world, where the disease is endemic in greater than 95% of the country and remains the leading cause of morbidity and mortality in Uganda. Ten to twelve million cases of malaria are estimated to occur each year in Uganda and the parasite prevalence was 42% among children less than five years of age based on a national survey conducted in 2009.

METHODS

Data

Outpatient health facility data collected by the Uganda Malaria Surveillance Project (UMSP) were used for this study. UMSP has adopted a sentinel site approach for monitoring malaria burden in
Uganda. The sentinel sites were established in a staggered fashion, starting in July 2006 with the final site opening in August 2008 (Figure 1). These sites were selected to represent the diversity of malaria transmission in Uganda and include six different health facilities that provide patient care free-of-charge, including diagnostic tests and medications. All sentinel sites staff trained in malaria diagnosis, case management, and data collection along with support for laboratory testing. Individual-level data collected from all patients presenting to the outpatient department include results of malaria diagnostic testing, diagnoses, treatments, as well as demographic information and parish of residence. Parishes are the second smallest administrative unit in Uganda with approximately 5,000 to 6,000 inhabitants. Data on the parish of residence for patients were used to determine the catchment area of each sentinel site.

Satellite sensor-derived environmental data were obtained from the Tropical Rainfall Measuring Mission (TRMM) and moderate resolution imaging spectroradiometer (MODIS) instruments onboard the Terra satellite. The TRMM product (TRMM3B42) provided daily rainfall estimates with a spatial resolution of 0.25° x 0.25° or 27.8 km x 27.8 km (at the equator). Daytime and nighttime temperature estimates were obtained from MODIS (MOD11A2) using 8-day composite images at a 1 km x 1 km resolution. The enhanced vegetation index (EVI) was also processed from MODIS (MOD13A1) using 16-day composite images at a 0.5 km x 0.5 km resolution.
Figure 6.1 Outpatient health facilities of the Uganda Malaria Surveillance Program (UMSP)

Measurement

We defined the primary outcome of interest (i.e., response series) as the weekly number of laboratory-confirmed malaria cases diagnosed at each sentinel site within the catchment area. We assessed the predictive power of variables representing the proportion of suspected cases (defined
by the presence of fever) that were tested for malaria, the number of suspected cases that tested negative for malaria, the type of antimalarial treatment prescribed, and the appropriateness of treatment (according to the Ugandan National Malaria Control Programme Policy). Each series was based upon the UMSP catchment areas of each sentinel site and the data were aggregated to a weekly frequency.

For each parish within the catchment area of a sentinel site, we calculated mean EVI and mean temperature (daytime and nighttime). We overlaid the parish boundaries with the raster environmental data (i.e., TRMM and MODIS) and calculated the weighted mean pixel value for each parish based on the proportion of the parish area contained within each pixel. Rainfall for a parish was calculated from the pixel that intersected with the center point of the parish polygon as the TRMM pixels were often larger than a parish. Total rainfall was the cumulative total of rainfall over a one-week period and other measures of rainfall were explored including logarithmically transformed total rainfall, maximum weekly rainfall, minimum weekly rainfall, and the weekly rainfall range. EVI and temperature values were interpolated to a weekly temporal resolution, given the different temporal frequencies. A linear spline was used to interpolate EVI and a quadratic spline to interpolate temperature measures. Approximately 9% of the observations were missing for nighttime temperature, and these values were imputed during the interpolation process using a quadratic spline. All polygons (parishes) were projected using the Universal Transverse Mercator system; zone 35 north (UTM35N).

Once weekly time series of environmental predictors were created for all parishes, a weekly average across all parishes within a catchment area was then calculated as a summary measure for each sentinel site. All environmental and clinical (i.e., UMSP data) predictor series began from the start of
consistent data collection (often coinciding with the site’s implementation date) at each sentinel site until May 31, 2013. All series are listed in Table 6.1.

Table 6.1 Potential predictors

<table>
<thead>
<tr>
<th>Series</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical data</strong></td>
<td></td>
</tr>
<tr>
<td>Confirmed malaria</td>
<td>Number of individuals with positive microscopy or rapid diagnostic test of malaria</td>
</tr>
<tr>
<td>Negative for malaria</td>
<td>Number of suspected (presence of fever) tested negative for malaria</td>
</tr>
<tr>
<td>Proportion tested</td>
<td>Number of individuals receiving appropriate antimalarial treatments based upon their malaria status and NMCP guidelines</td>
</tr>
<tr>
<td>Appropriate treatment</td>
<td>Number of individuals who were appropriately prescribed antimalarial arrests according to guidelines and malaria status</td>
</tr>
<tr>
<td>Inappropriate treatment</td>
<td>Number of individuals who were not prescribed an antimalarial arrest when they should have or were prescribed inappropriately</td>
</tr>
<tr>
<td>ACTs*</td>
<td>Number of individuals who did not receive appropriate antimalarial arrests based upon their malaria status and NMCP guidelines</td>
</tr>
<tr>
<td>Appropriate ACTs</td>
<td>Number of individuals who were appropriately prescribed ACTs according to guidelines and their malaria status</td>
</tr>
<tr>
<td>Inappropriate ACTs</td>
<td>Number of individuals who were not prescribed an ACT when they should have or were prescribed inappropriately</td>
</tr>
<tr>
<td>Quinine</td>
<td>Number of individuals who were appropriately prescribed quinine according to guidelines and malaria status</td>
</tr>
<tr>
<td>Appropriate quinine</td>
<td>Number of individuals who did not receive quinine when they should have or were prescribed quinine inappropriately</td>
</tr>
<tr>
<td>Inappropriate quinine</td>
<td>Number of quinine prescriptions</td>
</tr>
<tr>
<td>Chloroquine</td>
<td>Number of individuals who did not receive chloroquine when they should have or were prescribed chloroquine inappropriately</td>
</tr>
<tr>
<td>Inappropriate chloroquine</td>
<td>Number of chloroquine prescriptions</td>
</tr>
<tr>
<td><strong>Environmental data</strong></td>
<td></td>
</tr>
<tr>
<td>Daytime</td>
<td>LST at 3 pm (8-day composite image)</td>
</tr>
<tr>
<td>temperature</td>
<td>LST at 3 am (8-day composite image)</td>
</tr>
<tr>
<td>Nighttime</td>
<td>Cumulative sum of daily rainfall over a week period</td>
</tr>
<tr>
<td>temperature</td>
<td>Log of cumulative sum of daily rainfall over a week period</td>
</tr>
<tr>
<td>Total rainfall</td>
<td>Mean daily rainfall over a week period</td>
</tr>
<tr>
<td>Log total rainfall</td>
<td>Minimum daily rainfall over a week period</td>
</tr>
<tr>
<td>Mean rainfall</td>
<td>Maximum daily rainfall over a week period</td>
</tr>
<tr>
<td>Minimum rainfall</td>
<td>Difference between the maximum and minimum rainfall over a week period</td>
</tr>
<tr>
<td>Maximum rainfall</td>
<td></td>
</tr>
<tr>
<td>Rainfall range</td>
<td>Enhanced vegetation index (16-day composite image)</td>
</tr>
<tr>
<td><strong>Vegetation</strong></td>
<td></td>
</tr>
</tbody>
</table>

* artemisinin-based combination therapy (ACT)
Analysis

Multivariate transfer function models (ARIMAX) were developed for each sentinel site with a 52-week forecasting model developed per site. Data from the last 52 weeks of the confirmed malaria series (June 1, 2012-May 31, 2013) were reserved for testing the accuracy of the models, dividing the series into training (site implementation-May 31, 2012) and testing (June 1, 2012-May 31, 2013) series. To determine the lags at which the predictor series were correlated with the response series, the predictor and response series were pre-whitened and cross-correlations of the pre-whitened series were assessed. Pre-whitening involves fitting an ARIMA model to the input (predictor) series so that the residuals are “white noise” in that they exhibit random variation. The same ARIMA model is then applied to the output (response) series.\textsuperscript{119} The residuals from each series are subsequently used to estimate the cross-correlations between the predictor and response series. Typically, the lag of the pre-whitened predictor series that has the strongest correlation is included in the ARIMAX model. The pre-whitening process removes temporal autocorrelation, thus minimizing the occurrence of spurious correlations between the two series. Without pre-whitening, the estimated cross-correlation function may be biased and misleading.

The Akaike information criterion (AIC) was used to guide the model selection through a step-wise selection process, after the initial backward selection of potentially important predictor series. Model adequacy was assessed through inspection of residual autocorrelation diagnostics via the autocorrelation function (ACF), the partial autocorrelation function (PACF), and the Ljung-Box test. Further, histograms of the residuals and a normal quantile plots were used to assess the distribution of the residuals. Each model was used to generate weekly forecasts up 52 weeks in the future using the parameters that were chosen based upon the model selection. The forecasting (testing) period occurred from June 1, 2012 until May 31, 2013 and the models were implemented in a rolling fashion; after each weekly forecast, the model would be updated to include the observed malaria
counts and covariate values that occurred during that week before forecasting the malaria counts for the following week.

To enable comparison of forecasting errors across series and sites, we calculated the symmetric mean absolute percentage error (SMAPE) by forecast horizon:

\[ \text{SMAPE}_{\text{aveh}} = \frac{1}{n} \sum_{h=1}^{n} \frac{|Y_h - \hat{Y}_h|}{(Y_h + \hat{Y}_h)/2} \]

where SMAPE_{aveh} is the average SMAPE value for a horizon, \(Y_h\) is the observed value, \(\hat{Y}_h\) is the forecasted value, \(h\) is horizon, and \(n\) is the number of forecasts or observations for that horizon. Negative forecast values were set to zero and the upper bound of the SMAPE metric is 200%. For example, the error of all one-week ahead forecasts (horizon one) would be averaged to obtain an overall one-week ahead forecast error. In addition, the cumulative absolute error of each model was calculated using the entire series (52 weeks) of the forecasting series for each site.

RStudio v0.93 was used for data management, SAS v9.3 was used for the analyses, ERDAS Imagine v10.1 was used for processing the satellite images, and ArcGIS v10 was used for spatial analyses.

RESULTS

The Walukuba sentinel site experienced the largest number of malaria cases (n=29,664) relative to the other UMSP sites (Table 6.2). Nagongera had the youngest malaria cases and generally, there were slightly more female than male confirmed malaria cases at each site. On average, Nagongera had the highest temperatures (29.8°C) and received the most rain in 2012 (1.66 m).

The predictors included in the final models varied by sentinel site (Table 6.3). A commonly included category of predictor was drug treatment, with at least one treatment predictor series included in every model. Appropriate treatment and the number of ACTs were the most frequently
included treatment predictors. Total rainfall was the most commonly retained environmental predictor. Kamwezi’s models contained the smallest number of predictors, four, whereas Walukuba contained the most, fourteen. Approximately half the predictor series were lagged, ranging from lags of 1 to 52 weeks. A table of the predictors, noise terms, and lags are included in an appendix (Appendix A).

Table 6.2 UMSP site characteristics

<table>
<thead>
<tr>
<th>Site</th>
<th>Series start (# of weeks)*</th>
<th>Cumulative # of cases</th>
<th>Average age (years)</th>
<th>% female</th>
<th>Average daytime temperature (°C)</th>
<th>Average nighttime temperature (°C)</th>
<th>Cumulative rainfall for 2012 (meters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aduku</td>
<td>5-Nov-07 (291)</td>
<td>14,963</td>
<td>10.7</td>
<td>59%</td>
<td>29.1</td>
<td>17.6</td>
<td>1.29</td>
</tr>
<tr>
<td>Kamwezi</td>
<td>8-Sept-08 (247)</td>
<td>18,882</td>
<td>15.8</td>
<td>56%</td>
<td>27.7</td>
<td>15.8</td>
<td>1.01</td>
</tr>
<tr>
<td>Kasambya</td>
<td>10-Mar-08 (273)</td>
<td>20,636</td>
<td>13.7</td>
<td>57%</td>
<td>27.3</td>
<td>15.4</td>
<td>1.02</td>
</tr>
<tr>
<td>Kihhi</td>
<td>9-Jun-08 (261)</td>
<td>21,278</td>
<td>20.0</td>
<td>63%</td>
<td>27.3</td>
<td>16.5</td>
<td>1.05</td>
</tr>
<tr>
<td>Nagongera</td>
<td>16-Jun-08 (259)</td>
<td>20,716</td>
<td>8.4</td>
<td>56%</td>
<td>29.8</td>
<td>17.5</td>
<td>1.66</td>
</tr>
<tr>
<td>Walukuba</td>
<td>28-Apr-08 (266)</td>
<td>29,664</td>
<td>15.0</td>
<td>59%</td>
<td>28.6</td>
<td>16.9</td>
<td>1.22</td>
</tr>
</tbody>
</table>

*Total number of weeks or time points for the series (training and testing series)

Seasonality was assessed for each model although none of the models included seasonal terms as non-seasonal parameters and predictors were sufficient to model the seasonal variation. Autoregressive (AR) terms were included at lags that ranged from 1 to 43 weeks with some models including multiple AR terms whereas the moving average parameter was consistently 1 with a single order of differencing.
The short-term horizons (e.g., one to four weeks ahead) were better at predicting high-frequency variation in malaria cases (Figure 6.2) compared to longer-term horizons (Figure 6.3), although the short-term horizons predicted the peaks one to four weeks after they were observed. Kamwezi had the highest error with an average of 128% across all 52 forecast horizons and Nagongera had the lowest average error (27%) (Table 6.4). When examining the forecasting accuracy by forecast horizon, as expected, horizon one forecasts (i.e., one-week ahead forecasts), typically resulted in the smallest error. There were unexpected decreases in errors towards the end of the forecast horizons for all sites except for Kamwezi and Aduku and the SMAPE for each horizon was plotted (Figure 6.5). The weekly SMAPE error was highest when the observed counts were low or zero which occurred most often with the Kamwezi site (Figure 6.6). When examining the ability of models to predict the cumulative number of cases during the forecasting period, Nagongera had the lowest percent error at 2% (Table 6.5; Figure 6.7).
Figure 6.2 Forecast horizon 1 from the short-term models with observed counts of malaria by UMSP site from June 1, 2012 to May 31, 2013
Figure 6.3 Forecast horizon 20 from the long-term models with observed counts of malaria by UMSP site from June 1, 2012 to May 31, 2013
Table 6.4 Error for selected forecast horizons for the long-term forecast models, by UMSP site from June 1, 2012 to May 31, 2013

<table>
<thead>
<tr>
<th>Site</th>
<th>Horizon 1</th>
<th>Horizon 4</th>
<th>Horizon 12</th>
<th>Horizon 26</th>
<th>Horizon 52</th>
<th>Average*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aduku</td>
<td>31.6%</td>
<td>43.2%</td>
<td>62.1%</td>
<td>73.7%</td>
<td>99.5%</td>
<td>70.7%</td>
</tr>
<tr>
<td>Kamwezi</td>
<td>57.8%</td>
<td>117.0%</td>
<td>125.6%</td>
<td>147.1%</td>
<td>127.0%</td>
<td>127.8%</td>
</tr>
<tr>
<td>Kasambya</td>
<td>31.3%</td>
<td>42.8%</td>
<td>56.0%</td>
<td>42.9%</td>
<td>13.5%</td>
<td>46.6%</td>
</tr>
<tr>
<td>Kiihi</td>
<td>20.9%</td>
<td>31.1%</td>
<td>46.3%</td>
<td>31.2%</td>
<td>33.4%</td>
<td>37.0%</td>
</tr>
<tr>
<td>Nagongera</td>
<td>19.4%</td>
<td>27.8%</td>
<td>32.5%</td>
<td>31.9%</td>
<td>2.0%</td>
<td>26.3%</td>
</tr>
<tr>
<td>Walukuba</td>
<td>22.1%</td>
<td>30.7%</td>
<td>35.2%</td>
<td>37.3%</td>
<td>34.6%</td>
<td>30.8%</td>
</tr>
</tbody>
</table>

*Average error for horizons 1 to 52
Figure 6.4 Error by forecast horizon for long-term models, by UMSP site from June 1, 2012 to May 31, 2013
Figure 6.5 Weekly observed counts sized by SMAPE error for forecast horizon 1, by UMSP site from June 1, 2012 to May 31, 2013
Table 6.5 Percent error for cumulative case predictions, by UMSP site from June 1, 2012 to May 31, 2013

<table>
<thead>
<tr>
<th>Site</th>
<th>52 week burden estimate error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aduku</td>
<td>-8.4%</td>
</tr>
<tr>
<td>Kamwezi</td>
<td>-8.4%</td>
</tr>
<tr>
<td>Kasambya</td>
<td>-10.7%</td>
</tr>
<tr>
<td>Kihihi</td>
<td>22.1%</td>
</tr>
<tr>
<td>Nagongera</td>
<td>2.0%</td>
</tr>
<tr>
<td>Walukuba</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

Figure 6.6 Cumulative case predictions and observed cumulative cases by UMSP site from June 1, 2012 to May 31, 2013
DISCUSSION

The Abuja Declaration noted the importance of accurate disease prediction for targeting and evaluating control measures.\(^5\) For forecasting models to be useful for clinical and public health decision-making, models must produce accurate forecasts. We examined various predictors across six different settings in Uganda and consistently found that both environmental and clinical predictors were necessary to achieve the highest possible prediction accuracy. This is the first study, to the best of our knowledge, which examines clinical predictors in combination with environmental predictors for forecasting malaria. Future forecasting work should consider clinical predictors given the likelihood of their relevance across different endemic settings and an increased emphasis on the importance of laboratory testing for malaria.

Incorporating clinical predictors such as antimalarial treatment, the proportion of individuals screened for malaria, and the number of malaria negative individuals, produced models with the highest forecast accuracy across a range of settings and forecast horizons. These factors may be associated with malaria in different ways. For instance, inappropriate antimalarial treatment could result in cases subsequently returning to the health facility or could result in ongoing transmission of the parasite.\(^{22,120,121}\) The potential relationship between malaria-negative cases and screening for malaria is less clear although there may be a relationship between the number of non-malaria and malaria cases, relating to capacity of the clinic.\(^{122-124}\) Laboratory testing may be related to malaria in a manner similar to inappropriate treatment; if a person has malaria and is not tested, their malaria remains untreated causing a subsequent visit to the facility.\(^{125}\)

The accuracy of the models varied widely between the sites, with models at some sites influenced by low and zero counts in the response series, leading to large relative error measures (200%). The consistency in forecast accuracy over longer horizons (Figure 4) demonstrates that our models and predictors are stable across forecast horizons although at longer forecast horizons, we
would expect to see significant deterioration in forecast accuracy. One possible answer to the improved accuracy towards the end of the forecast horizons is that the models captured the trend and the observed series returned towards the trend in the later forecast horizons.

We do not know if the observed cases were incident or recrudescent. Inclusion of recrudescent cases in the outcome series would weaken the predictive ability of environmental covariates\textsuperscript{126}, which have a stronger relationship with incident cases, although inclusion of recrudescent cases may strengthen the predictive ability of certain treatment predictors.\textsuperscript{22,120,121} There are different ways in which measurement error could have influenced our findings. Remote sensing data was used in lieu of ground observations due to data availability, and these remote sensing observations are subject to measurement error.\textsuperscript{127-132} The treatment data were based upon prescriptions and not on dispensed antimalarial medication, which may have introduced noise into the series, and facility-level factors likely influenced the accuracy of the observed counts of confirmed malaria. Finally, we did not incorporate other predictors such as humidity and intervention data (e.g., insecticide-treated nets, indoor residual spraying), which may further improve the forecasting accuracy. The Aduku region, for example, has been subject to rounds of indoor residual spraying\textsuperscript{53}, which likely accounts for some of the unexplained variation. All of these factors have likely resulted in measurement error, increasing the noise of the different series and decreasing their ability to predict malaria.

There are different potential users of malaria forecasts. Health facilities could use the forecasts to plan for patient visits, for example, in ensuring that sufficient diagnostic and treatment materials are available. Policy makers and those involved with malaria control strategy planning could use the information to understand the burden of malaria in a particular location for the coming year, to inform the procurement of antimalarials and diagnostic equipment, and also in informing malaria control strategy such as targeting intervention efforts. With the increasing availability of electronic
medical records and electronic systems, clinical predictors could be collected and analyzed in real-time in conjunction with remote sensing data, if meteorological data are not an option. Using malaria forecasting models in practice would also allow us to understand how accurate a model needs to be, in order to be useful. Potential barriers to the utility of the models include the supply chain management approach, if supply decisions are made at the national level through a national store (‘push’ system) versus at the health-facility level as well as a lack of resources required to guide community-tailored prevention measures.

Conclusions

Clinical data such as drug treatment can be used to improve the accuracy of malaria predictions in a highly endemic setting when coupled with environmental predictors. Future research should consider other non-environmental predictor series and the practical implications of accuracy should be examined to determine the impact of forecast accuracy on disease control decisions. Accurate malaria forecasting models are needed to guide efficient allocation of resources for prevention and response; further exploration of malaria forecasting is necessary to improve its value in practice.

Funding

This research was supported by doctoral awards granted by the International Development Research Centre, Fonds de recherche santé du Québec, and McGill University Health Centre Research Institute.
### 6.3 Appendix A: Description of final models

<table>
<thead>
<tr>
<th>Location</th>
<th>ARIMA((p,d,q))</th>
<th>Predictor (lag)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aduku</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>(p=1,6,8,9,14,30,43,44; d=1, q=1)</td>
<td>total rainfall (16), rainfall range (3), maximum rainfall (3), daytime temperature (50), nighttime temperature, proportion tested (1), negative for malaria (10), appropriate treatment, ACTs, appropriate ACTs</td>
</tr>
<tr>
<td>Long-term</td>
<td>(p=1,6,43; d=1; q=1)</td>
<td>log total rainfall, total rainfall (26), daytime temperature (50), nighttime temperature (8), proportion tested (1), negative for malaria (10), appropriate treatment, appropriate ACTs</td>
</tr>
<tr>
<td><strong>Kamwezi</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>(p=4,24; d=1, q=1)</td>
<td>total rainfall (34), vegetation, negative for malaria (27), inappropriate ACTs (7)</td>
</tr>
<tr>
<td>Long-term</td>
<td>(p=4,19; d=1; q=1)</td>
<td>total rainfall (34), negative for malaria, ACTs, inappropriate chloroquine</td>
</tr>
<tr>
<td><strong>Kasambya</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>(p=1; d=1; q=1)</td>
<td>total rainfall (9), rainfall range (9), rainfall maximum (9), daytime temperature (16), nighttime temperature (42), vegetation (34), proportion tested, negative for malaria, appropriate treatment, ACTs, appropriate ACTs, appropriate quinine</td>
</tr>
<tr>
<td>Long-term</td>
<td>(p=3,4,7; d=1; q=1)</td>
<td>log total rainfall, daytime temperature (42), vegetation (36), negative for malaria</td>
</tr>
<tr>
<td><strong>Kihihi</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>(p=1,4; d=1; q=1)</td>
<td>total rainfall (18), log total rainfall, rainfall range (4), maximum rainfall (4), daytime temperature (10), nighttime temperature, vegetation (35), proportion tested, appropriate treatment, inappropriate treatment, inappropriate ACTs, chloroquine</td>
</tr>
<tr>
<td>Long-term</td>
<td>(p=4; d=1, q=1)</td>
<td>total rainfall (43), rainfall range (4), maximum rainfall (18), daytime temperature (12), nighttime temperature, proportion tested (52), appropriate treatment, ACTs, inappropriate quinine, chloroquine</td>
</tr>
<tr>
<td><strong>Nagongera</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>(p=1; d=1; q=1)</td>
<td>total rainfall, log total rainfall, nighttime temperature (26), vegetation (5), negative for malaria, inappropriate treatment</td>
</tr>
<tr>
<td>Long-term</td>
<td>(p=1,3,14,34,34; d=1, q=1)</td>
<td>log total rainfall, rainfall range, nighttime temperature (10), proportion tested, appropriate treatment, appropriate ACTs</td>
</tr>
<tr>
<td><strong>Walukuba</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>(p=10; d=1; q=1)</td>
<td>total rainfall (37), log total rainfall (39), rainfall range (37), rainfall maximum (37), daytime temperature (15), nighttime temperature (33), vegetation, proportion tested, negative for malaria, appropriate treatment, ACTs, appropriate ACTs, appropriate quinine, inappropriate quinine (43)</td>
</tr>
<tr>
<td>Long-term</td>
<td>(p=1,13; d=1; q=1)</td>
<td>daytime temperature (15), vegetation, proportion tested, negative for malaria, appropriate treatment, inappropriate treatment (52), ACTs, appropriate ACTs, quinine (43), appropriate quinine, chloroquine (43), inappropriate chloroquine (39)</td>
</tr>
</tbody>
</table>
6.4 Supplement 1: Accuracy measure comparisons

To enable comparison of forecasting errors across series and across sites, we calculated multiple accuracy measures, including the symmetric mean absolute percentage error (SMAPE), the mean absolute scaled error (MASE), and the mean absolute percentage error (MAPE).

\[
\text{SMAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|Y_t - \hat{Y}_t|}{(Y_t + \hat{Y}_t)/2}
\]
\[
\text{MASE} = \frac{1}{n-1} \sum_{t=1}^{n} |Y_t - \hat{Y}_{t-1}|
\]
\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|Y_t - \hat{Y}_t|}{Y_t}
\]

where \(Y_t\) is the observed value, \(\hat{Y}_t\) is the forecasted value, and \(Y_{t-1}\) is the observed value at lag -1. The upper limit of the SMAPE metric is restricted to 200%. The MASE compares the error of the fitted model (numerator) to the error of a naïve forecast, which is the previous observation, within the training series.\(^{102}\)

Here, the average forecasted value of each week was used as the basis for calculating the forecast error. The forecasted value across all horizons for each week was averaged:

\[
\hat{Y}_{\text{ave}} = \frac{1}{n} \sum_{w=1}^{n} \hat{Y}_w
\]

where \(\hat{Y}_{\text{ave}}\) is the averaged forecasted value by week, \(\hat{Y}_w\) is the forecasted value, \(w\) is week, and \(n\) is the number of horizons that are included in the averaging.

**Table S1 Overall errors by short-term and long-term model using average forecasted value, by UMSP site from June 1, 2012 to May 31, 2013**

<table>
<thead>
<tr>
<th>Site</th>
<th>Aduku</th>
<th>Kamwezi</th>
<th>Kasambya</th>
<th>Kiihi</th>
<th>Nagongera</th>
<th>Walukuba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall SMAPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 week</td>
<td>35.9%</td>
<td>96.1%</td>
<td>35.2%</td>
<td>23.8%</td>
<td>20.3%</td>
<td>26.9%</td>
</tr>
<tr>
<td>52 week</td>
<td>44.0%</td>
<td>90.4%</td>
<td>39.8%</td>
<td>32.0%</td>
<td>24.7%</td>
<td>30.5%</td>
</tr>
<tr>
<td>Overall MAPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 week</td>
<td>37.8%</td>
<td>*</td>
<td>127.1%</td>
<td>26.1%</td>
<td>21.7%</td>
<td>28.7%</td>
</tr>
<tr>
<td>52 week</td>
<td>43.8%</td>
<td>*</td>
<td>119.0%</td>
<td>37.4%</td>
<td>25.7%</td>
<td>35.2%</td>
</tr>
<tr>
<td>Overall MASE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 week</td>
<td>0.7</td>
<td>1.4</td>
<td>3.4</td>
<td>1.7</td>
<td>1.9</td>
<td>0.7</td>
</tr>
<tr>
<td>52 week</td>
<td>0.9</td>
<td>1.4</td>
<td>4.2</td>
<td>2.3</td>
<td>2.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

* cannot compute due to zero denominator
The SMAPE and MAPE values are very similar unless there are very small and non-zero values in the observed series that are not reflected in the forecasted series (zero values in the observed series do not permit MAPE to be calculated). For example, in Kasambya at one time point, 3 counts of malaria occur and correspond to an average forecasted value of 138 counts of malaria. This results in a very large MAPE value of 4,514% and smaller SMAPE value of 191%. Small counts can greatly influence an overall measure of accuracy for a series using MAPE. A limitation to both measures, however, is that the errors are not symmetrical for under- and over-forecasts. To overcome this limitation, we also calculated the MASE. The main limitations with this approach is that it does not allow for scale variance and the MASE also relates the forecasting error (ARIMA model) to a one-step ahead naïve error, which may not be suitable for multi-step ahead forecasts. Based on these limitations, we therefore recommend that there is more discussion and investigation of measures of forecast accuracy, which will better facilitate the comparison of model accuracy across studies.
6.5 Supplement 2: Intermediate forecasting models

Intermediate-term forecasting models (26 weeks) were developed in addition to the short-term (4 weeks) and long-term (52 weeks) models. The intermediate-term models were developed in the same fashion as the short-term models; the last 26 weeks of the training series were reserved to calculate the absolute error (the absolute difference between the cumulative total of observed cases and the cumulative total of predicted cases during that period of time). The smallest absolute error (AE) combined with the AIC was used to guide the model selection for the intermediate models.

The symmetric mean absolute percentage error (SMAPE) was calculated by forecast horizon:

$$\text{SMAPE}_{\text{aveh}} = \frac{1}{n} \sum_{h=1}^{n} \frac{|Y_h - \hat{Y}_h|}{(Y_h + \hat{Y}_h)/2}$$

where SMAPE_{aveh} is the average SMAPE value for a horizon, $Y_h$ is the observed value, $\hat{Y}_h$ is the forecasted value, $h$ is horizon, and $n$ is the number of forecasts or observations for that horizon.

The composition of the final intermediate-term models was similar to those of the short- and long-term models (Table S2-1). The intermediate-term models had the lowest errors for horizons 1 to 4 relative to the short-term and long-term models (Table S2-2). This further supports the assertion that additional exploration of model selection methods is required in forecasting studies, such as using the AIC or different accuracy measures.

Table S2-1 Predictors included in final models

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Aduku</th>
<th>Kamwezi</th>
<th>Kasambya</th>
<th>Kihii</th>
<th>Nagongera</th>
<th>Walukuba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainfall</td>
<td>✓</td>
<td>✓</td>
<td>✓ (not I)</td>
<td>✓</td>
<td>✓</td>
<td>✓ (S only)*</td>
</tr>
<tr>
<td>Temperature</td>
<td>✓</td>
<td>✓</td>
<td>✓ (not I)</td>
<td>✓</td>
<td>✓ (not I)</td>
<td>✓</td>
</tr>
<tr>
<td>Vegetation</td>
<td>✓</td>
<td>✓ (not I)</td>
<td>✓</td>
<td>✓ (S only)</td>
<td>✓ (S only)</td>
<td>✓</td>
</tr>
<tr>
<td>Treatment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Suspected</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓ (S only)</td>
<td>✓</td>
</tr>
<tr>
<td>Proportion screened</td>
<td>✓</td>
<td>✓ (S only)</td>
<td>✓</td>
<td>✓ (not S)</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Table S2-2 SMAPE for forecast horizons 1 to 4 for short-term, intermediate-term, long-term models, by UMSP site from June 1, 2012 to May 31, 2013

<table>
<thead>
<tr>
<th>Site</th>
<th>Horizon 1</th>
<th>Horizon 2</th>
<th>Horizon 3</th>
<th>Horizon 4</th>
<th>Average*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aduku</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>31.5%</td>
<td>36.9%</td>
<td>44.3%</td>
<td>47.5%</td>
<td>40.1%</td>
</tr>
<tr>
<td>Intermediate-term</td>
<td>30.5%</td>
<td>35.9%</td>
<td>39.2%</td>
<td>41.9%</td>
<td>36.9%</td>
</tr>
<tr>
<td>Long-term</td>
<td>31.6%</td>
<td>37.5%</td>
<td>40.4%</td>
<td>43.2%</td>
<td>38.2%</td>
</tr>
<tr>
<td>Kamwezi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>58.9%</td>
<td>91.5%</td>
<td>108.5%</td>
<td>122.7%</td>
<td>95.4%</td>
</tr>
<tr>
<td>Intermediate-term</td>
<td>54.3%</td>
<td>80.5%</td>
<td>99.1%</td>
<td>112.2%</td>
<td>86.5%</td>
</tr>
<tr>
<td>Long-term</td>
<td>57.8%</td>
<td>85.8%</td>
<td>107.4%</td>
<td>117.0%</td>
<td>92.0%</td>
</tr>
<tr>
<td>Kasambya</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>34.3%</td>
<td>42.4%</td>
<td>49.1%</td>
<td>47.6%</td>
<td>43.4%</td>
</tr>
<tr>
<td>Intermediate-term</td>
<td>29.4%</td>
<td>35.2%</td>
<td>38.4%</td>
<td>40.2%</td>
<td>35.8%</td>
</tr>
<tr>
<td>Long-term</td>
<td>34.8%</td>
<td>42.8%</td>
<td>47.0%</td>
<td>48.7%</td>
<td>43.3%</td>
</tr>
<tr>
<td>Kihih</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>21.7%</td>
<td>26.6%</td>
<td>29.5%</td>
<td>30.7%</td>
<td>27.1%</td>
</tr>
<tr>
<td>Intermediate-term</td>
<td>21.1%</td>
<td>25.8%</td>
<td>28.6%</td>
<td>31.2%</td>
<td>26.7%</td>
</tr>
<tr>
<td>Long-term</td>
<td>20.9%</td>
<td>25.9%</td>
<td>28.4%</td>
<td>31.1%</td>
<td>26.6%</td>
</tr>
<tr>
<td>Nagongera</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>18.8%</td>
<td>20.9%</td>
<td>22.1%</td>
<td>24.7%</td>
<td>21.6%</td>
</tr>
<tr>
<td>Intermediate-term</td>
<td>18.2%</td>
<td>20.3%</td>
<td>23.2%</td>
<td>24.6%</td>
<td>21.6%</td>
</tr>
<tr>
<td>Long-term</td>
<td>19.4%</td>
<td>23.4%</td>
<td>27.0%</td>
<td>27.8%</td>
<td>24.4%</td>
</tr>
<tr>
<td>Walukuba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short-term</td>
<td>21.9%</td>
<td>25.8%</td>
<td>29.9%</td>
<td>32.9%</td>
<td>27.6%</td>
</tr>
<tr>
<td>Intermediate-term</td>
<td>21.6%</td>
<td>25.7%</td>
<td>28.6%</td>
<td>30.2%</td>
<td>26.5%</td>
</tr>
<tr>
<td>Long-term</td>
<td>22.1%</td>
<td>25.6%</td>
<td>28.8%</td>
<td>30.7%</td>
<td>26.8%</td>
</tr>
</tbody>
</table>

*Average SMAPE for horizons 1 to 4
6.6 Supplement 3: Comparison to Holt-Winters smoothing methods

To serve as a simple comparison to the ARIMA models, forecasts were generated using the Holt-Winters exponential smoothing method:

\[ \hat{Y}_t = (a + bt)s(t) + \epsilon \]

where \( a \) is the overall mean term, \( bt \) the trend, and \( s(t) \) represents the seasonal parameter. Both seasonal and non-season models were tested using the Winters methods. Forecasts were generated at the weekly level for 52 weeks.

When comparing the ARIMA models to the Holt-Winters models, the cumulative and weekly forecasts for the first 4-week period (June 1, 2012 to June 30, 2012) and the 52-week cumulative period (June 1, 2012 to May 31, 2013) was compared between the ARIMA and Holt-Winters models (Table S3). The ARIMA models performed better than the Holt-Winters models for both the weekly and cumulative case predictions. Figure S3 illustrates the forecasts that were generated by each model and by UMSP site for June 2012. For each site, the best (most accurate) smoothing model was chosen as a comparator to the ARIMA models. Simple exponential smoothing was the most accurate approach for four of the six sites: Kasambya, Kihhi, Nagongera, Walukuba. In contrast, Aduku performed best with a Holt-Winters including a seasonal term and Kamwezi performed best with a Holt-Winters without a seasonal term.

This work demonstrates how the forecast accuracy for malaria in Uganda is better with ARIMAX models compared to simple smoothing models. Future work should compare the ARIMAX to other methods, including mathematical models and GLMs.
Table S3 SMAPE for short-term ARIMA models and 4-week smoothing models, and cumulative case error for long-term ARIMA models and 52-week smoothing models

<table>
<thead>
<tr>
<th>Site</th>
<th>ARIMA Weekly SMAPE (4 weeks)</th>
<th>Smooth ARIMA Cumulative case error (52 weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aduku</td>
<td>21.6%</td>
<td>37.2%</td>
</tr>
<tr>
<td>Kamwezi</td>
<td>24.8%</td>
<td>32.2%</td>
</tr>
<tr>
<td>Kasambya</td>
<td>24.0%</td>
<td>69.6%</td>
</tr>
<tr>
<td>Kihhi</td>
<td>13.2%</td>
<td>109.6%</td>
</tr>
<tr>
<td>Nagongera</td>
<td>6.9%</td>
<td>126.6%</td>
</tr>
<tr>
<td>Walukuba</td>
<td>16.8%</td>
<td>100.1%</td>
</tr>
</tbody>
</table>

Figure S3 Short-term ARIMA model and Holt-Winters model for June 2012, by UMSP site
CHAPTER 7: Conclusions

7.1 Summary

The goal of my thesis work was to develop evidence to guide the selection and use of malaria forecasting methods that use environmental and clinical predictors across different settings in a highly endemic country. In order to achieve this goal, the objectives were as follows: 1) to systematically examine and summarize the existing body of evidence on malaria prediction models by identifying methods of prediction and evaluation, as well as current research gaps; 2) to evaluate different methods of defining health facility catchment areas; and 3) to identify significant predictors of malaria across different settings and forecast horizons.

In the first manuscript of this thesis titled “A scoping review of malaria forecasting: past work and future directions”, we presented a systematic summary of previous malaria forecasting studies. The forecasting approaches of the 29 studies that met the inclusion criteria included statistical modeling, mathematical modeling, and machine learning methods. Climate-related predictors were used consistently in the forecasting models and model accuracy was measured in a variety of ways including mean-squared error and correlation coefficients. We could not compare the forecast accuracy of models from the different studies as the evaluation measures differed across the studies. Due to these findings, we developed recommendations for future forecasting work including:

- When performing forecasting, it is important to understand the assumptions of each method as well as the associated advantages and disadvantages.

- Using common accuracy measures is essential as they will facilitate the comparison of findings between studies and methods.

- Applying different forecasting methods to the same data and exploring the predictive ability of non-environmental variables, including interventions to reduce transmission will help determine the optimal approach and predictors for malaria forecasting.
The second manuscript, “Determining health-care facility catchment areas in Uganda using data on malaria-related visits”, was a necessary step for the forecasting work as it provided the geographic limits for the forecasting models. Due to the limitations of current catchment definition methods, I developed a new method for defining catchment areas of health facilities, which was termed the cumulative case ratio approach. My method defines a different, and I believe more accurate, catchment area for facilities when compared to the straight-line or road network approach, as it is based upon the actual residential location of patients. The commonly used distance approaches do not consider where patients live and therefore they often include parishes where there were no cases of malaria. Incorrectly identifying a catchment area can lead to inefficiencies and inadequacies of services, misspecification of the catchment population, and potentially flawed decisions regarding other facilities (e.g., placement of future facilities).

The third manuscript, “Short and long-term forecasting of malaria using environmental and clinical predictors in a highly endemic country”, identified significant environmental and clinical predictors across different settings and forecast horizons in Uganda. When examining the forecasting accuracy by horizon, forecasts for shorter horizons typically resulted in the smallest error although there were unexpected decreases in errors towards the end of the forecast horizons for all sites except for two UMSP sites. The long-term models performed well in predicting the cumulative number of cases during the forecasting period (52 weeks), with an average error of 9% across all sites. Our work determined that clinical predictors such as antimalarial treatment, the proportion of individuals screened for malaria, and the number of malaria negative individuals, produced models with the highest forecast accuracy across a range of settings and forecast horizons in Uganda when coupled with environmental predictors. These predictors are likely relevant predictors in other malaria endemic countries. The manuscript also touched upon model selection issues, as outlined in
the first manuscript of this thesis, such as using the AIC or forecast accuracy to guide model development.

7.2 Implications for research and practice

The work of this thesis resulted in the development the cumulative case ratio approach to defining a health facility’s catchment area. This approach is simple and very relevant in resource-limited settings where information on the attendance of all health facilities may be limited.

The scoping review included recommendations for future forecasting work that, if followed, should lead to improved methods and increased forecast accuracy for malaria but also for other infectious diseases. The forecasting work demonstrated the importance of considering clinical predictors, such as antimalarial treatment, when developing models to forecast malaria. Future research should consider other non-environmental predictor series and using different forecasting methods on the same data. Also, there needs to be further investigation of model selection methods – comparing the AIC approach to an accuracy approach (based on a subset of the training data or using cross-validation) when choosing the predictors. Moreover, there should be further discussion about what accuracy measure(s) should be used and how they should be applied when reporting the results of forecasting work.

There are different potential users of malaria forecasts. Health facilities could use the forecasts to plan for patient visits, for example, in ensuring that sufficient diagnostic and treatment materials are available. Policy makers and those involved with malaria control strategy planning could use the information to understand the burden of malaria in a particular location for the coming year, to inform the procurement of antimalarials and diagnostic equipment, and also to inform malaria control strategy, such as targeting intervention efforts. Increasing the use of malaria forecasting models in practice would also allow evaluation of the practical implications of accuracy, such as understanding how accurate a model needs to be in order to be useful in practice. Also, working
with organizations that are responsible for malaria control interventions would also provide valuable information in terms of model parameters. Meaning, what temporal and spatial resolutions are most useful to inform malaria control strategies and/or interventions? How far ahead should malaria be forecasted in order to maximize its utility? What are the barriers to using a forecasting model in practice? Understanding these conditions would provide a framework for the forecasting model and to help ensure its value in practice.

7.3 Conclusions

The Abuja Declaration noted the importance of accurate disease prediction for targeting and evaluating control measures. The collective work of this thesis should advance the field of malaria forecasting in a number of ways: in providing methodological guidelines for future forecasting studies, in providing a simple method for catchment definition that can be applied to define the geographic limits of a forecasting model, in demonstrating the importance of clinical predictors for forecasting malaria, in providing an example of forecasting models with high spatial and temporal resolutions, and in raising important points of consideration for future forecasting work.
# Appendix A Case report form

## Patient Record Form

<table>
<thead>
<tr>
<th>Date</th>
<th>OPD Number</th>
<th>Patient’s Last Name</th>
<th>First Name</th>
<th>New attendance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Parties**
  - Village
  - Age: __________
  - Sex: Male / Female
- **History & Exam Findings**
  - Fever or chills: Yes / No
  - History of cough: Yes / No

<table>
<thead>
<tr>
<th>Blood group</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>O</th>
<th>Rh positive</th>
<th>Rh negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV status</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

- **Diagnosis**
  - SIRS
- **Urea**

## Diagnosis (Check all that apply)

<table>
<thead>
<tr>
<th>Category</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infectious Disease</td>
<td>Typhoid Fever</td>
</tr>
<tr>
<td>Infections</td>
<td>Bacterial Infections</td>
</tr>
<tr>
<td>Cancer</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>Other</td>
<td>Other</td>
</tr>
</tbody>
</table>

## Treatment (Check all that apply)

<table>
<thead>
<tr>
<th>Drug</th>
<th>Dose</th>
<th>Drug</th>
<th>Dose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibiotics</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antimalarials</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antivirals</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antifungals</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antituberculosis</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antineoplastic</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antiepileptics</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anticonvulsants</td>
<td>Other Drugs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Referrals and additional notes

- **TB Drug Regimen**
  - Initial Phase
  - Second Line
  - Other services
Appendix B Observed counts of malaria by UMSP facility
Appendix C Interpolated and observed environmental predictor series

Figure C-1 Vegetation index interpolated and observed series by UMSP site
Figure C-2 Daytime temperature (in Celisus) interpolated and observed series by UMSP site
Figure C-3 Nighttime temperature (in Celsius) interpolated and observed series by UMSP site
Appendix D Variability of forecasted malaria counts by week (pooled across horizon)
Appendix E Variability of forecasted error by horizon
Appendix F SMAPE figures for short-term models

Figure F-1 Weekly observed counts sized by SMAPE error for forecast horizon 2 of the short-term models, by UMSP site from June 1, 2012 to May 31, 2013
Figure F-2 Weekly observed counts sized by SMAPE error for forecast horizon 3 of the short-term models, by UMSP site from June 1, 2012 to May 31, 2013
Figure F-3 Weekly observed counts sized by SMAPE error for forecast horizon 4 of the short-term models, by UMSP site from June 1, 2012 to May 31, 2013
References


36 Lindsay SW, Martens WJ. Malaria in the African highlands: past, present and future. *Bull World Health Organ* 1998; **76**: 33–45.


50 Lengeler C. Insecticide-treated bed nets and curtains for preventing malaria. *Cochrane Database Syst Rev* 2004; : CD000363.

51 Balayo C. Overview of the implementation of the insecticide treated nets (ITNs) programme in Uganda. Ministry of Health, 2005.


Chatfield C. The analysis of time Series: an introduction, 6 ed. London, Chapman & Hall,
2004.


100 Shmueli G. To explain or to predict? Statist Sci 2010; 25: 289–310.


106 Gilmour SJ. Identification of hospital catchment areas using clustering: an example from the NHS. *Health Serv Res* 2010; **45**: 497–513.

107 Diesfeld HJ. The definition of the hospital catchment area and its population as a denominator for the evaluation of hospital returns in developing countries. *Int J Epidemiol* 1973; **2**: 47–53.


110 Boscoe FP, Henry KA, Zdeb MS. A nationwide comparison of driving distance versus straight-line distance to hospitals. *Prof Geogr* 2012; **64**: 188–96.


115 Williams F, Boren SA. The role of the electronic medical record (EMR) in care delivery development in developing countries: a systematic review. *Inform Prim Care* 2008; **16**: 139–45.


117 Blaya JA, Fraser HSF, Holt B. E-health technologies show promise in developing countries. *Health Aff* 2010; **29**: 244–51.


119 Chatfield C. The analysis of time series: an introduction.


130 Huete AR. Vegetation indices, remote sensing and forest monitoring. *Geography Compass* 2012; **6**: 513–32.


132 Adeyewa ZD, Nakamura K. Validation of TRMM radar rainfall data over major climatic regions in Africa. *J Appl Meteorol* 2003; **42**: 331–47.

