STATISTICAL MODELLING APPROACHES FOR LONGITUDINAL MULTIPLE OUTCOME DATA FROM IMMUNO-EPIDEMIOLOGICAL STUDIES IN ENTEBBE, UGANDA

Lawrence Lubyayi

A Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of

Doctor of Philosophy

AUGUST 2021

Division of Epidemiology and Biostatistics

School of Public Health, Faculty of Health Sciences

University of the Witwatersrand,

Johannesburg, South Africa

Funded by

The DELTAS Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB)

Research group affiliation

Immunomodulation and Vaccines Programme, MRC/UVRI and LSHTM Uganda Research Unit

Declaration

I, Lawrence Lubyayi declare that this thesis is my original work and has not been submitted before for any degree or award at this or any other university. Where information has been derived from other sources, I confirm that this has been duly acknowledged. Parts of this work have been published as open access articles, I (and my co-authors) retained the copyright for that work.

Lawrence Lubyayi

30th August 2021

Dedicated to my wife Esther, our children, the Lubyayi family (Dad, Mum, Martin, Fred, Deo, Fr. Tom, Grace, Flavia) and to the Sempuma family (Jjajja Margaret, Uncle Godfrey, Auntie Charity, Nami, Asanka and Paul)

Original papers

1. **Lubyayi L**, Mawa PA, Nabakooza G, Nakibuule M, Tushabe JV, Serubanja J, Aibo D, Akurut H, Tumusiime J, Hasso-Agopsowicz M, Kaleebu P, Levin J, Dockrell HM, Smith S, Webb EL, Elliott AM, Cose S. **Maternal latent *Mycobacterium tuberculosis* does not affect the infant immune response following BCG at birth: An observational longitudinal study in Uganda.** Frontiers in Immunology. 2020 May 14; **11**(929). DOI: https://doi.org/10.3389/fimmu.2020.00929

   My contribution to the paper

   Structuring of the paper, planning and running all statistical analyses, writing all drafts of the paper, addressing co-authors' and peer reviewers' comments and checking proofs for the final accepted and published manuscript.

2. **Lubyayi L**, Mawa PA, Cose S, Elliott AM, Levin J, Webb EL. **Analysis of multivariate longitudinal immuno-epidemiological data using a pairwise joint modelling approach.** (Under review at BMC Immunology).

   My contribution to the paper

   Conceptualisation of research idea, structuring of the paper, planning and running all statistical analyses, writing all drafts of the paper, addressing co-authors' comments.

3. **Lubyayi L**, Mpairwe H, Nkurunungi G, Lule SA, Nalwoga A, Webb EL, Levin J, Elliott AM. **Early childhood exposure to infections is associated with reduced risk of allergy related disease outcomes in later childhood: results from a Ugandan birth cohort.** (Under review at the eLife journal).

   My contribution to the paper

   Conceptualisation of research idea, structuring of the paper, planning and running all statistical analyses, writing all drafts of the paper, addressing co-authors' comments.

Conference proceedings

1. **Lubyayi L**, Mawa PA, Nabakooza G, Nakibuule M, Serubanja J, Aibo D, Dockrell HM, Levin J, Webb EL, Elliott AM, Cose S. **Immune sensitisation and evolution of cytokine responses in infants of mothers with and without latent *Mycobacterium tuberculosis* infection.** Keystone Symposium on Tuberculosis: Mechanisms, Pathogenesis and Treatment, Banff, Alberta, Canada, January 17-21, 2019 (short talk and poster presentation).
2. **Lubyayi L**, Mawa PA, Nabakooza G, Nakibuule M, Serubanja J, Aibo D, Dockrell HM, Levin J, Webb EL, Elliott AM, Cose S. **The impact of maternal infection with latent *Mycobacterium tuberculosis* on infant immune sensitisation and evolution of immune responses following immunisation with BCG at birth.** DELTAS Africa Scientific Conference, Dakar, Senegal, July 15-17, 2019 (oral presentation).
3. **Lubyayi L**, Mawa PA, Nabakooza G, Nakibuule M, Tushabe JV, Serubanja J, Aibo D, Akurut H, Tumusiime J, Hasso-Agopsowicz M, Kaleebu P, Levin J, Dockrell HM, Smith S, Webb EL, Elliott AM, Cose S. **Maternal latent *Mycobacterium tuberculosis* has no impact on immune responses following BCG at birth: results from longitudinal studies in Uganda and the United Kingdom.** Infectious Diseases in Africa Symposium, Cape Town, South Africa, October 7-12, 2019 (poster presentation).
4. **Lubyayi L**, Mawa PA, Cose S, Elliott AM, Levin J, Webb EL. **Analysis of multivariate longitudinal immuno-epidemiological data using a pairwise joint modelling approach.** 30th International Biometric Conference (IBC 2020), Virtual, July - August, 2020 (oral presentation).
5. **Lubyayi L**, Mpairwe H, Nkurunungi G, Lule SA, Webb EL, Levin J, Elliott AM. **A latent class analysis of early childhood infection-exposure and its association with allergy-related diseases in later childhood: data from the Entebbe Mother and Baby Study.** 2020 DELTAS Africa SSACAB Virtual Scientific Conference, 19th – 23rd October, 2020 (oral presentation).

Abstract

Background

Longitudinal studies are indispensable for studying changes in variables over time, and are increasingly common in many fields of research. However, in comparison with other study designs, longitudinal studies have some practical challenges including increased time requirements and associated costs. Many challenges arise when analysing data from longitudinal studies including multivariate measures with complex correlation and random error structures, time-varying covariates and missing data. Such challenges increase the complexity of longitudinal data analysis, and this is particularly the case for immuno-epidemiological studies.

Immuno-epidemiological studies investigate the influence of immune responses on the epidemiology of conditions such as infectious diseases, cancer, hypersensitivity and autoimmunity. These studies have increased recently as a result of technological developments in sample processing and analyte measurement allowing for the simultaneous measurement of many immunological parameters, generating data for the exploration of both simple and complex associations between disease, immunity, social, environmental, genetic or other factors. Consequently, the complexity of associations between several immunological parameters poses a major challenge of how to select the most appropriate statistical approach to extract the maximum relevant information from such complex datasets whilst avoiding spurious findings.

A small number of journal articles have given an overview of application of statistical techniques to immuno-epidemiological data, however, these focus mainly on cross-sectional data. Published guidance on the analysis of longitudinal immunological data remains limited, particularly for

longitudinal data with multivariate outcomes. This PhD project therefore aimed to investigate and extend appropriate statistical modelling methods for longitudinal data, and to apply them to longitudinal multiple-outcome data from immuno-epidemiological studies conducted in Entebbe, Uganda.

Methods

This PhD study utilised secondary data from two immuno-epidemiological cohort studies, carried out in Entebbe, Uganda. The Infant BCG study (IBS), an observational longitudinal study, aimed to examine the timing, magnitude and quality of the initial response to BCG immunisation and to determine the effect of prenatal exposure to maternal latent *Mycobacterium tuberculosis* infection (LTBI) on these outcomes. The Entebbe Mother and Baby Study (EMaBS), a birth cohort study, aimed to investigate the impact of worms and their treatment on responses to immunisation, illness and allergy-related diseases (ARDs).

For the IBS, latent variable modelling approaches involving dimension reduction including principal component analysis (PCA) and three-way component analysis, in combination with mixed effects models, were employed to study the evolution of multivariate cytokine responses over time and how that evolution depends on infant characteristics such as maternal LTBI status and age, sex, birthweight or other covariates. For the EMaBS, latent class analysis (LCA) was employed to characterise the infection experience of study participants and how this relates to susceptibility to ARDs at 9 years. To handle missing data, multiple imputation and direct likelihood approaches were incorporated in the analyses for both the IBS and EMaBS.

Results

The IBS showed that there was remarkably high early sensitisation to mycobacterial antigens *in utero*, in Uganda, but this sensitisation had no impact on the infant response to immunisation with BCG. PCA in cord blood showed IL-10 and TNF tending to group separately from the other cytokines, but there was no evidence that maternal LTBI was associated with a differing pattern of response, or that differences in these cord blood profiles impacted the subsequent infant response to BCG.

The pairwise joint modelling approach was applied to the IBS data in a situation where fitting of a full multivariate mixed model had failed due to model complexity. Parameter estimates from the pairwise approach had better precision than those from the univariate mixed models. Initial patterns from PCA of cord blood responses were not reflected in the profile of responses that developed post-BCG as indicated by the PCA on the correlation matrix of random intercepts. This improved our interpretation of the data by suggesting that probably the effect of BCG immunisation was potent enough to over-ride patterns established *in utero*.

LCA of data from the EMaBS identified two distinguishable groups of children, during each of the first five years of life, based on their early infection-exposure. In the first year of life, latent class (LC) 1 was characterised by higher probabilities of malaria, diarrhoea and LRTI, and membership of this class was associated with lower proportions of wheeze, eczema, rhinitis and SPT positivity for *Dermatophagoides* at 9 years. Associations between LC membership and ARDs were less consistent for infections in years two to five, suggesting that infancy may be a critical period during which the risk of ARDs in later childhood is established.

Conclusions

This PhD study adapts and applies recently developed robust statistical analysis methods in novel analyses of longitudinal immuno-epidemiological data with case studies from Entebbe, Uganda. Notably, this work shows that maternal latent *M.tb* infection is not the reason for reduced effectiveness of BCG in tropical countries and therefore, all infants are likely to benefit from BCG immunisation regardless of their maternal LTBI status. This work also shows that exposure to common infections, especially during the first year of life, was associated with lower proportions of ARDs in later childhood. The statistical analysis approaches used, including the pairwise joint modelling approach and latent class analysis, improve our understanding and interpretation of longitudinal immuno-epidemiological data.

Preface

This thesis is submitted to the University of the Witwatersrand, Faculty of Health Sciences, School of Public Health, in fulfilment of the requirements for the degree of Doctor of Philosophy. Doctoral work has been performed under the supervision of Professor Jonathan Levin from the Division of Epidemiology and Biostatistics, with co-supervision from Dr Emily L. Webb from the MRC International Statistics and Epidemiology Group, Department of Infectious Disease Epidemiology at the London School of Hygiene and Tropical Medicine, United Kingdom and Professor Alison M. Elliott from the Immunomodulation and Vaccines Program at the MRC/UVRI and LSHTM Uganda Research Unit.

Financial assistance was through a PhD fellowship from the DELTAS Africa Initiative SSACAB (grant number 107754). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (grant number 107754/Z/15/Z) and the UK Government. The views expressed in this study are those of the authors and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust or the UK government.

Acknowledgements

I am forever grateful to my supervisors for the success of this PhD journey. It all began when I joined the MRC/UVRI and LSHTM Uganda Research Unit as a Statistician, specifically to work on projects under the Immunomodulation and Vaccines (IVac) Program headed by Professor Alison Elliott. In one of our initial discussions about ongoing studies and available datasets, upon my proposition for longitudinal data analysis approaches for some of the cohort studies, Prof. Elliott asked whether I was interested in pursuing those analyses as part of my PhD and thus the journey started, many thanks Prof. Elliott. I am also profoundly honoured to have been mentored by Dr Emily Webb before and during my PhD journey, I will forever revere your technical refinement of my work and selfless guidance throughout. I am also very grateful to Prof. Jonathan Levin who unwaveringly accepted to supervise me and not only provided technical guidance but also made sure that my stay in South Africa was always comfortable.

I am very grateful to all the study participants, field, clinic and laboratory staff of the Infant BCG Study (IBS) and the Entebbe Mother and Baby Study (EMaBS) under the IVac Program at the MRC/UVRI and LSHTM Uganda Research Unit. Special thanks to Dr Steve Cose, Prof. Alison Elliott, Prof. Pontiano Kaleebu, Dr Emily Webb, Dr Steve Smith and Prof. Hazel Dockrell who conceived the IBS and secured funding. Many thanks to Dr Patrice Mawa who was a key investigator and a major resource in terms of the immunological aspects of the IBS. Added thanks to Prof. Elliott, Dr Harriet Mpairwe and the various project leaders and staff who have overseen the EMaBS since 2003 to date. I also benefitted greatly from the peer support structure of the IVac program, the

weekly journal clubs and monthly program meetings, all of which provided avenues for me to discuss my PhD results and to obtain invaluable reviews.

I thank my PhD sponsors, the DELTAS Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB), led by Prof. Tobias Chirwa and other co-leads. Without this funding, it would have taken much longer for me to realise this PhD dream. In addition to the PhD fellowship, the SSACAB secretariat organised annual conferences which involved knowledge sharing, review and feedback from senior mentors. To this extent, I am very thankful to Prof. Jim Todd, Prof. Eustasius Musenge, Prof. Samuel Manda and Prof. Henry Mwambi who reviewed and provided a lot of constructive feedback along my PhD journey.

Many thanks to the University of the Witwatersrand School of Public Health, the head of the PhD programme Prof. Kathleen Kahn, the coordinator Dr Jude Igumbor and the administrator Mr. Paul Bohloko, for organising protocol presentations, interim seminars and several workshops to facilitate a smooth PhD journey.

Lastly, I acknowledge the unwavering support of my family throughout my studies to date. To my lovely wife Esther, who had to endure several prolonged spells of my absence from home and the hours on the computer while at home, thanks for your patience and understanding. My parents, Mr. and Mrs. Lubyayi, for supporting and encouraging me all the way, I am forever grateful. My siblings Martin, Fred, Deo, Fr. Tom, Grace and Flavia, many thanks for encouraging me and for ensuring that there was some fun along the way. I cannot forget my sisters-in-law (Rose, Barbara and Leocadia) and my nieces (Martina, Martha, Paula and Eleanor) for cheering me on. The Sempuma family (Jjajja Margaret, Uncle Godfrey, Aunt Charity, Nami, Asanka and

Paul), thanks a lot for sharing your space with me and for making me feel at home in a foreign land (South Africa) throughout my studies, I greatly owe my success to you.

Above all, I praise and thank God for my miraculous journey thus far.

Contents

List of Tables

1. List of tables as presented in Research Papers

2. List of tables in other parts of the thesis

Title	Page

List of Figures

1. List of figures as presented in Research Papers

2. List of figures in other parts of the thesis

Title	Page

Acronyms / Abbreviations

ALS	Alternating Least Squares
ARDs	Allergy-related diseases
asIgE	Allergen-specific Immunoglobulin E
AUC	Area under the curve
BCG	Bacille Calmette Guerin
CFP10	Culture Filtrate Protein 10
CMV	Cytomegalovirus
CP	Candecomp/Parafac
EBV	Epstein Barr Virus
EMaBS	Entebbe Mother and Baby Study
ESAT6	Early Secretory Antigenic Target 6
GPA	Generalised Procrustes Analysis
HICs	High-income countries
HSV	Herpes Simplex Virus
IBS	Infant BCG study
IL	Interleukin
IFN	Interferon
LCA	Latent Class Analysis
LRTI	Lower Respiratory Tract Infections
LSHTM	London School of Hygiene and Tropical Medicine
LTA	Latent Transition Analysis
LTBI	Latent *Mycobacterium tuberculosis* infection
MAR	Missing at random
MCAR	Missing completely at random
M.tb	*Mycobacterium tuberculosis*
MI	Multiple Imputation

ML	Maximum Likelihood
MNAR	Missing not at random
MRC	Medical Research Council
PCA	Principal component analysis
PPD	Purified Protein Derivative
REML	Restricted Maximum Likelihood
SPT	Skin Prick Test
SVD	Singular Value Decomposition
T3	Tucker3
Th1	T helper 1
Th2	T helper 2
TNF	Tumor necrosis factor
URTI	Upper Respiratory Tract Infections
UVRI	Uganda Virus Research Institute

Chapter 1. Background

1.1. Longitudinal studies and immuno-epidemiology

Longitudinal studies are common in many fields of research, including health, biology, agriculture, education, marketing, social and behavioural sciences. They are indispensable when studying changes in variables over time. Through making measurements on study participants on many occasions over time, longitudinal studies provide for the direct analysis of changes within individual participants and the factors that affect these changes[1]. Since the study of changes in variables over time is relevant to almost every discipline, there has been a steady increase in the number of studies using longitudinal designs[1]. However, in comparison with other study designs such as cross-sectional studies, longitudinal studies are subject to some practical challenges, for both participants and researchers, including increased time and expenses, the potential for larger sample size requirements, and attrition.

Many challenges arise when analysing data from longitudinal studies[2]. Naturally, the repeated observations or measures arising from such studies are multi-dimensional and have complex correlation and random-error structures that should be properly accounted for during analysis[1]. These studies also allow for repeated observations of both outcomes and covariates[1]. The other issue which complicates analyses, in such studies, is the presence of time-varying covariates, which pose many intricate analytic issues[3-5]. Missing data and attrition are also common occurrences in these studies; yet proper handling of missing data remains a major challenge during analysis of longitudinal data[1, 6, 7]. Such challenges, amongst many others, increase the complexity of longitudinal data analysis, and this is particularly the case for immuno-epidemiological studies.

Immuno-epidemiological studies investigate the influence of population immunity on the epidemiology of conditions such as infectious diseases, cancer, hypersensitivity and autoimmunity[8, 9]. Such studies are likely to have a large number of, often correlated, outcomes measured repeatedly over time. Studies on immunological mechanisms affecting human disease and the identification of related immunological markers have increased in the recent past[10]. Recent technological developments in sample processing and analyte measurement allow for the simultaneous measurement of many immunological parameters, generating data which enable the exploration of both simple and complex associations between disease, immunity, social, environmental, genetic and many other factors[10]. A case in point is the multiplex assay which simultaneously measures dozens of analytes, or more, in a single run, resulting in a large amount of data being generated simultaneously. As a result, the complexity of the associations between several immunological parameters poses a major challenge: how to select the most appropriate statistical approach to extract the maximum relevant information from such complex datasets whilst avoiding spurious findings[10].

In immuno-epidemiological studies, a number of scientific questions might be of interest, for instance to determine the effect of an intervention or exposure on the joint evolution of all outcomes together, or to categorise outcomes that change in the same way. Modelling the different outcomes separately may not be appropriate because it ignores the underlying relationships between them and the relationships between outcomes may change over time. In such situations, a joint modelling strategy is necessary[11].

1.2. Literature review: analysis approaches for immuno-epidemiological data

In many cross-sectional immuno-epidemiological studies, simple statistical approaches are applied frequently even when complex patterns of inter-relationships between parameters are anticipated, instead of more appropriate multivariate approaches which can analyse multiple measurements on an individual simultaneously[10, 12]. For instance, Härtel et al. used a simple bivariate analysis approach, non-parametric correlation, to investigate associations between age and cytokine expression (IFN-γ, TNF-α, IL-2, IL-4, IL-5, IL-10, IL-12)[13]. Geiger et al. used the non-parametric Wilcoxon rank test and the non-linear Spearman's rank test, with a Bonferroni-Holm adaptation for multiple testing, to study differences in cytokine production (IL-12, IFN-γ, IL-5, IL-10, IL-13 and TNF-α) between hookworm infected and non-infected endemic controls[14]. Tanaka et al. used the Student's paired t-test to study the role of stimulator/responder ratio on the ability of human monocyte-derived dendritic cells to promote T helper type 1 (Th1) or Th2 differentiation[15]. Rachmiel et al. used Wilcoxon's signed rank test to compare ratios of distinct immune responses (Th1/Th2) between patients with both type 1 diabetes mellitus (T1DM) and asthma with those who had only T1DM or only asthma[16]. Tisato et al. analysed each of 29 markers separately by comparing pre- and post-surgery cytokine levels in patients who had chronic venous disease, using the Student's t-test and the Mann-Whitney rank-sum test, without adjustment for multiple testing[17]. Other studies[18-20] have used the approach of including several markers, as multivariate exposures in multivariable regression models, neglecting any a priori knowledge about possible underlying biological mechanisms.

For longitudinal immuno-epidemiological data, simple analysis methods remain in common use. Several articles report the use of separate analysis at each time point, for instance, Soares et al.

analysed longitudinal changes (separately at each time point) in CD4+ T-cell memory responses of new-borns[21], Lalor et al.[22] and Mawa et al.[23] analysed longitudinal infant responses to BCG separately at each time point. Nematollahi et al. reported the use of a summary measure, area under the curve (AUC), to describe the levels of pro-inflammatory cytokines in response to insulin induced hypoglycemic stress in healthy subjects[24], while Lalor et al. used nonparametric Mann-Whitney tests to compare fold differences between UK and Malawian cytokine responses[22]. The major disadvantages of these simple analysis approaches are that a lot of information is lost and they do not allow for drawing of conclusions about the way the endpoint has been reached. As such, important patterns or associations over time might be missed. In addition, there are further technical challenges, for example, AUC summary measures have been shown to be biased in the presence of missing data[25]. The separate analysis of longitudinal data at each time point has been referred to as invalid[26] since it assumes independence of measurements at each time point hence losing the advantages of undertaking a longitudinal study.

A number of articles have given an overview of application of statistical techniques to cross-sectional immuno-epidemiological data[10, 12, 27, 28]. Genser et al. provided a guide for the simultaneous analysis of multiple study variables using conceptual frameworks, latent variables and multivariate techniques, reflecting on the complexity of modern immunological research study questions[10]. Genser et al. also provided a systematic analytical approach for the statistical analysis of multiple correlated immune markers and highlighted a framework for incorporating prior knowledge about underlying immunological mechanisms[12]. McGuinness et al. discussed the robustness and limitations of normal parametric methods for the analysis of highly skewed immune response data and they explored the use of bootstrap resampling techniques for

checking normal analysis and also as a direct method of analysis[27]. Bennett and Riley offered some guidelines on the analysis of immuno-epidemiological data, focussing on the appropriateness of logarithmic transformations, use of the stimulation index and the use of the proportion of responders for comparing groups of individuals[28]. These articles focus mainly on cross-sectional data. Published guidance on the analysis of longitudinal immunological data is limited, for instance, Panda et al. discussed a mixed models approach for repeated measures of immunological data[29], however, they focus on only one longitudinal outcome. A gap remains for guidance on the analysis of longitudinal immunological data with multivariate outcomes or exposures.

A large body of literature (both books and journal articles) have been published, describing the use of multivariate techniques for statistical analysis in other scientific fields, including for multivariate longitudinal data[1, 6, 11, 30-36]. For instance, Fitzmaurice et al.[1] and Diggle et al.[6] provided comprehensive details of how to analyse longitudinal data, using linear mixed effects models; Verbeke et al.[11] and Fieuws et al.[35] provided reviews for the analysis of multivariate longitudinal data and discussed the pairwise fitting of joint models for multivariate longitudinal profiles; Oort[30, 31], Timmerman and Kiers[34] and Kroonenberg[36] published about multiway data analysis including the use of three-way component models; while Lanza et al.[37], Goodman[38] and Collins[39] described applications of latent class and latent transition analysis models for multivariate exposure data. Unfortunately, for the applied immunologists who may not have detailed knowledge of methodological statistics, this published literature may not be easily accessible[10] and might be challenging to apply to their specific datasets. Despite the extensive literature cited in this paragraph, none of the articles includes any applications to immuno-

epidemiological data. This PhD study aimed to address this problem and to illustrate application of some of the advanced statistical analysis techniques to multivariate longitudinal immuno-epidemiological data.

1.3. Immuno-epidemiological cohort studies in Entebbe

The studies described in this sub-section yielded the datasets which were used to explore the proposed statistical analysis approaches for longitudinal multiple-outcome data and to investigate how best these methods can be used to answer the corresponding scientific questions.

The Infant BCG study (IBS) was a birth cohort, based at Entebbe General Hospital, Uganda, which aimed to examine the timing, magnitude and quality of the initial response to BCG immunisation and determine the effect of prenatal exposure to maternal latent *Mycobacterium tuberculosis* infection (LTBI)[40]. In this study, 132 LTBI positive and 150 LTBI negative mothers and their babies were enrolled. Infants were followed up to age one year, with blood sampled at multiple time points, in order to assess the establishment of BCG-specific immunological memory. Recruitment of participants commenced on 16th June 2014 and follow-up ended on 30th October 2016. Infants were randomly assigned to two separate blood sampling strategies: 150 infants (75 from mothers with LTBI, and 75 from mothers without LTBI) to give blood at 1, 6, 14 and 52 weeks while the remaining infants were randomly assigned to give blood at 4, 10, 24 and 52 weeks. Immunological parameters, including 17 representative cytokine/chemokine responses, and antibody responses (total immunoglobulin (Ig)G) to 4 vaccine-specific antigens, were assessed at each of these time points. The IBS sought to understand the dynamics of vaccine responses (to BCG), how they differ

between infants born of mothers with or without LTBI and their relationship to postnatal acquisition of infections such as cytomegalovirus (CMV), Epstein Barr Virus (EBV) and malaria. The statistical challenge in the IBS study was how to analyse the multiple measurements of the same parameters at each of seven time-points in an appropriate way and be able to discern overall differences between infants born of mothers with or without LTBI.

The Entebbe Mother and Baby Study (EMaBS), based at Entebbe General Hospital, Uganda, was a trial (with enrolment running from 2003 to 2005) of three randomized, double-blind, placebo-controlled interventions at two times, in a 2x2(x2) factorial design: women were randomized to single-dose albendazole versus placebo and single-dose praziquantel versus placebo during their second or third trimester of pregnancy; their children were randomized to quarterly albendazole versus placebo from age 15 months to 5 years[41, 42]. Trial primary and secondary outcomes were both immunological and clinical, including vaccine responses, infectious diseases and allergy-related diseases (ARDs) initially assessed at various time points up to five years of age. These children have continued to be followed-up annually and through visits to the study clinic when ill, and had detailed data on ARDs collected at age nine years. The EMaBS recruited 2507 women during pregnancy [42], leading to 2345 live births. Of the 2345 live births, 1214 (52%) children were seen at 9 years[43]. Similar to findings from high-income countries (HICs)[44], we have previously shown that allergy-related disease (such as eczema, wheeze and rhinitis) declines with age, despite increasing atopy (as assessed by skin prick positivity to common allergens)[43]. An outstanding question for the EMaBS was how to use the array of longitudinal illness and infection data available to explore the “hygiene hypothesis” further. The hygiene hypothesis states that a lack of early childhood exposure to infectious agents, symbiotic micro-organisms and parasites

increases susceptibility to allergic disease by modulating the natural development of the immune system[45-47]. We therefore sought to explore the role of common early childhood infections, and of the age at which they occurred, in the development of atopy and allergy-related diseases, using appropriate statistical modelling methods.

1.4. Problem Statement

The studies described above provided a precious resource for exploring statistical analysis approaches for studies of a longitudinal nature, and for investigation of best practice in handling multiplicity of immunological outcome data. Various analyses, predominantly cross-sectional, have been conducted to address the objectives of these studies, but the longitudinal nature of much of the outcome data has received less attention. The studies above exhibit complex longitudinal designs of unbalanced nature, due to commonly occurring phenomena in longitudinal studies, such as intermittent missingness or attrition. Unbalancedness can also occur by design, for instance in the IBS where participants were randomised to provide blood samples following two different schedules. Multivariate responses and time-varying covariates are the other characteristic features of these data.

Some articles exist in the literature, giving an overview of application of statistical techniques to immuno-epidemiological data[10, 12, 27-29], however these focus mainly on cross-sectional data. Published guidance on the analysis of longitudinal immunological data is limited[28], particularly for longitudinal data with multivariate outcomes. This PhD project aimed to address this gap.

1.5. Research Objectives

The aim of this research was to investigate and extend appropriate statistical modelling methods for longitudinal data, and to apply them to longitudinal multiple-outcome data from immuno-epidemiological studies in Entebbe, Uganda.

Specifically, the research aimed to:

1. Use latent variables in combination with longitudinal analysis approaches to understand the dynamics of vaccine responses to BCG and how they differ between infants born of mothers with and without LTBI, in the Infant BCG Study.
2. Use longitudinal data and latent classes of participants to investigate which infections, and at which times, affect atopy and allergy-related disease outcomes during childhood, in the Entebbe Mother and Baby Study.
3. Consider statistical approaches for dealing with missing data from longitudinal multiple-outcome immuno-epidemiological studies.

1.6. Conceptual framework

Figure 1 gives an overview of the approaches for the specific objectives of the study. It highlights the multivariate longitudinal outcomes for the Infant BCG Study and shows that latent variable modelling approaches which involve dimension reduction in combination with mixed effects models can be used to study the evolution of cytokine responses over time and how that evolution depends on infant characteristics such as maternal LTBI status and age, sex, birthweight or other covariates. It also highlights the longitudinal exposure variables from the EMaBS and shows that latent class analysis can be employed to characterise the infection status of study

participants and how this relates to susceptibility to ARDs at 9 years. Objective 3 involved consideration of missing data approaches for both studies and these are incorporated in the analysis approaches for both the IBS and EMaBS.

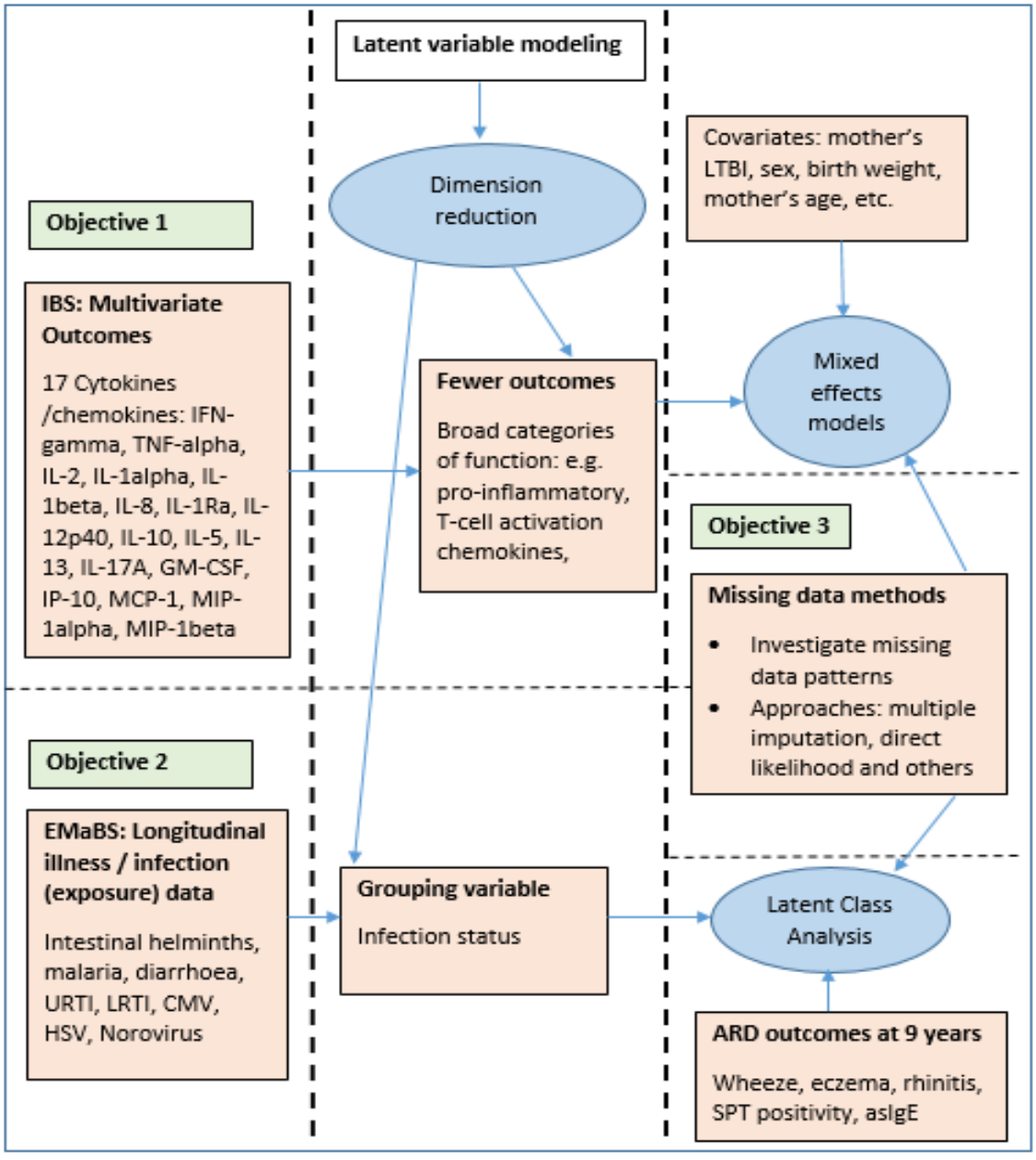


Figure 1: Statistical analysis approaches for longitudinal multiple-outcome/exposure data from immuno-epidemiological studies

1.7. Thesis structure

Chapter 1 is the background of the thesis. It highlights the growing use of multivariate longitudinal studies in the field of immuno-epidemiology, and the analytical challenges involved with such studies. It includes a literature review, gives an overview of the common analysis approaches for immuno-epidemiological data and identifies some gaps in the literature especially for published guidance on the analysis of multivariate longitudinal immunological data. The chapter then describes two immuno-epidemiological case studies from Entebbe, Uganda, highlights the problem statement, research objectives, conceptual framework and ends with a description of the layout of the entire thesis.

Chapter 2 describes the general methodology including the study setting, data sources, statistical methods and ethical approvals. Missing data approaches (thesis objective 3) are discussed at length in line with their application to both the IBS and EMaBS datasets. Certain elements of the general methodology are duplicated in subsequent chapters because the research papers comprised in this thesis were written as stand-alone manuscripts.

Chapter 3 (Research paper 1) presents results for the first objective of this PhD study. It highlights the use of longitudinal data analysis approaches to understand the dynamics of vaccine responses and how they differ between infants born of mothers with or without LTBI, in the Infant BCG study.

Chapter 4 (Research paper 2) describes a pairwise joint modelling approach for the analysis of multivariate longitudinal immuno-epidemiological data and discusses its benefit over simpler statistical analysis approaches in common use.

Chapter 5 (Research paper 3) presents results for the second objective of this PhD study. It describes the use of latent class analysis to investigate the association between early childhood exposure to infections and reduced allergy-related disease outcomes in later childhood, using data from the Entebbe Mother and Baby Study.

Chapter 6 summarises the key findings from this PhD project, discusses strengths and limitations and presents concluding remarks.

Appendices are included after the references. **Appendix 1** contains a copy of the ethical approval form for conducting my PhD study, **Appendix 2** contains copies of ethical approvals for the IBS while **Appendix 3** contains copies of ethical approvals for the EMaBS.

Chapter 2. Methodology

2.1. Study setting

Both the IBS and EMaBS are based at the Entebbe Hospital and at the MRC/UVRI and LSHTM Uganda Research Unit under the Immunomodulation and Vaccines (I-Vac) Programme[48]. The MRC/UVRI and LSHTM Uganda Research Unit is an internationally recognized Centre of Excellence for research on HIV, other infectious diseases, neglected, endemic, emerging and re-emerging infections and non-communicable diseases. The Unit is based at the Uganda Virus Research Institute (UVRI) at the Entebbe campus with outposts in Masaka, Kalungu, Kampala and Wakiso Districts[49]. The I-Vac Programme at the Unit seeks to understand the impact of chronic and repeated infection on human immune responses, and the implications for outcomes including vaccine response and efficacy, and non-communicable diseases such as asthma and allergy.

Entebbe General Hospital is situated in the town of Entebbe, in Wakiso district. It is a public hospital funded by the Uganda Ministry of Health and serves patients from Wakiso and nearby districts including Entebbe Town, and neighbouring islands in Lake Victoria. Figure 2 shows the location of Entebbe General Hospital and the MRC/UVRI and LSHTM Uganda Research Unit in Entebbe town.

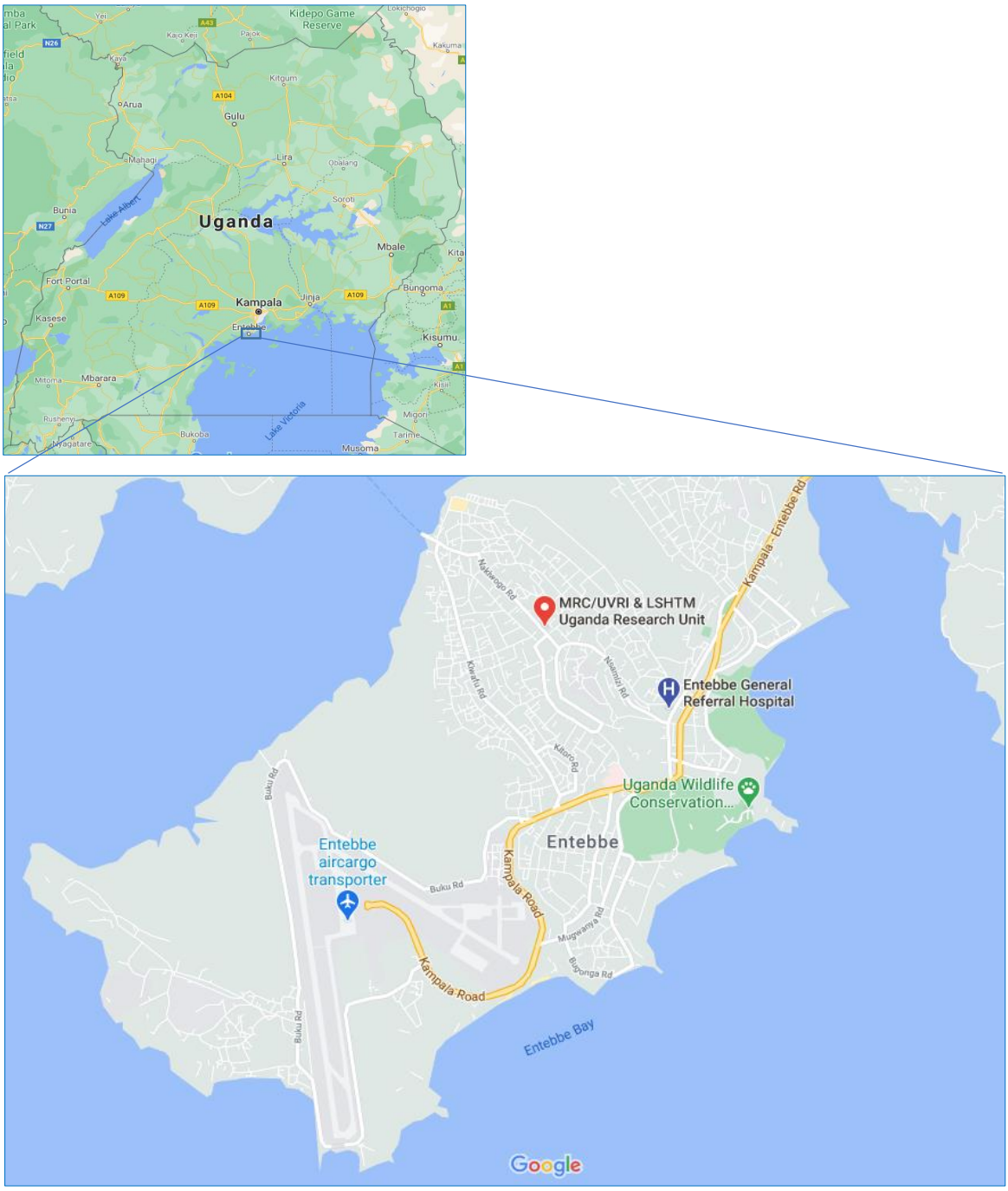


Figure 2: Map of Entebbe peninsula showing the location of Entebbe General Hospital and the MRC/UVRI and LSHTM Uganda Research Unit

2.2. Data sources

For the IBS, the primary outcomes were 17 cytokine/chemokine responses to the *M.tb* purified protein derivative (PPD) at the various time (weeks 1, 4, 6, 10, 14, 24 and 52) points. Table 1 below shows the 17 cytokines/chemokines and broad categories of function. Secondary outcomes were cytokine/chemokine responses to the early secretory antigenic target and culture filtrate protein antigens (ESAT6/CFP10) and antimycobacterial antibody responses at the various time points.

Table 1: IBS outcomes

Category	**Cytokine/Chemokine**	**Name in full**
Proinflammatory	IFN-γ	Interferon γ
	TNF-α	Tumor necrosis factor α
	IL-2	Interleukin 2
	IL-1α	Interleukin 1α
	IL-1β	Interleukin 1β
	IL-8	Interleukin 8
	IL-1Ra	Interleukin 1 receptor antagonist
T cell activation	IL-12p40	Interleukin 12p40
T cell regulation	IL-10	Interleukin 10
Th2	IL-5	Interleukin 5
	IL-13	Interleukin 13
Th17	IL-17A	Interleukin 17
Growth factor	GM-CSF	Granulocyte macrophage colony-stimulating factor
Chemokine	IP-10	IFN-γ-inducible protein 10
	MCP-1	Monocyte chemotactic protein 1
	MIP-1α	Macrophage inflammatory protein 1α
	MIP-1β	Macrophage inflammatory protein 1β

Baseline characteristics of mothers and infants who were enrolled in the study were tabulated by exposure group i.e. infants born of mothers with (LTBI positive) and without latent TB (LTBI

negative). Variables pre-specified as potential confounders for the association between LTBI and immune responses to vaccines included sex of the baby, child's birth weight, mother's age, number of pregnancies, malaria infection during pregnancy, maternal helminth infections, maternal tribe and maternal alcohol consumption.

For the EMaBS, primary outcomes of interest were allergy-related diseases (ARDs) at nine years, these included recent reported wheeze, eczema, rhinitis, Skin Prick Test (SPT) positivity to one or more allergens and detectable allergen-specific Immunoglobulin E (asIgE). Exposure illnesses/infections considered included clinical illnesses (malaria, diarrhoea, Upper and Lower Respiratory Tract Infections (URTI and LRTI), intestinal helminths) and seroprevalence of Herpes Simplex Virus (HSV), cytomegalovirus (CMV) and norovirus, in the first five years of life. Variables for maternal area of residence, maternal and household socio-economic status (SES) were investigated as potential confounders for the association between infection exposure and latent class membership.

2.3. Statistical methods

This section provides an overview of the statistical methods used for this PhD research. The overarching theme of the work was multivariate longitudinal analysis methods for immuno-epidemiological data and these are detailed below for each objective. These methods are further summarised in the individual papers.

2.3.1. Statistical methods for objective 1

The IBS presented many outcomes from multiplex data of 17 cytokine/chemokine responses to PPD and ESAT6/CFP10 at various time (weeks 1, 4, 6, 10, 14, 24 and 52) points; this raised the challenges of longitudinal outcomes and multiple testing arising from the fact that many cytokines at many different time points were examined. Cytokine/chemokine responses were skewed, therefore transformations were investigated to improve these distributions and the log transformation was considered and applied.

To reduce the dimensionality of the multiple response vector (17 outcome variables), approaches that involve principal component analysis were used and extensions for longitudinal data were considered. Statistical methods used to address objective 1 are as follows in sections 2.3.1.1. to 2.3.1.4.

2.3.1.1. Principal component analysis

Principal component analysis (PCA) is a technique used when there is a desire to reduce the dimensionality of multivariate datasets. It involves replacing the many correlated original variables with a smaller number of uncorrelated variables, referred to as principal components, which represent linear combinations of the original variables[50]. These fewer principal components often retain most of the variability of the original variables. Details of application of the PCA technique have been described in many publications[51-53]. In this study, we used PCA to investigate relationships between cord blood cytokine responses.

2.3.1.2. Three-way component analysis

Three-mode (three-way) component analysis[36] is an extension of principal component analysis (PCA) to three-mode data. Two-mode data consist of participants measured on several variables on a single occasion or set of conditions while three-mode data consist of participants (first mode) measured on the same variables (second mode) at several occasions, or conditions (third mode). The IBS data was a good example of three-mode data with infants (first mode) having 17 different cytokine/chemokine response variables (second mode) measured at several time points (third mode) post-BCG.

Three-way component analysis concerns the analysis of three-mode data. Several research questions can be handled via three-way analysis including i) What are the relations between variables? ii) What trends can be discovered over time? iii) Are there different categories of participants?[36] Each of these first three questions relates to only one of the three modes. More complex questions can be handled by three-way analysis; such questions could be iv) Do the relations between variables change over time? v) Do the relations between variables change over time in a different way for different categories of participants?[36]

Literature proposes several approaches for performing three-way component analysis, with the most popular techniques being the Tucker3 (T3)[54] and Candecomp/Parafac (CP)[55, 56] models. Here, the focus is on the T3 model which was used in this study and its methodological background is described. The CP model was not used because it is a constrained version of the T3.

Suppose the data array **X** is of order (*I* x *J* x *K*) where $i \in \{1, \dots, I\}$ denotes the number of participants, $j \in \{1, \dots, J\}$ denotes the number of variables and $k \in \{1, \dots, K\}$ denotes the number of occasions. For generality we can refer to the A-, B- and C- modes with *I*, *J* and *K* entities, respectively. The Tucker3 (T3)[54] model summarises **X** by extracting different components for each mode. The T3 model with *P*, *Q* and *R* components for the A-, B- and C- modes respectively, can be written as

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A (\mathbf{C}' \otimes \mathbf{B}') + \mathbf{E}_A ,$$

where A, B and C are component matrices for the A-, B- and C- modes respectively, with generic elements a_{ip}, b_{jq} and c_{kr} which express the component scores of the *i*-th entity on the *p*-th component for the A-mode, of the *j*-th entity on the *q*-th component for the B-mode, and of the *k*-th entity on the *r*-th component for the C-mode, respectively. $\mathbf{X}_A = [\mathbf{X}_{..1}, \dots, \mathbf{X}_{..k}, \dots \mathbf{X}_{..K}]$, where $\mathbf{X}_{..k}$ stands for the matrix frontal slab or slice of **X** i.e. $\mathbf{X}_A$ is the matrix with *I* rows and *JK* columns obtained by juxtaposing next to each other the frontal slabs of **X** (A-mode matricisation of **X**). $\mathbf{G}_A$ is a matricised version of the so-called core array **G** of order (*P* x *Q* x *R*) with generic element g_{pqr} which gives the interaction among components of the three modes. $\mathbf{E}_A$ denotes the matricised array of errors and the $\otimes$ symbol denotes the Kronecker product between two matrices[57]. A scalar formulation of the T3 model can be written as:

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}$$

With the T3 model, a limited number of components is sought for all three modes. However, it may be useful to reduce to only two modes or just one. In such cases, the Tucker2 (T2) or Tucker1 (T1) models may be introduced. Optimal parameter matrices of the T3, T2 and T1 models are

found by minimising $||\boldsymbol{E}_A||^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} e_{ijk}^2$, using an Alternating Least Squares (ALS) algorithm for T2 and T3 models or by singular value decomposition for T1 models.

For the IBS, three-way component analysis was used to explore relations between cytokine responses, following BCG, taking into account associations over all time points.

2.3.1.3. Linear Mixed Effects Models

The linear mixed effects model[58-62] has been shown to analyse repeated measurements in continuous longitudinal data in an easy, valid and flexible manner[63]. Suppose the random variable Y_{ij} denotes an outcome of interest (e.g. TNF response from the IBS), for the *i*th individual, measured at time t_{ij}, *i* = 1, 2, …, *N*, $j = 1, 2, \dots, n_i$, and suppose $\boldsymbol{Y}_i$ is the n_i-dimensional vector of all repeated measurements for the *i*th individual, according to Laird and Ware[64], the linear mixed effects model can be defined as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i\,; \quad \boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D}); \quad \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i); \quad \boldsymbol{b}_i, \boldsymbol{\varepsilon}_i \; are \; independent$$

where $\boldsymbol{Y}_i$ is the n_i dimensional response vector for the *i*th subject, $1 \leq i \leq N$, *N* is the number of subjects, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are (n_i x *p*) and (n_i x *q*) dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a *p*-dimensional vector of fixed effects, $\boldsymbol{b}_i$ is a *q*-dimensional vector of random effects, $\boldsymbol{\varepsilon}_i$ is an n_i-dimensional vector of residual components, $\boldsymbol{D}$ is a covariance matrix of random effects and $\boldsymbol{\Sigma}_i$ is a covariance matrix of residuals.

Model building followed guidelines proposed by Verbeke and Molenberghs[65], starting with exploration of preliminary mean structures of the cytokine responses over time, followed by investigation of serial correlation, reduction of random effects and reduction of mean structures consecutively.

To assess the need for serial correlation, models with the same mean structure and random effects but different serial correlation, including Exponential, Power, Gaussian and Simple, were assessed. Restricted maximum likelihood (REML) was employed while checking for serial correlation because it reduces the finite sample bias in the estimation of the covariance[60].

Inclusion of random effects was assessed using mixtures of chi-square distributions, due to the boundary problem. The boundary problem arises because these tests involve variances which cannot be negative. Therefore, a likelihood ratio test of the null hypothesis that a variance is zero is testing a hypothesis that is on "the boundary of the parameter space" for a variance. The consequence is that the usual null distribution for the likelihood ratio test is no longer valid and instead the null distribution is a mixture of chi-squared distributions[60].

To discover the most parsimonious mean structure, F-tests and likelihood ratio tests were employed under maximum likelihood estimation.

For the IBS, changes in log-transformed cytokine responses over time were studied, by mother's LTBI status, using univariate linear mixed effects models adjusted for factors that showed baseline differences between the two groups.

2.3.1.4. The pairwise joint modelling approach

Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles was introduced by Fieuws and Verbeke[35], to help avert computational problems which arise due to the dimension of the joint covariance matrix of random effects which results from increased number of outcomes or random effects considered. This approach has advantages over univariate mixed models especially when association structures of several longitudinal outcomes

are of interest, or when it is desired to study the joint effect of a treatment on several outcomes together.

The pairwise joint modelling approach resolves computational problems of high dimensional mixed effects models by first fitting all pairwise bivariate models separately, then in a second step, all estimates are averaged to obtain one single estimate for each parameter of the full joint model, and finally in a third step, pseudo-likelihood theory is applied in order to correctly calculate the sampling variability of the estimates from the pairwise approach[35, 66].

For the IBS dataset, the pairwise joint modelling approach was used in order to properly account for and to study the strength of association between the evolutions of cytokine responses over time and to test for the joint effect of maternal LTBI on evolution of all outcomes together.

2.3.2. Statistical methods for objective 2

For the EMaBS, I aimed to use latent variable modelling techniques to group participants based on their infection experience during the first five years of life and to study the association between early infection-exposure and ARDs at nine years. Infection experience (exposure) is an unobservable (latent) construct which can be inferred from multiple observed infections. Using multiple observed variables as indicators of a latent variable provides a more complete picture of the underlying construct and allows estimation of measurement error, leading to better measurement of the latent variable[37]. The Latent Class Analysis (LCA)[38, 67, 68] framework was used because it has been shown to flexibly model the relationship between discrete observed variables and a discrete latent variable[37].

2.3.2.1. Latent Transition Analysis

The initial analysis approach investigated was latent transition analysis (LTA), a longitudinal extension of LCA[39], because infection-exposure data were collected at different time-points over the first five years of life for EMaBS participants. LTA allows for estimation of prevalence of latent classes as well as the incidence of transitions to different latent classes over time and prediction of both latent class memberships and transitions over time[37, 69, 70]. Several applications of LTA have been reported for smoking behaviour[69], pubertal development[71], substance use, sexual behaviour among injection drug users, design and evaluation of intervention programs[37] and to assess effectiveness of interventions by incorporating treatment as a grouping variable[72].

A major requirement for LTA is longitudinal measurement invariance of item-response probabilities[37, 70]. This measurement invariance requirement ensures that the latent classes preserve their characteristics and meaning across time points[37, 70]. The LTA mathematical model has been discussed in detail elsewhere[37, 39]. For the EMaBS objective 2 analysis, the formal test for measurement invariance was done by comparing a model with all item-response probability parameters estimated freely to one in which those parameters are constrained to be equal across time[37]. In this study, since the requirement for longitudinal measurement invariance did not hold, LCA models were fitted separately at each time point, as recommended[70].

2.3.2.2. Latent Class Analysis at each time point

The LCA mathematical model has been described[37, 39] as follows. Suppose there are $j = 1, 2, \ldots, J$ observed variables and variable j has $r_j = 1, 2, \ldots, R_j$ response categories; and that the latent variable has $c = 1, 2, \ldots, C$ latent classes; $\boldsymbol{y}$ represents a particular response pattern (a vector of

possible responses to observed variables), and **Y** represents the array of all possible ***y***s. The estimated proportion of a particular response pattern P(**Y** = ***y***) can be expressed as a function of two parameters namely the latent class prevalences, γ parameters, and item-response probabilities, ρ parameters. The γ parameters sum up to 1 because latent classes are mutually exclusive and exhaustive. The ρ parameters express the relationship between observed and latent variables.

Suppose an indicator function $I(y_j = r_j)$ equals 1 when the response to variable $j = r_j$, and 0 otherwise, the probability of observing a particular response pattern can be written as

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)}$$

where γ_c is the probability of membership in latent class c and $\rho_{j,r_j|c}^{I(y_j=r_j)}$ is the probability of response r_j to observed variable j, conditional on membership in latent class c. The γ parameters sum up to 1 and represent a vector of latent class membership probabilities while the ρ parameters express a matrix of item-response probabilities conditional on latent class membership[37].

Two assumptions are critical when defining the latent class model. First, all individuals in a latent class are assumed to have the same item-response probabilities for the observed variables. Second, there is an assumption of local independence which states that conditional on latent class, observed variables are independent. Consequently, because the full dataset is a mixture of

several latent classes, this assumption does not imply that the observed variables are independent in the full sample[39].

For the EMaBS, prospectively collected data on malaria, diarrhoea, Upper and Lower Respiratory Tract Infections (URTI and LRTI), intestinal helminths, seroprevalence of Herpes Simplex Virus (HSV), cytomegalovirus (CMV) and norovirus were used to characterise the latent construct of infection exposure. Models were fitted, successively increasing the number of classes up to a seven-class solution, separately for each of the first five years, with no a priori restrictions to find consistent classes across the different years. The Bayesian Information Criterion (BIC), adjusted BIC (ABIC), consistent Akaike Information Criterion (CAIC) and entropy were used to guide the choice of optimal number of classes[67, 73].

Multiple-group LCA[67] was conducted to establish whether measurement invariance across sex holds. Multiple-group LCA allows class membership and item response probabilities to vary across groups and a formal statistical test for measurement invariance across groups can be conducted[37].

LCA also allows for incorporation of covariates in the model in order to predict latent class membership and these are added to the model via multinomial logistic regression[67]. In this study, we included the covariates maternal area of residence, and maternal and household socio-economic status (SES) at enrolment, to establish whether these impacted the probability of latent class membership.

Another vital extension of LCA which was applied in this study was the opportunity to predict distal outcomes from latent class membership. Two broad classes of approaches of LCA with

distal outcomes have been discussed in the literature namely the classify-analyse[74-76] and model-based[77] approaches. The classify-analyse approaches involve assigning latent class membership and conducting the outcome analysis in separate steps. Consequently, they ignore the uncertainty related to class membership and they also impute the latent variable under a model that is not sufficiently general and this may lead to attenuated estimates of the relation between the latent variable and the distal outcome[77]. The model-based approach has been shown to be consistent and to produce less biased estimates of the association between latent class membership and distal outcomes[77]. For the EMaBS, ARDs at nine years were predicted from latent class membership, using a SAS macro[78] implementing a model-based approach as described[77].

2.3.3. Statistical methods for objective 3

Missing data were common in both the IBS and EMaBS. In the IBS, infants were randomly assigned to two separate sampling strategies: 150 infants (75 from mothers with latent *Mycobacterium tuberculosis* infection (LTBI), and 75 from mothers without LTBI) to give blood at 1, 6, 14 and 52 weeks while the remaining 150 infants were randomly assigned to give blood at 4, 10, 24 and 52 weeks. This means that at the time points where only one group were sampled, and at two time points which were introduced at a later stage (namely 14 and 24 weeks), there was missing data by design. In addition, there was some attrition throughout the study period.

In the EMaBS, similar issues of attrition arise with decreasing proportions of participants being seen over time, for instance 61% and 52% of live-born participants were seen at the 5 and 9 year annual visits, respectively.

Inadequate handling of missing data in a statistical analysis may lead to biased or inefficient estimates of parameters and biased standard errors resulting in incorrect confidence intervals and significance tests[79]. During statistical analysis, assumptions are often made about the missing data. Rubin[80] introduced the taxonomy of missing data mechanisms, classifying missing data as being missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). The MCAR mechanism potentially depends on observed covariates, but neither on observed nor unobserved outcomes; a MAR mechanism depends on the observed outcomes and perhaps also on the covariates, but not further on unobserved measurements while a MNAR mechanism depends on unobserved measurements, maybe in addition to dependencies on covariates and/or on observed outcomes[7]. It is not possible to distinguish between MAR and MNAR from the observed data alone, although the MAR assumption can be approximated by collecting more explanatory variables and including them in the analysis[79].

In this section, the missing data approaches considered and applied within the framework of the various statistical analysis methods used for the IBS and EMaBS data are highlighted.

2.3.3.1. Multiple imputation

Multiple imputation (MI) gives unbiased and valid estimates if data are MAR, but the method can also handle both MCAR and MNAR[81]. MI was introduced over 40 years ago[82] and is now a very important and popular approach in the statistical analysis of missing data[83]. MI has been shown to be a convenient and flexible paradigm for analysing data with missing values[84], with a major goal of providing statistically valid inference[85]. The idea of MI is to use the distribution of the observed data to estimate a set of plausible values for the missing data. Multiple data sets are

created and then analysed individually but identically to obtain a set of parameter estimates, then the estimates are combined to obtain the overall estimates, variances and confidence intervals[79]. MI was used while fitting three-way (T3) models to the IBS data as described below.

The first step of analysis with MI involves generation of several complete datasets with imputed values for the missing data. An appropriate imputation model is required in order to generate imputed values which closely mimic the structure of the observed data[86]. The IBS data were of a hierarchical or multilevel nature since study participants had several measurements (cytokine responses) taken at various time points; they were also characterised by missing data by design as described earlier. The joint modelling approach to performing MI for multilevel missing data was implemented. This approach uses a single statistical model for imputing all incomplete variables simultaneously[87] and can be implemented using the R package pan[88]. The underlying statistical model, for the pan R package, is a multivariate extension of the linear mixed effects model, representing multiple dependent variables simultaneously[87, 88]. For the IBS data, the imputation model included the cytokine response variables as dependent variables, fixed effects of time (in weeks) and LTBI status as predictor variables and random intercepts were considered for the random effects structure.

The pan algorithm uses Markov Chain Monte Carlo (MCMC) methods to sample replacements for missing values, drawing from a posterior predictive distribution of the missing data[89]. The procedure is characterised by two phases: a burn-in phase and an imputation phase. The burn-in phase involves several iterations, before any replacements are drawn, with a sole aim of ensuring that model parameters converge to stationary distributions. The imputation phase then follows, allowing for a number (m) of imputed data sets to be drawn, each spread a number of iterations

apart. Drawing imputations from iterations that are not consecutive ensures that autocorrelation is avoided and the imputed datasets form independent random draws from the posterior predictive distribution. For the IBS dataset, pan was run for 50,000 burn-in iterations, then m = 100 imputed data sets were drawn, each spread 5,000 iterations apart. Several studies recommend generating such large numbers of imputations, especially when large portions of the data are missing[88, 90, 91]. The pan procedure offers some convergence diagnostics to help ensure that the algorithm converges and that the imputed datasets are approximately independent draws from the posterior predictive distribution. These diagnostics include the potential scale reduction factor ($\bar{R}$)[92] and diagnostic (trace, autocorrelation and density) plots[88].

The second step of data analysis with MI is to analyse the multiply imputed datasets individually but identically. Three-way component analysis was therefore applied to each of the 100 imputed datasets using the function T3 of the R package ThreeWay[57].

In a third step, the parameter estimates from each of the imputed datasets are pooled to give the final parameter estimates. Often, it is sufficient to simply average over the multiple values of a statistic derived from the multiply imputed datasets, however, Van Ginkel and Kroonenberg[93] showed that rather than averaging loadings, Generalised Procrustes Analysis (GPA)[94] should be used to combine the PCA loadings from multiply imputed datasets. GPA is a generalisation of a Procrustes rotation[95], in which matrices are aligned towards each other and the centroid of the optimally aligned matrices serves as the average[86]. GPA was therefore used to pool the solutions from the three-mode analysis conducted on each of the 100 imputed datasets.

2.3.3.2. Direct likelihood methods for handling missing data

Under the MAR assumption, valid inferences can be obtained through likelihood based analyses without the need for modelling the dropout process[7]. Consequently, linear or generalised linear mixed effects models can be used without additional complication[7]. Direct likelihood methods have the advantage of utilising all the available data rather than using only the complete cases. This approach was applied to the IBS study, using linear mixed effects models as described earlier.

The LCA procedures, used for the EMaBS dataset as described above, employ a maximum-likelihood routine which allows for MAR missingness, but not for MNAR missingness. Several studies have indicated that parameter recovery for data that are MCAR or MAR is not substantially biased regardless of the amount of missing data, the sample size, or the latent class model[37]. The maximum-likelihood missing data procedure is most successful if all variables relevant to missingness are included in the model. If there are relevant variables that cannot be included, a multiple imputation approach may be preferable[96]. This approach was considered appropriate for the EMaBS dataset because of the considerably large amount of data available.

2.4. Statistical software

Most of the data processing and descriptive analyses were conducted using Stata 15.0 (College Station, Texas, USA) and R version 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria). Estimation of the T3 model parameters was done using the function T3 of the R package ThreeWay[57] while multiple imputation was done using the R package pan[88]. The function "panImpute" is the main interface of the "mitml" package to run the pan algorithm in R.

Statistical Analysis System (SAS) software version 9.4 (SAS Institute Inc., Cary, NC, USA) was used to fit Linear Mixed Effects Models, Latent Class Analysis and the pairwise joint modelling approach. LCA models were estimated using PROC LCA[67] and a SAS macro[78] was used to implement the model-based approach for LCA with distal outcomes. The pairwise joint modelling approach was also implemented using a SAS macro[97], adapted to suit the IBS dataset.

2.5. Ethical approvals

The IBS and EMaBS were granted ethical approval from the Uganda Virus Research Institute Research Ethics Committee, the Uganda National Council for Science and Technology (UNCST), and the London School of Hygiene and Tropical Medicine. Further clearance for this PhD study was granted by the Human Research Ethics Committee of the University of Witwatersrand. In both studies, participants provided signed informed consent. For minors, participants' parents or guardians provided consent. For EMaBS participants at the 9-year annual visit, minors provided assent. Consent and assent forms were translated into the local language (Luganda) and participants chose whichever language they understood better. Copies of ethical approval forms are included as appendices 1 to 3.

Chapter 3. A longitudinal data analysis approach for understanding the dynamics of vaccine responses and how they differ in response to selected exposures

3.1. Introduction

This chapter illustrates the use of longitudinal data analysis approaches, including the linear mixed model and three-way component analysis, to study the dynamics of vaccine responses and how they differ in response to selected exposures, using data from the IBS (**thesis objective 1**). The linear mixed model handles missing data via a likelihood based approach under the MAR assumption while multiple imputation was used when fitting three-way models (**thesis objective 3**). Results are presented in Research paper 1, entitled "Maternal latent *Mycobacterium tuberculosis* does not affect the infant immune response following BCG at birth: an observational longitudinal study in Uganda", this has been published as an original article in the journal *Frontiers in Immunology*.

Differences in immune sensitisation between infants born of mothers with or without LTBI are studied. In addition, patterns of post-BCG cytokine and antibody responses to mycobacterial antigens are compared between infants of mothers with or without LTBI. These analyses show a first instance of extension of appropriate statistical analysis methods to longitudinal immuno-epidemiological data in line with the main aim of this thesis. Further extensions of appropriate statistical models for longitudinal multiple-outcome/exposure data are illustrated in subsequent chapters.

3.2. Research Paper 1: Maternal latent *Mycobacterium tuberculosis* does not affect the infant immune response following BCG at birth: an observational longitudinal study in Uganda

frontiers in Immunology

ORIGINAL RESEARCH
published: 14 May 2020
doi: 10.3389/fimmu.2020.00929

Maternal Latent *Mycobacterium tuberculosis* Does Not Affect the Infant Immune Response Following BCG at Birth: An Observational Longitudinal Study in Uganda

Lawrence Lubyayi [1,2†], Patrice A. Mawa [1,3†], Grace Nabakooza [1], Marjorie Nakibuule [1], John Vianney Tushabe [1], Joel Serubanja [1], Dorothy Aibo [1], Hellen Akurut [1], Josephine Tumusiime [1], Mateusz Hasso-Agopsowicz [4], Pontiano Kaleebu [1,3], Jonathan Levin [2], Hazel M. Dockrell [4], Steven Smith [5], Emily L. Webb [6], Alison M. Elliott [1,7] and Stephen Cose [1,7]*

[1] Immunomodulation and Vaccines Programme, Medical Research Council/Uganda Virus Research Institute and London School of Hygiene and Tropical Medicine Uganda Research Unit Entebbe, Entebbe, Uganda, [2] Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa, [3] Department of Immunology, Uganda Virus Research Institute, Entebbe, Uganda, [4] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom, [5] Department of Life Sciences, Brunel University London, London, United Kingdom, [6] MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom, [7] Department of Clinical Research, London School of Hygiene and Tropical Medicine, London, United Kingdom

OPEN ACCESS

Edited by:
Pere-Joan Cardona, Germans Trias i Pujol Health Science Research Institute (IGTP), Spain

Reviewed by:
Carmen Alvarez-Dominguez, Marqués de Valdecilla Health Research Institute (IDIVAL), Spain
Rachel Tanner, University of Oxford, United Kingdom
Warwick Britton, University of Sydney School of Medicine, Australia

***Correspondence**:
Lawrence Lubyayi
lawrence.lubyayi@mrcuganda.org; lawrencelby@gmail.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Vaccines and Molecular Therapeutics, a section of the journal Frontiers in Immunology

Received: 24 December 2019
Accepted: 21 April 2020
Published: 14 May 2020

Citation:
Lubyayi L, Mawa PA, Nabakooza G, Nakibuule M, Tushabe JV, Serubanja J, Aibo D, Akurut H, Tumusiime J, Hasso-Agopsowicz M, Kaleebu P, Levin J, Dockrell HM, Smith S, Webb EL, Elliott AM and Cose S (2020) Maternal Latent Mycobacterium tuberculosis Does Not Affect the Infant Immune Response Following BCG at Birth: An Observational Longitudinal Study in Uganda. Front. Immunol. 11:929. doi: 10.3389/fimmu.2020.00929

Background: BCG has low efficacy in tropical countries. We hypothesized that maternal latent *Mycobacterium tuberculosis (M.tb)* infection (LTBI) results in fetal tolerance to mycobacterial antigens and impaired responses to BCG immunization.

Methods: We enrolled 132 LTBI-positive and 150 LTBI-negative mothers and their babies in Entebbe, Uganda. Infants were BCG-immunized at birth. Cord blood and samples at weeks 1, 4, 6, 10, 14, 24, and 52 were analyzed for cytokine/chemokine responses to *M.tb* antigens by Luminex 17-plex assay in 6-day whole blood cultures and antibody responses by ELISA. Of the 17 Luminex analytes, seven (IL-2, IL-5, IL-10, IL-13, IL-17A, TNF, and IFN-γ) were included in the main analysis as they were considered most likely to represent T cell responses. Immune sensitization was defined as a detectable cord blood cytokine response to PPD for any of the seven cytokines. Patterns of cytokine and antibody responses were compared between infants of mothers with and without LTBI using linear mixed models adjusting for confounders.

Results: Most infants (73%) were sensitized *in utero* to *M.tb* antigens, with no overall difference seen between infants born to mothers with or without LTBI. Patterns of post-BCG cytokine and antibody responses to mycobacterial antigens were similar between the two infant groups.

Conclusions: Our data do not support the hypothesis that maternal LTBI results in an impaired response to BCG immunization, in Ugandan infants. BCG vaccination at or shortly after birth is likely to be beneficial to all infants, irrespective of maternal LTBI status.

Keywords: latent *Mycobacterium tuberculosis* infection, maternal infection, BCG vaccine, cytokine responses, antibody responses

INTRODUCTION

Bacille Calmette-Guerin (BCG) is the only licensed vaccine against tuberculosis (TB). It protects against tuberculous meningitis and miliary TB in infants (1), but its protective efficacy against pulmonary TB varies between populations. Meta-analyses of BCG vaccine trials have shown that latitude is an important factor for responses in adolescents and adults, with lower protection closer to the equator (2–5).

Modification of the protective effect of BCG through sensitization to non-tuberculous mycobacteria (NTMs) has been suggested as a reason for variable BCG efficacy, and its association with latitude (6, 7). The protective effects of BCG might be blocked by exposure to NTMs, or NTMs might provide equivalent protection to BCG, thus masking the benefit provided by BCG (8). Although NTMs have a variable distribution by latitude (9), NTM exposure may not fully explain this variability (10).

In TB endemic areas, BCG is administered to new-borns at birth, in accordance with WHO recommendations (11). BCG elicits different profiles of immune response in Africa compared with the UK when given early in life (12). Prior sensitization, perhaps due to early exposure to *Mycobacterium tuberculosis (M.tb)* itself, or to environmental mycobacteria has been reported in infants immunized some months after birth (13). However, *in utero*, rather than early life, sensitization (14) may result in a more substantial modification of responses in exposed infants.

In TB endemic areas, a high proportion of adults harbor latent *M.tb* infections (LTBI). A dynamic relationship between mycobacteria and the immune system is thought to exist during LTBI. Individuals with LTBI may have circulating antigens and higher concentrations of TB-specific antibodies, plasmablasts, and memory B cells than those without infection (15, 16). Mycobacterial antigens cross the placenta in murine models (17). Thus, maternal LTBI might lead to exposure to mycobacterial antigens *in utero* driving a modified profile of sensitization (18), or inducing tolerance in the fetus (14, 19). Alternatively, passive transfer of maternal anti-mycobacterial antibodies (by providing passive immunity) or maternal anti- idiotype antibodies (mimicking antigen) (20), might influence the ability of neonatal BCG vaccine to elicit protective immune responses. The maternal and placental immunological milieu could also be influenced non-specifically by maternal LTBI, with consequences for fetal and neonatal response following immunization (21). For other pathogens, maternal infections have been shown to induce either tolerization or sensitization in the fetus, with subsequent differences in susceptibility to infection (22). We previously showed impaired mycobacteria-specific T-cell responses following BCG immunization of infants born to LTBI-positive mothers, although this effect appeared to be transient (23).

We hypothesized that maternal LTBI influences the neonatal response to mycobacteria, impairing the response to BCG and *M.tb*. To investigate this we followed a cohort of infants of LTBI infected or uninfected mothers over the first year of life. We measured cellular responses induced by neonatal BCG immunization using a whole blood assay (24). As well, the evolution of anti-mycobacterial antibody responses was assessed, since these have recently gained new recognition for a potential role in protective immunity against tuberculosis (25–28). We measured responses to both the relatively *M.tb*-specific antigens (ESAT6 and CFP10) and to the broadly mycobacterial-specific purified protein derivative (PPD) in order to evaluate exposure to *M.tb*, and acquisition of responses to *M.tb*, *in utero* and during the first year of life, as distinct from the response to BCG.

MATERIALS AND METHODS

Study Design and Participants

Healthy mothers and their infants were recruited at Entebbe General Hospital between June 2014 and October 2016. Women who were willing to participate in the study, had a normal singleton pregnancy, resided in Entebbe municipality or neighboring Katabi sub-county, and were HIV negative were eligible for inclusion. They were excluded if cord blood was not obtained, delivery was not normal, the mother was unwilling to undergo a repeat HIV test or was found to be HIV-positive on repeat testing, birth weight was $<$2,500 g, the neonate was unwell as judged by the midwife, the mother had indeterminate LTBI status (as described below), or the neonate presented with significant congenital abnormalities likely to impair the child's general health and development. Enrolled infants received all vaccines recommended by the Expanded Programme on Immunization. Infants were included in the study based on their mother's LTBI status, targeting equal numbers of infected and uninfected women. All infants were immunized at birth or within the first week of life with a single dose of intradermal BCG (Statens Serum Institut (SSI), Denmark).

Blood Sampling Strategy

Up to 7 ml of cord blood was collected. Infants were then randomly assigned in a 1:1 ratio (stratified by LTBI status) to two sampling strategies to reduce the blood-sampling burden on individual infants. Half gave 2 ml venous blood at 1, 6, and 14 weeks, the remainder at 4, 10, and 24 weeks. All gave blood (5 ml) at 52 weeks. Blood draws at weeks 14 and 24 were introduced when the study was already underway, resulting in lower sample numbers at these times.

Tests for Latent TB Infection

Women were investigated for LTBI at approximately 1 week post-delivery using the tuberculin skin test (TST) (PPD RT23 SSI, Copenhagen, Denmark) and T-SPOT.TB assay (Oxford Immunotec, Abingdon, UK) (29). The TST was performed in the mothers after bleeding for the T-SPOT.TB assay and was read 48–72 h later, and defined as positive if ≥10 mm in diameter. Women positive on both tests were considered LTBI-positive; those negative on both tests were considered LTBI-negative. Those with indeterminate and discordant results were excluded in order to optimize our ability to determine effects of LTBI (as distinct from other mycobacterial exposures). LTBI-positive mothers were investigated for active tuberculosis by symptoms, sputum examination (if available), and chest x-ray. No cases of active TB were detected.

Whole Blood Assays for Cytokine/Chemokine Responses

Whole blood (including cord and infant venous blood) was diluted 1 in 5 in RPMI 1640 (Invitrogen) supplemented with 2 mM L-glutamine (Invitrogen) and cultured under 5% carbon dioxide at 37°C for 6 days in 96-well U-bottomed plates (final volume 200 µl). Duplicate wells were incubated with medium alone (negative control), PPD (Statens Serum Institut, catalog #RT50) (10 µg/ml), or a combination of ESAT6 and CFP10 antigens (BEI Resources, calatologue #sNR14868 and NR-49425) (5 µg/ml).

After 6 days, plates were centrifuged at 400 g for 5 min. Supernatants were removed from duplicate wells, pooled, and stored at −80°C prior to analysis. Thawed supernatants were randomized across plates and subjected to multiplex bead array analysis using the human cytokine/chemokine MilliplexTM MAP 17-plex pre-mixed kit (Merck Millipore), following the manufacturer's instructions. The pre-mixed bead set included interleukin (IL)-1α, IL-1β, IL-1Ra, IL-2, IL-5, IL- 8, IL-10, IL-12p40, IL-13, IL-17A, interferon (IFN)-γ, IFN-γ-inducible protein (IP)-10, monocyte chemotactic protein (MCP)-1, macrophage inflammatory protein (MIP)-1α, MIP-1β, tumor necrosis factor (TNF), and granulocyte macrophage colony-stimulating factor (GM-CSF). Data were acquired using the Biorad Luminex® 200 system and Bioplex Manager Software version 6.1 (Biorad).

ELISA for Anti-mycobacterial IgG Antibodies

Total immunoglobulin (Ig)G against PPD, ESAT6, CFP10, and Ag85A were assayed in plasma of a random sample of infants, at each time point, by ELISA as described elsewhere (23). Briefly, flat-bottomed 96-well microlon plates (Greiner Bio-one, Germany) were coated with purified IgG standard (GenScript, NJ, USA) and mycobacterial antigens. After overnight incubation, the plates were blocked and samples diluted 1 in 100 were added to the plates and left overnight at 4°C. Polyclonal anti-human IgG Horse Radish Peroxidase (Poly HRP, 0.5 µg/ml, Dako, Denmark) was added and plates incubated. 0-Phenylenediamine (OPD, Sigma-Aldrich, MO, USA) substrate mixture (3 mg OPD, 0.1M citric acid, 0.2M Na2HPO4, 3 µL 30% hydrogen peroxide in distilled water) was then added and the reaction stopped with 2M Sulphuric acid and read at test wavelength 490 nm and reference wavelength 630 nm using a MRX1.1 plate reader and Gen5 1.07 Software (BioTek Instruments, Inc., VT, USA). The lowest standard concentration above which antibody was detectable (0.01 µg/ml) was set as the sensitivity of the assay.

Statistical Methods

Analyses studied the time course of PPD- and ESAT6/CFP10-specific responses at 1, 4, 6, 10, 14, 24, and 52 weeks after BCG immunization, and the influence of maternal LTBI on infant responses.

We aimed to recruit 150 women with LTBI and 150 without, to give 80% power to detect a difference of 0.35log10 (assuming a standard deviation of 0.9log10) (30) in infant cytokine response at 52 weeks between the two groups, and a difference of 0.5log10 at other time points (with 75 infants in each group).

Unstimulated cytokine response values were subtracted from antigen-stimulated results. Values <3.2 pg/mL (lower detection limit of assay) were assigned as 3.2 pg/mL. Values above 11,000 pg/mL (the upper detection limit) were assigned 11,000 pg/mL.

Baseline characteristics of participants were summarized, by LTBI status, using percentages, means and standard deviations, and medians and interquartile ranges. Of the 17 Luminex analytes, seven (IL-2, IL-5, IL-10, IL-13, IL-17A, TNF, and IFN-γ) were included in the main analysis as they were considered most likely to represent T cell responses. Results from the remaining ten cytokines were included in supplementary analyses.

Principal components analysis (PCA), a procedure which transforms several (possibly) correlated variables into a smaller number of uncorrelated variables (principal components), was used on cord blood outcomes to investigate relationships between the seven cytokine responses. We defined immune sensitization as a detectable cytokine response to PPD, in cord blood, for any of the seven cytokines. We summarized proportions of infants sensitized based on each of the seven cytokines and on groups discovered by PCA, and compared these proportions by maternal LTBI status using chi-squared tests.

Changes in log-transformed cytokine responses over time, by mother's LTBI status, were studied using linear mixed models adjusted for factors that showed baseline differences between the two groups. Profile plots showing the mean concentration and 95% confidence intervals (CI) were used to visualize the differences. Similar plots were generated for antibody responses.

Three-way component analysis (31) was used to explore relations between cytokine responses, following BCG, taking into account trends over all time points.

Data analysis was conducted using Stata 14.1 (College Station, Texas, USA) and R version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria). A 5% significance level was used for all analyses.

Ethical Approvals

Ethical approval was given by the Uganda Virus Research Institute Research Ethics Committee (reference GC/127/13/09/16 and GC/127/16/03/434), Uganda National Council for Science and Technology (reference HS 1526) and the London School of Hygiene & Tropical Medicine (reference 7104). Written informed consent was received from all mothers, for their own and their baby's participation.

RESULTS

Participants' Characteristics

The study flowchart is shown in **Figure 1**. Between June 2014 and October 2016, 1134 women were invited to participate, 798 (70%) of whom consented and were tested for LTBI. 390 women tested negative on both TST and TSPOT.TB while 133 were positive on both tests. Among the 390 LTBI-negative women, systematic sampling was done to randomly select those for enrolment:

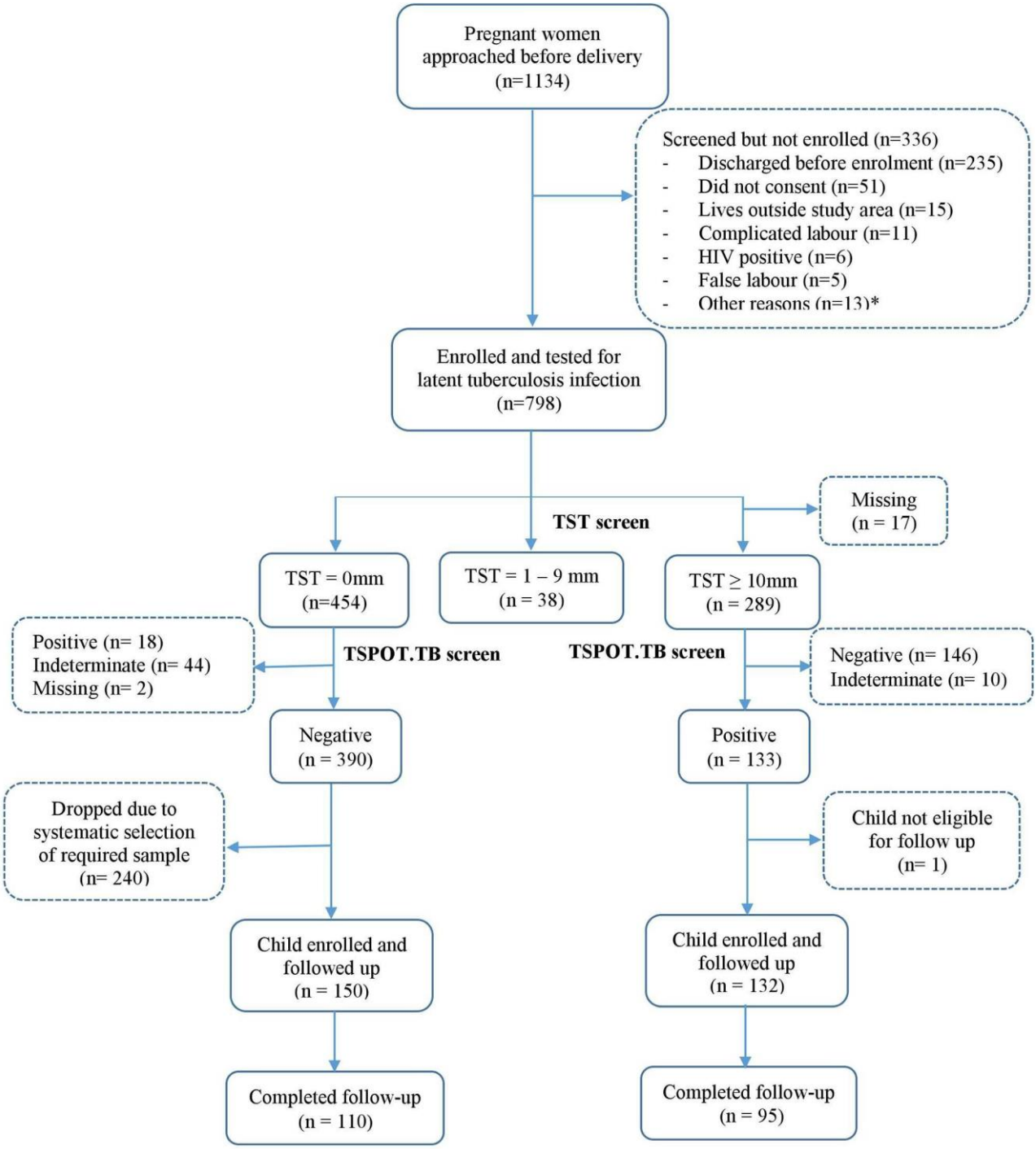


FIGURE 1 | Study flowchart. *Other reasons include discharged before consenting procedures, low birthweight babies, cord blood not taken, asphyxia, child not vaccinated, and birth to twins.

initially we enrolled every second woman, later every third woman, to ensure contemporaneous recruitment with LTBI-positive women. Of the 133 LTBI-positive women, 132, with their infants, were eligible for follow up and enrolled into the study.

Baseline characteristics of participants were similar between the two groups (**Table 1**), except that LTBI-positive mothers were on average older, more likely to have lived with someone who had TB, to drink alcohol and to originate from the central region of

TABLE 1 | Characteristics of study participants.

	LTBI Negative (*n* = 150)		LTBI Positive (*n* = 132)	
Maternal characteristics				
Mean mother's age (SD) (mv 0, 3)[a]	23.65	(3.67)	25.53	(4.99)
Median number of pregnancies (IQR) (mv 1, 1)	2	(1-3)	2	(2-4)
Median number of births (IQR) (mv 1, 1)	2	(1-3)	2	(1-3)
Positive malaria test during pregnancy (mv 1, 1)	38	25.5%	29	22.1%
Ever lived with someone with TB (mv 2, 1)	3	2.0%	19	14.5%
BCG scarring (mv 1, 2)	106	71.1%	95	73.1%
Ever taken any medicine for worms (mv 1, 2)	135	90.6%	116	89.2%
Current marital status (mv 2, 4)				
Single	28	18.9%	20	15.6%
Married/living as married	120	81.1%	108	84.4%
Highest level of education attained (mv 2, 3)				
Never attended school	5	3.4%	2	1.5%
Primary	50	33.8%	44	33.9%
Secondary	81	54.7%	66	50.8%
Tertiary	12	8.1%	17	13.1%
Smoked in the past (mv 1, 1)	1	0.7%	1	0.8%
Drink alcohol (mv 3, 4)	18	12.2%	28	21.9%
Schistosoma mansoni infected (mv 33, 36)	13	11.1%	9	9.4%
Hookworm infected (mv 33, 36)	9	7.7%	2	2.1%
Trichuris infected (mv 33, 36)	1	0.9%	1	1.0%
Any helminth infection (mv 33, 36)	20	17.1%	12	12.5%
Mother's tribe grouping (mv 5, 3)				
Central	56	38.6%	72	55.8%
Other	89	61.4%	57	44.2%
Father's tribe grouping (mv 3, 2)				
Central	58	39.5%	75	57.7%
Other	89	60.5%	55	42.3%
Infant characteristics				
Sex of the baby, male	77	51.3%	77	58.3%
Mean birth weight in kg (SD)	3.24	(0.43)	3.21	(0.40)

Data are mean (SD), median (IQR), or n (%). SD, standard deviation; IQR, interquartile range; mv, missing values. [a]*Figures in parentheses indicate missing values in the LTBI-Negative and LTBI-Positive groups, respectively.*

Uganda. Infant characteristics (sex and birth weight) were similar between the two groups.

Due to the study design, not all infants provided samples at all-time points. **Supplementary Table 1** shows sample numbers assayed at each time point.

Cord Blood Outcomes and Immune Sensitization

Based on cord blood responses to PPD for any of the seven cytokines IL-2, IL-5, IL-10, IL-13, IL-17A, TNF, and IFN-γ, 73% of the infants made a positive cytokine response to *M.tb* antigens, with no overall difference seen between infants born to mothers with or without LTBI (78 vs. 69%, respectively; $p = 0.119$). Considering individual cytokines, a larger proportion of infants born of LTBI-positive mothers had PPD-specific TNF responses in cord blood as compared to those born of LTBI-negative mothers (74 vs. 61%, respectively; $p = 0.038$) (**Table 2**).

TABLE 2 | Immune sensitization based on cord blood responses to PPD.

Cytokines	LTBI-Negative (*n* = 138)		LTBI-Positive (*n* = 116)		*p*-value
Individual cytokines					
IL2	12	8.7%	10	8.6%	0.983
IL5	7	5.1%	7	6.0%	0.738
IL10	38	27.5%	45	38.8%	0.057
IL13	21	15.2%	17	14.7%	0.900
IL17A	13	9.4%	7	6.0%	0.318
TNF	84	61.3%	84	73.7%	0.038
IFN-γ	17	12.3%	14	12.1%	0.952
Based on any of the 7 cytokines					
IL2, IL5, IL10, IL13, IL17A, TNF and IFN-γ	95	68.8%	90	77.6%	0.119
Based on PCA grouping					
IL2, IL5, IL13, IL17A and IFN-γ	30	21.7%	26	22.4%	0.897

For ESAT6/CFP10 responses, there was no overall difference in responses between infant groups, and >90% of all infants had a positive TNF response (**Supplementary Table 2**).

PCA component loadings for cord blood cytokine responses to both PPD and ESAT6/CFP10, are shown in **Figure 2**. The cytokine responses to PPD, IL-10 and TNF clustered separately from the other cytokines. The cytokine responses to ESAT6/CFP10, IL-10, TNF, and IFN-γ clustered separately from the other cytokines. The PCA clustering profiles described did not differ between the two infant groups.

Maternal LTBI and Infant Cytokine Responses Over Time

Cytokine responses to both PPD and ESAT6/CFP10 were similar, at all-time points, between the two infant groups (**Figures 3**, **4** and **Supplementary Table 3**). The peak of the infant response to PPD was sustained from 10 to 24 weeks of age (**Figure 3**). From univariate linear mixed models, adjusted for sex of the infant and factors that showed baseline differences between the two groups (mother's age, household TB contact, alcohol consumption, parental tribe), there was no evidence of interaction between time and mother's LTBI status for any of the seven outcomes for either PPD or ESAT6/CFP10 responses (**Figures 5**, **6**), demonstrating that the association between LTBI status and the responses did not change over time.

Cytokine responses to ESAT6/CFP10 varied over time. IFN-γ and TNF responses started high, increased up to about 24 weeks and then plateaued; IL-13 and IL-17A responses increased up to 4 weeks, declined up to 10 weeks and then plateaued; IL-2 and IL-5 responses were consistently very low, whilst IL-10 responses declined up to 10 weeks, increased up to 14 weeks and then plateaued (**Figure 4**). These responses also showed no difference between the groups over time.

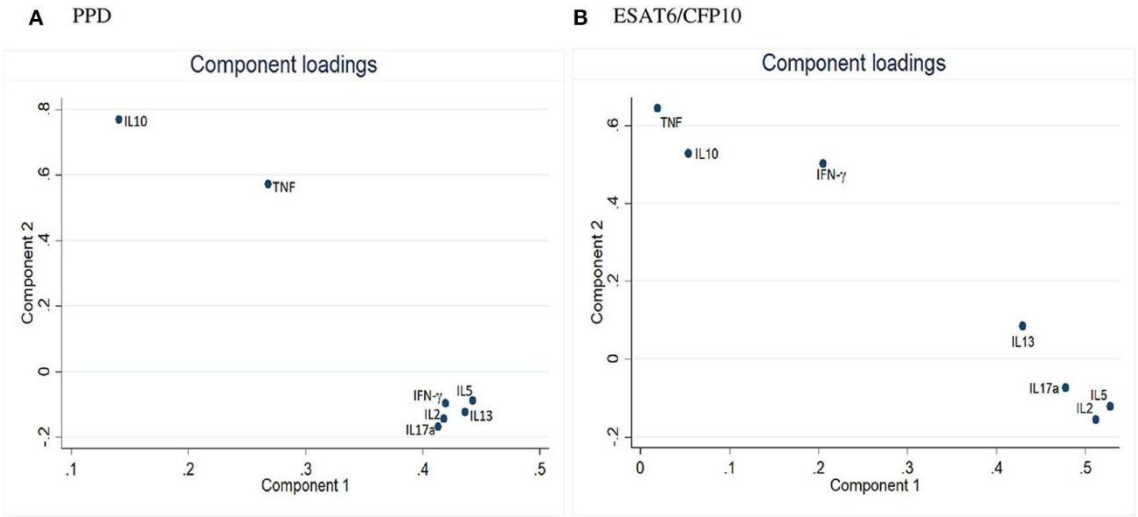


FIGURE 2 | Principal Component Analysis of cord blood outcomes in Ugandan infants. **(A)** PCA for cord blood cytokine responses to PPD. **(B)** PCA for cord blood cytokine responses to ESAT6/CFP10.

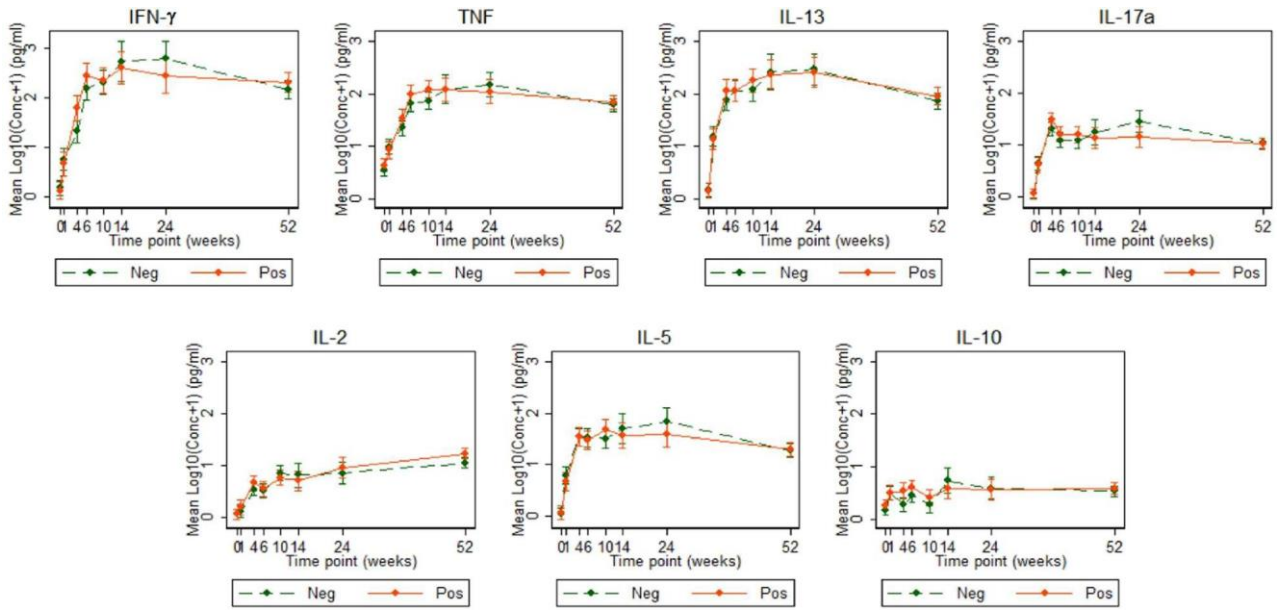


FIGURE 3 | Cytokine responses to PPD in Ugandan infants born of mothers with or without LTBI. PPD-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent 95% confidence intervals around the mean. The solid and dashed lines represent concentrations from children born of LTBI-positive and LTBI-negative mothers, respectively.

PPD and ESAT6/CFP10 responses for the other 10 cytokines/chemokines included in the array (IL-8, MCP- 1, MIP-1α, MIP-1β, IL-1α, IL-1β, IL-12, IL-1Ra, GM-CSF, and IP-10) were also similar between the two infant groups (**Supplementary Figures 1, 2**).

Cord Blood Sensitisation and Infant Cytokine Responses Over Time

There was no difference in evolution of PPD responses based on cord blood response profiles (**Supplementary Figure 3**).

Antibody Responses Over Time

Antibody concentrations were similar, at all-time points, between the two infant groups. Anti-ESAT6 and anti-Ag85A IgG antibody concentrations declined up to 10 weeks and then gradually increased, consistent with maternal antibodies (in both groups) waning and then infants generating their own antibodies. There was no difference in antibody concentrations over time, based on cord blood response profiles. Antibody levels to PPD, Ag85A, and CFP10 were generally higher than those to ESAT6 (**Figure 7**).

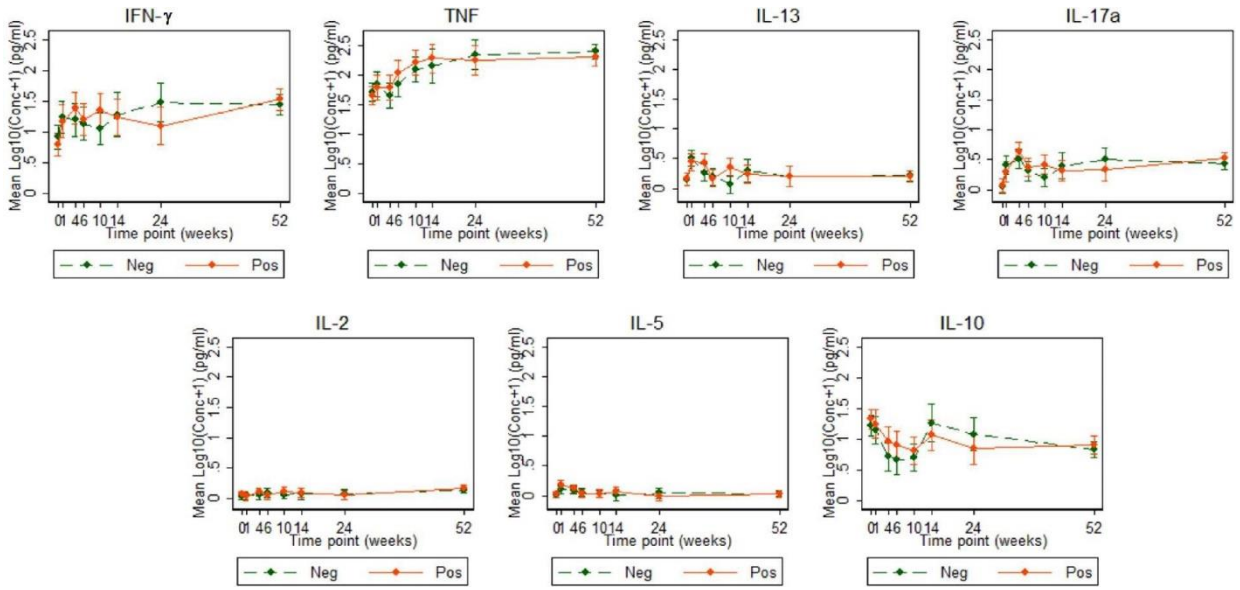


FIGURE 4 | Cytokine responses to ESAT6/CFP10 in Ugandan infants born of mothers with or without LTBI. ESAT6/CFP10-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent the 95% confidence intervals around the mean. The solid and dashed lines represent concentrations from children born of LTBI-positive and LTBI-negative mothers, respectively.

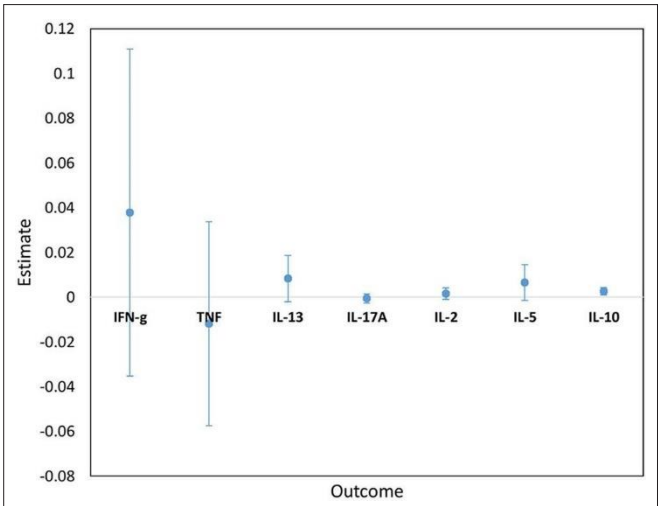


FIGURE 5 | Estimates for the interaction terms between time and LTBI status for cytokine responses to PPD. Estimates and 95% confidence intervals for the interaction term between time and LTBI from univariate linear mixed models for cytokine responses to PPD.

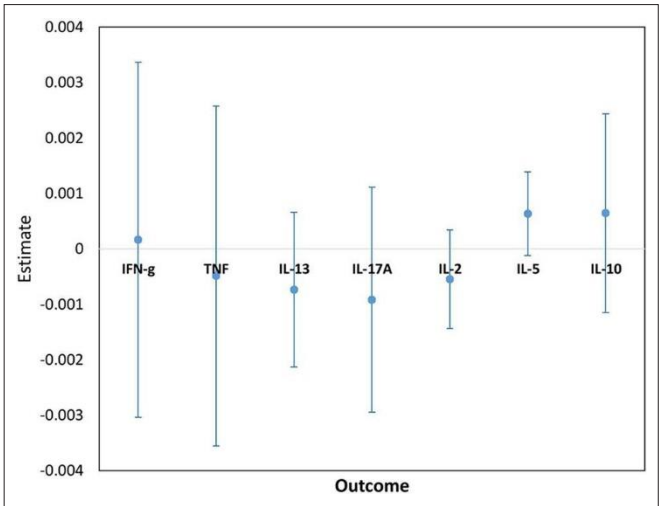


FIGURE 6 | Estimates for the interaction terms between time and LTBI status for cytokine responses to ESAT6/CFP10. Estimates and 95% confidence intervals for the interaction term between time and LTBI from univariate linear mixed models for cytokine responses to ESAT6/CFP10.

Associations Between Cytokine Response Dynamics Over Time

Results from three-way component analyses show that IL- 10, IL-2, and IL-5 responses to PPD evolved in a similar way (**Supplementary Figure 4A**), and TNF and IL-17 responses evolved similarly to each other, whereas IL-13 and IFN-γ responses separated out independently from the other cytokines. Responses to ESAT6/CFP10 clustered for IL-13, IL-5, and IL-17 (**Supplementary Figure 4B**), indicating similar evolution patterns over time; TNF, IL-10, IFN-γ, and IL-2 were separated from the other cytokines indicating unique evolution patterns over time. However, there were no differences in clustering patterns between the groups based on mother's LTBI status.

DISCUSSION

In this study, we show that Ugandan infants were sensitized to mycobacterial antigens *in utero*, regardless of maternal LTBI status, and that maternal LTBI had no effect on evolution of immune responses to BCG over time.

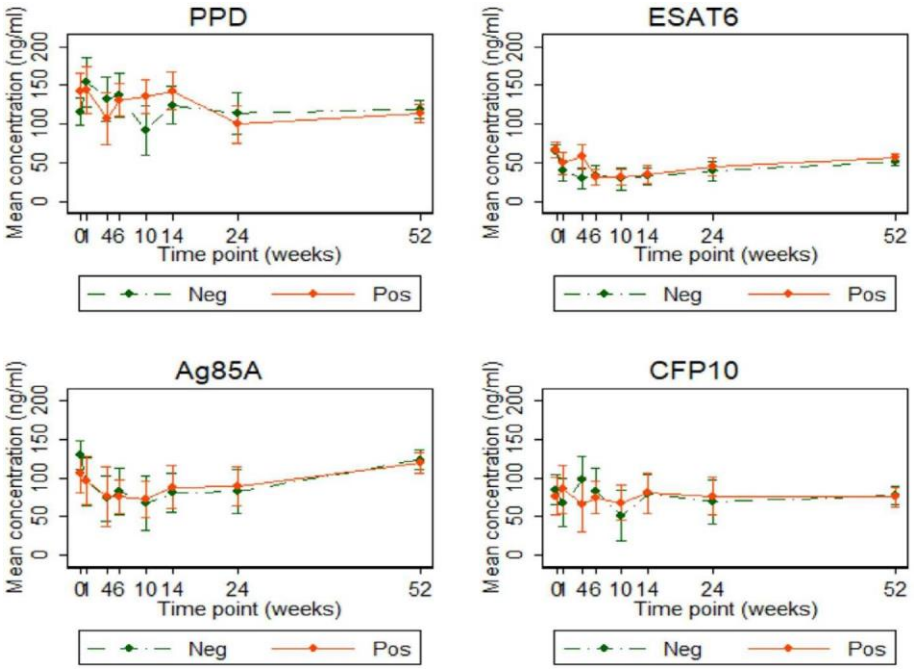


FIGURE 7 | Antibody responses to PPD, ESAT6, Ag85A, and CFP10 in infants born of mothers with or without LTBI. Antibody responses to PPD, ESAT6, Ag85A, and CFP10 (ng/ml). Points represent mean values and the bars represent 95% confidence intervals around the mean. The solid and dashed lines represent concentrations from children born of LTBI-positive and LTBI-negative mothers, respectively.

Maternal infections during pregnancy, with various pathogens, sensitize the fetus *in utero* (14, 18, 32, 33), and prenatal *M.tb* sensitization has long been recognized among mothers from TB-endemic settings (14). In addition, maternal helminth co-infection appears to modify the infant response to subsequent infection with the same pathogen (22), and to BCG immunization (34). In a small preliminary study, we demonstrated impaired mycobacteria-specific T-cell responses following BCG immunization of infants born to LTBI-positive mothers, although this effect was transient (23). We therefore postulated that maternal *M.tb* infection would strongly influence fetal sensitization and the neonatal response to BCG immunization. Our new results refute this hypothesis. A previous study from South Africa showed that maternal HIV infection had an effect on infant immune responses in the presence of maternal *M.tb* sensitization at birth, however, these effects were not maintained post immunization with BCG (35). Our study design differed from the South African study in terms of earlier BCG immunization (at birth, rather than age 6 weeks), larger sample size, exclusion of HIV positive mothers, use of two tests (TST and T-SPOT.TB) rather than one (QuantiFERON-TB Gold In-Tube) to rigorously distinguish LTBI-positive or negative mothers, and more frequent sampling during infancy. Together, the two studies provide clear evidence that maternal LTBI does not impact the infant response to BCG in endemic settings.

PCA in cord blood showed IL-10 and TNF tending to group separately from the other five cytokine responses, but there was no evidence that maternal LTBI was associated with a differing pattern of response, or that differences in these cord blood profiles impacted the subsequent infant response to BCG.

This conclusion is surprising in view of the recognized effects of prenatal exposure to other pathogens on neonatal immune responses. One possible explanation lies in the observation that, in our study, a very high proportion of infants demonstrated cord blood cytokine responses to mycobacterial antigens, regardless of maternal LTBI status. The mechanism of fetal sensitization may involve transfer of either antigen or antibody. Interestingly, antibody to PPD and to ESAT6/CFP10 was found in the cord blood of almost all our study infants, in considerable concentrations, implying that exposure to mycobacteria, whether NTM or *M.tb*, was almost universal. Passive transfer of antibodies appears to have occurred regardless of maternal LTBI status. Nevertheless, how fetal sensitization to mycobacterial antigen occurs in the absence of maternal LTBI needs further investigation.

A second possible explanation for the lack of difference in response between infants of mothers with and without LTBI is that the BCG stimulus is sufficiently strong to override effects of prior *in utero* exposures. Our principal component analysis of cord blood responses suggested distinct groups with IL-10 (potentially suggestive of tolerization) separate from Th1 or Th2 cytokine responses for PPD, and Th1 vs. Th2 biased groups for ESAT6/CFP10. However, these initial sensitization patterns were not reflected in the profile of response that developed following BCG. In addition to driving antigen-specific responses, it is possible that BCG-induced increases in function of innate immunity, through trained immunity, could have contributed

to a lack of difference in responses between infants born of mothers with or without LTBI (36). However, additional analyses using techniques such as intracellular cytokine staining would be needed to assess the extent of cytokine production from innate cells compared to antigen-specific lymphocytes, whether T cells or B cells.

In our study, the peak infant T cell response to PPD was sustained from age 10 to 24 weeks, later than the 6–10 week period previously reported (37). This is important, as the peak for BCG-induced responses should inform future prime- boost strategies regarding timing of the boost—perhaps boosters should be given much later than has previously been thought. Our study differed from the South African study (37) in terms of the geographical location and the use of whole blood stimulation with Luminex assays to measure analytes in supernatant, as compared to flow cytometry.

Comparing UK BCG-vaccinated and un-vaccinated infants, BCG vaccination induced several cytokines and chemokines (IFN-γ, TNF, IL-2, IL-6, IL-1α, IL-4, IL-5, IL-13, IL-10, IL-8, IP-10, MIP-1α, G-CSF, and GM-CSF) (38). Our results are comparable: infants in our study produced similar cytokines, although it was not possible to determine which cells produced them.

This was the largest study of its kind to date. Although recruitment (and hence statistical power) was slightly below target, there was no suggestion of any consistent or persistent differences between the two groups. Thus, the small reduction in power is unlikely to explain the lack of differences seen. Use of a whole blood assay means that we cannot be sure about the cellular source of the analytes measured; for this analysis we have focussed on those most likely to be of T-cell origin; moreover, given that analytes were measured after 6 days of culture.

In conclusion, our data suggest remarkably high early exposure to mycobacterial antigens *in utero* in Uganda, but no impact of this exposure on the infant response to BCG. Our data do not support the hypothesis that prenatal exposure to antigens or antibodies resulting from maternal LTBI results in impaired cytokine responses induced by BCG immunization. The implication of our findings is that maternal LTBI is not the reason for reduced effectiveness of BCG in the tropics and all infants are likely to benefit from BCG immunization regardless of their maternal LTBI status.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Uganda Virus Research Institute Research Ethics Committee (reference GC/127/13/09/16 and GC/127/16/03/434), Uganda National Council for Science and Technology (reference HS 1526) and the London School of Hygiene & Tropical Medicine (reference 7104). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AE, SC, PK, EW, SS, and HD conceived the study and secured funding. AE, SC, SS, HD, EW, LL, and PM reviewed the data and wrote the initial drafts of the manuscript. LL, EW, and JL provided statistical expertise and support. PM, GN, MN, JTus, and MH-A performed all laboratory assays. JS, DA, HA, and JTum were members of the clinical team involved in recruitment, follow up and clinical reporting. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by a project grant from the UK Medical Research Council, grant number MR/K019708. LL was supported by a Ph.D. fellowship through the DELTAS Africa Initiative SSACAB (Grant No. 107754). SC was supported by funds from the DELTAS Initiative MUIIplus (Grant No. 107743). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and was supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the UK Government. The views expressed in this study are those of the authors and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust or the UK government. PM was supported by a Commonwealth Scholarships Commission PhD scholarship (UGCS-2012–602) and a training fund from the MRC/UVRI and LSHTM Uganda Research Unit. SS, MH-A and HD received support from the EU Horizon 2020 programme (TBVAC2020, agreement number 643381) and MH-A was supported by an MRC London Intercollegiate Doctoral Training Partnership studentship. EW received funding from MRC Grant Reference MR/K012126/1; this grant and the MRC/UVRI and LSHTM Uganda Research Unit are jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and are also part of the EDCTP2 programme supported by the European Union.

ACKNOWLEDGMENTS

We thank the study participants who took part in this study, the staff of the Immunomodulation and Vaccines (I-Vac) Programme at the MRC/UVRI and LSHTM Uganda Research Unit and the midwives of the Entebbe General Hospital Maternity Department.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.00929/full#supplementary-material

REFERENCES

1. Trunz BB, Fine P, Dye C. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta- analysis and assessment of cost-effectiveness. *Lancet*. (2006) 367:1173–80. doi: 10.1016/S0140–6736(06)68507–3
2. Abubakar I, Pimpin L, Ariti C, Beynon R, Mangtani P, Sterne J, et al. Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette-Guérin vaccination against tuberculosis. *Health Technol Asses.* (2013) 17:1. doi: 10.3310/hta17370
3. Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, et al. Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *JAMA*. (1994) 271:698–702. doi: 10.1001/jama.1994.03510330076038
4. Fine PE. Variation in protection by BCG: implications of and for heterologous immunity. *Lancet*. (1995) 346:1339–45. doi: 10.1016/S0140–6736(95)92348–9
5. Mangtani P, Abubakar I, Ariti C, Beynon R, Pimpin L, Fine PE, et al. Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. *Clin Infect Dis*. (2013) 58:470–80. doi: 10.1093/cid/cit790
6. Edwards LB, Acquaviva F, Livesay VT, Cross FW, Palmer CE. An atlas of sensitivity to tuberculin, PPD-B, and histosplasmin in the United States. *Am Rev Respir Dis.* (1969) 99:1–132.
7. Weir R, Black G, Nazareth B, Floyd S, Stenson S, Stanley C, et al. The influence of previous exposure to environmental mycobacteria on the interferon-gamma response to bacille Calmette–Guérin vaccination in southern England and northern Malawi. *Clin Exp Immunol*. (2006) 146:390–9. doi: 10.1111/j.1365–2249.2006.03222.x
8. Andersen P, Doherty TM. Learning from BCG: designing a better tuberculosis vaccine. *Discovery medicine*. (2009) 5:383–7. Available online at: http://www.discoverymedicine.com/Peter-Andersen/2009/07/26/learning-from-bcg-designing-a-better-tuberculosis-vaccine/
9. Hoefsloot W, Van Ingen J, Andrejak C, Ängeby K, Bauriaud R, Bemer P, et al. The geographic diversity of nontuberculous mycobacteria isolated from pulmonary samples: an NTM-NET collaborative study. *Eur Respir J*. (2013) 42:1604–13. doi: 10.1183/09031936.00149212
10. Barreto ML, Pilger D, Pereira SM, Genser B, Cruz AA, Cunha SS, et al. Causes of variation in BCG vaccine efficacy: examining evidence from the BCG REVAC cluster randomized trial to explore the masking and the blocking hypotheses. *Vaccine*. (2014) 32:3759–64. doi: 10.1016/j.vaccine.2014.05.042
11. Hesseling A, Cotton M, Fordham von Reyn C, Graham S, Gie R, Hussey, G. Consensus statement on the revised World Health Organization recommendations for BCG vaccination in HIV-infected infants: Submitted on behalf of the BCG Working Group, Child Lung Health Section, International Union Against Tuberculosis and Lung Disease, 38th Union World Conference on Lung Health, Cape Town, 8–12 November 2007 [Official statement]. *Int J Tubercul Lung Dis.* (2008) 12:1376–9. Available online at: http://hdl.handle.net/10019.1/8816
12. Lalor MK, Floyd S, Gorak-Stolinska P, Ben-Smith A, Weir RE, Smith SG, et al. BCG vaccination induces different cytokine profiles following infant BCG vaccination in the UK and Malawi. *J Infect Dis*. (2011) 204:1075–85. doi: 10.1093/infdis/jir515
13. Burl S, Adetifa UJ, Cox M, Touray E, Whittle H, McShane H, et al. The tuberculin skin test (TST) is affected by recent BCG vaccination but not by exposure to non-tuberculosis mycobacteria (NTM) during early life. *PLoS ONE.* (2010) 5:e12287. doi: 10.1371/journal.pone.0012287
14. Malhotra I, Ouma J, Wamachi A, Kioko J, Mungai P, Omollo A, et al. *In utero* exposure to helminth and mycobacterial antigens generates cytokine responses similar to that observed in adults. *J Clin Investig*. (1997) 99:1759–66. doi: 10.1172/JCI119340
15. Sebina I, Biraro IA, Dockrell HM, Elliott AM, Cose S. Circulating B-lymphocytes as potential biomarkers of tuberculosis infection activity. *PLoS ONE.* (2014) 9:e106796. doi: 10.1371/journal.pone. 0106796
16. Kimuda SG, Nalwoga A, Levin J, Franken KL, Ottenhoff TH, Elliott AM, et al. Humoral Responses to Rv1733c, Rv0081, Rv1735c, and Rv1737c DosR Regulon-Encoded Proteins of Mycobacterium tuberculosis in Individuals with Latent Tuberculosis Infection. *J Immunol Res*. (2017) 2017:1593143. doi: 10.1155/2017/1593143
17. Rahman MJ, Degano IR, Singh M, Fernandez C. Influence of maternal gestational treatment with mycobacterial antigens on postnatal immunity in an experimental murine model. *PLoS ONE*. (2010) 5:e9699. doi: 10.1371/journal.pone.0009699
18. Potian JA, Rafi W, Bhatt K, McBride A, Gause WC, Salgame P. Preexisting helminth infection induces inhibition of innate pulmonary anti-tuberculosis defense by engaging the IL-4 receptor pathway. *J Exp Med*. (2011) 208:1863–74. doi: 10.1084/jem.20091473
19. Gebreegziabiher D, Desta K, Desalegn G, Howe R, Abebe M. The effect of maternal helminth infection on maternal and neonatal immune function and immunity to tuberculosis. *PLoS ONE*. (2014) 9:e93429. doi: 10.1371/journal.pone.0093429
20. Novato-Silva E, Gazzinelli G, Colley, D. Immune responses during human schistosomiasis mansoni. XVIII. Immunologic status of pregnant women and their neonates. *Scand J Immunol.* (1992) 35:429–37. doi: 10.1111/j.1365–3083.1992.tb02878.x
21. Holt PG, Strickland DH. Soothing signals: transplacental transmission of resistance to asthma and allergy. *J Exp Med*. (2009) 206:2861–4. doi: 10.1084/jem.20092469
22. Malhotra I, Mungai PL, Wamachi AN, Tisch D, Kioko JM, Ouma JH, et al. Prenatal T cell immunity to Wuchereria bancrofti and its effect on filarial immunity and infection susceptibility during childhood. *J Infect Dis*. (2006) 193:1005–13. doi: 10.1086/500472
23. Mawa PA, Nkurunungi G, Egesa M, Webb EL, Smith SG, Kizindo R, et al. The impact of maternal infection with Mycobacterium tuberculosis on the infant response to bacille Calmette–Guérin immunization. *Philos Trans R Soc B Biol Sci*. (2015) 370:20140137. doi:10.1098/rstb.2014.0137
24. Weir R, Fine P, Nazareth B, Floyd S, Black G, King E, et al. Interferon-γ and skin test responses of schoolchildren in southeast England to purified protein derivatives from *Mycobacterium tuberculosis* and other species of mycobacteria. *Clin Exp Immunol*. (2003) 134:285–94. doi: 10.1046/j.1365–2249.2003.02272
25. Li H, Wang XX, Wang B, Fu L, Liu G, Lu Y, et al. Latently and uninfected healthcare workers exposed to TB make protective antibodies against *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA.* (2017) 114:5023–8. doi: 10.1073/pnas.1611776114
26. Lu LL, Chung AW, Rosebrock TR, Ghebremichael M, Yu WH, Grace PS, et al. A functional role for antibodies in tuberculosis. *Cell*. (2016). 167:433–43. e414. doi: 10.1016/j.cell.2016.08.072
27. Sebina I, Cliff JM, Smith SG, Nogaro S, Webb EL, Riley EM, et al. Long-lived memory B-cell responses following BCG vaccination. *PLoS ONE*. (2012) 7:e51381. doi: 10.1371/journal.pone.0051381
28. Fletcher HA, Snowden MA, Landry B, Rida W, Satti I, Harris SA, et al. T-cell activation is an immune correlate of risk in BCG vaccinated infants. *Nat Commun.* (2016) 7:11290. doi: 10.1038/ncomms11290
29. Nkurunungi G, Lutangira JE, Lule SA, Akurut H, Kizindo R, Fitchett JR, et al. Determining Mycobacterium tuberculosis infection among BCG-immunised Ugandan children by T-SPOT. TB and tuberculin skin testing. *PLoS ONE.* (2012) 7:e47340. doi: 10.1371/journal.pone.0047340
30. Webb EL, Mawa PA, Ndibazza J, Kizito D, Namatovu A, Kyosiimire-Lugemwa J, et al. Effect of single-dose anthelmintic treatment during pregnancy on an infant's response to immunisation and on susceptibility to infectious diseases in infancy: a randomised, double-blind, placebo-controlled trial. *Lancet*. (2011) 377:52–62. doi: 10.1016/S0140–6736(10)61457–2
31. Kroonenberg PM. *Applied Multiway Data Analysis.* Hoboken, NJ: John Wiley and Sons (2008). doi: 10.1002/9780470238004
32. Eloi-Santos S, Novato-Silva E, Maselli V, Gazzinelli G, Colley D, Correa-Oliveira R. Idiotypic sensitization in utero of children born to mothers with schistosomiasis or Chagas' disease. *J Clin Investig*. (1989) 84:1028–31. doi: 10.1172/JCI114225
33. Weil GJ, Hussain R, Kumaraswami V, Tripathy SP, Phillips K, Ottesen EA. Prenatal allergic sensitization to helminth antigens in offspring of parasite-infected mothers. *J Clin Investig*. (1983) 71:1124–9. doi: 10.1172/JCI110862
34. Malhotra I, Mungai P, Wamachi A, Kioko J, Ouma JH, Kazura JW, et al. Helminth-and Bacillus Calmette-Guerin-induced immunity in children sensitized in utero to filariasis and schistosomiasis. *J Immunol*. (1999) 162:6843–8.
35. Jones CE, Hesseling AC, Tena-Coki NG, Scriba TJ, Chegou NN, Kidd M, et al. The impact of HIV exposure and maternal *Mycobacterium tuberculosis* infection on infant immune responses to bacille Calmette-Guérin vaccination. *AIDS.* (2015) 29:155. doi: 10.1097/QAD.0000000000000536

36. Netea MG, Quintin J, Van Der Meer JW. Trained immunity: a memory for innate host defense. *Cell Host Microbe*. (2011) 9:355–61. doi: 10.1016/j.chom.2011.04.006
37. Soares AP, Kwong Chung CK, Choice T, Hughes EJ, Jacobs G, van Rensburg EJ, et al. Longitudinal changes in CD4+ T-cell memory responses induced by BCG vaccination of newborns. *J Infect Dis*. (2013) 207:1084–94.doi: 10.1093/infdis/jis941
38. Lalor MK, Smith SG, Floyd S, Gorak-Stolinska P, Weir RE, Blitz R, et al. Complex cytokine profiles induced by BCG vaccination in UK infants. *Vaccine*. (2010) 28:1635–41. doi: 10.1016/j.vaccine.2009.11.004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

3.3. Supplementary information for Research Paper 1 (Also available in the article's online repository at https://www.frontiersin.org/articles/10.3389/fimmu.2020.00929/full#supplementary-material)

3.3.1. Supplementary tables for Research Paper 1

Supplementary table 1. Number of available samples at each time point for PPD and ESAT6/CFP10 responses

Time point	Responses to PPD			Responses to ESAT6/CFP10		
(weeks)	LTBI-Negative (n=150)	LTBI-Positive (n=132)	Total (n=282)	LTBI-Negative (n=150)	LTBI-Positive (n=132)	Total (n=282)
0	138	116	254	72	78	150
1	71	58	129	41	37	78
4	66	53	119	38	39	77
6	57	56	113	38	41	79
10	57	51	108	41	39	80
14	21	32	53	21	32	53
24	27	29	56	27	29	56
52	110	95	205	109	95	204

Table shows the number of samples assayed at each time point. Half of the infants were expected at time points other than weeks 0 and 52. Weeks 14 and 24 have lower numbers because they were introduced later.

Supplementary table 2. Immune sensitisation based on cord blood responses to ESAT6/CFP10

Cytokines	LTBI-Negative (n = 72)		LTBI-Positive (n = 78)		p-value
Individual cytokines					
IL2	3	4.2%	8	10.3%	0.213
IL5	1	1.4%	2	2.6%	1.000
IL10	60	83.3%	66	84.6%	0.831
IL13	19	26.4%	15	19.2%	0.295
IL17A	6	8.3%	7	9.0%	0.889
TNF	68	95.8%	72	94.7%	1.000
IFN-γ	46	63.9%	44	56.4%	0.350
Based on any of the 7 cytokines					
IL2, IL5, IL10, IL13, IL17A, TNF & IFN-γ	71	98.6%	76	97.4%	0.608
Based on PCA grouping					
TNF & IL10	69	95.8%	75	96.2%	1.000
IL2, IL5, IL13 & IL17A	22	30.6%	23	29.5%	0.887

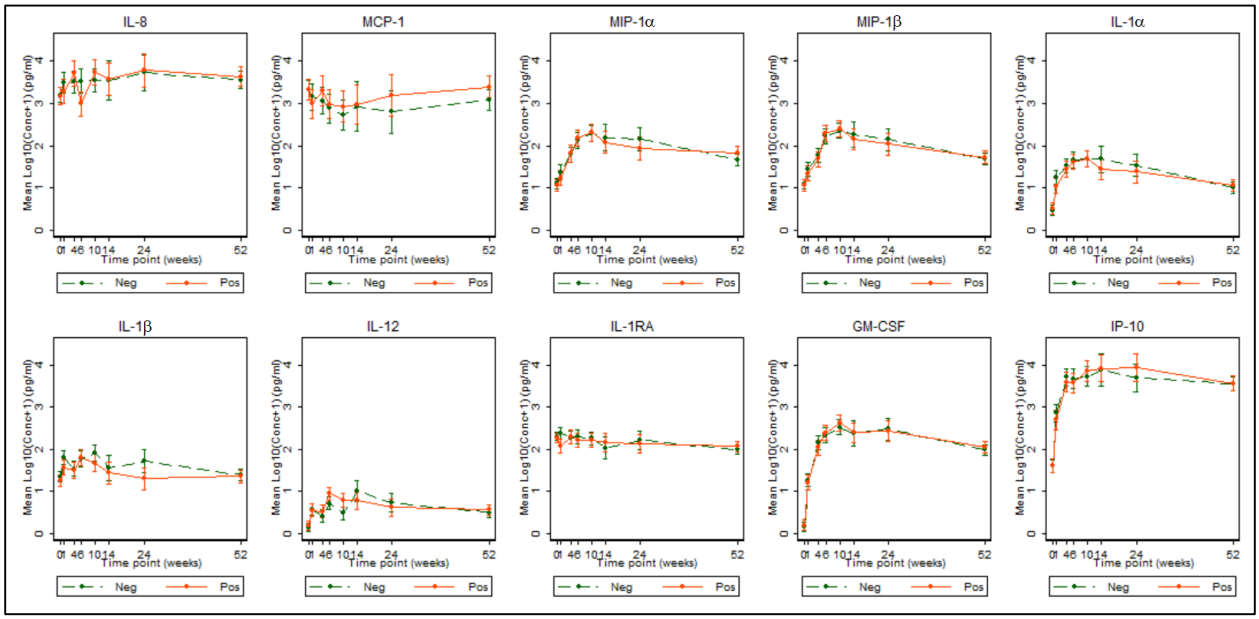
Supplementary table 3. Proportion of infants who developed cytokine responses to PPD at each time point

Cytokines	LTBI-Negative		LTBI-Positive	
Week 1	**(n = 71)**		**(n = 58)**	
IL2	16	22.5%	14	24.1%
IL5	44	62.0%	33	56.9%
IL10	32	45.1%	27	46.6%
IL13	51	71.8%	47	81.0%
IL17A	42	59.2%	33	56.9%
TNF	60	84.5%	44	75.9%
IFN-γ	37	52.1%	31	53.4%
Week 4	**(n = 66)**		**(n = 53)**	
IL2	38	57.6%	37	69.8%
IL5	57	86.4%	43	81.1%
IL10	23	34.8%	28	52.8%
IL13	59	89.4%	49	92.5%
IL17A	59	89.4%	48	90.6%
TNF	60	90.9%	50	94.3%
IFN-γ	45	68.2%	46	86.8%
Week 6	**(n = 57)**		**(n = 56)**	
IL2	34	59.6%	31	55.4%
IL5	52	91.2%	50	89.3%
IL10	25	43.9%	32	57.1%
IL13	53	93.0%	50	89.3%
IL17A	50	87.7%	52	92.9%
TNF	53	93.0%	53	94.6%
IFN-γ	48	84.2%	50	89.3%
Week 10	**(n = 57)**		**(n = 51)**	
IL2	43	75.4%	34	66.7%
IL5	54	94.7%	49	96.1%
IL10	19	33.3%	22	43.1%
IL13	55	96.5%	49	96.1%
IL17A	51	89.5%	48	94.1%
TNF	54	94.7%	49	96.1%
IFN-γ	52	91.2%	46	90.2%
Week 14	**(n = 21)**		**(n = 32)**	
IL2	14	66.7%	24	75.0%
IL5	19	90.5%	30	93.8%
IL10	16	76.2%	25	78.1%
IL13	20	95.2%	32	100.0%
IL17A	19	90.5%	29	90.6%
TNF	20	95.2%	31	96.9%
IFN-γ	20	95.2%	32	100.0%
Week 24	**(n = 27)**		**(n = 29)**	
IL2	24	88.9%	25	86.2%
IL5	26	96.3%	28	96.6%
IL10	20	74.1%	19	65.5%
IL13	26	96.3%	29	100.0%

IL17A	26	96.3%	27	93.1%
TNF	27	100.0%	29	100.0%
IFN-γ	26	96.3%	29	100.0%
Week 52	**(n = 110)**		**(n = 95)**	
IL2	92	83.6%	84	88.4%
IL5	92	83.6%	78	82.1%
IL10	66	60.0%	59	62.1%
IL13	102	92.7%	87	91.6%
IL17A	91	82.7%	80	84.2%
TNF	106	96.4%	91	95.8%
IFN-γ	101	91.8%	89	93.7%

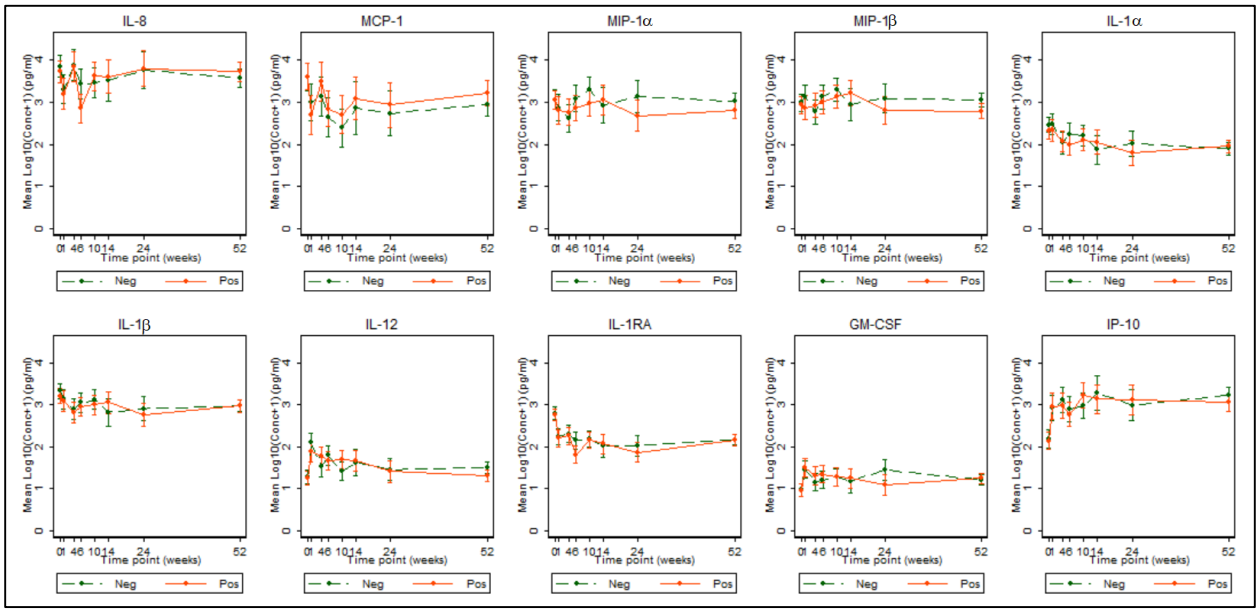
3.3.2. Supplementary figures for Research Paper 1

Supplementary Figure 1. Cytokine responses to PPD in infants born of mothers with or without LTBI



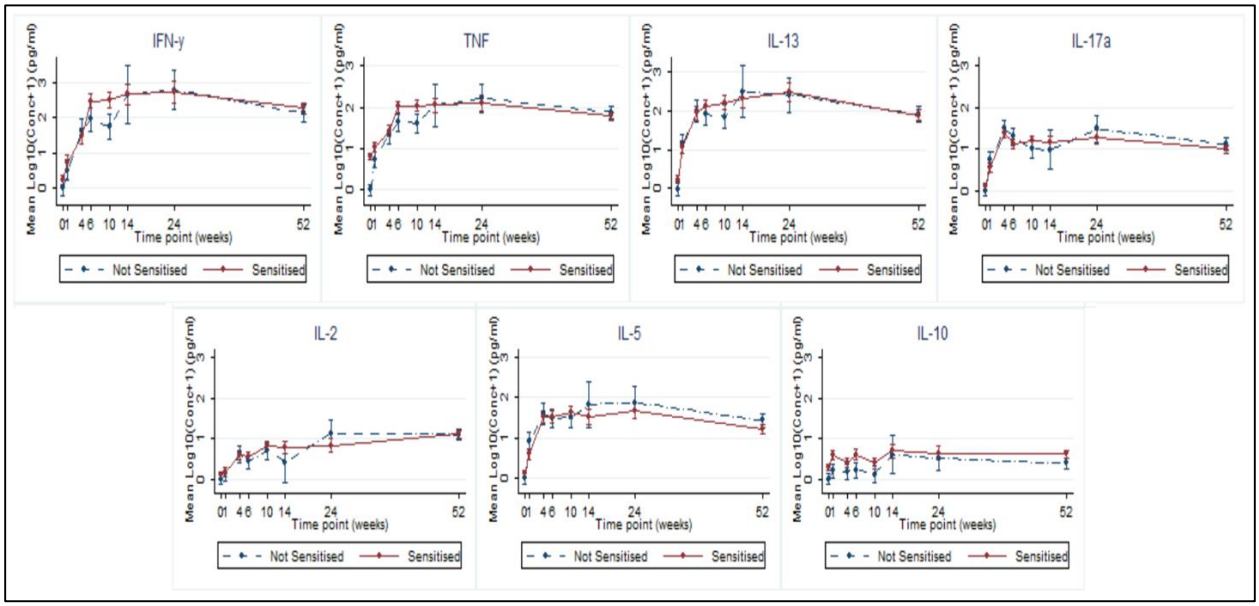
PPD-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent the 95% confidence intervals around the mean. The solid and dashed lines represents concentrations from children born of LTBI-positive and LTBI-negative mothers respectively.

Supplementary Figure 2. Cytokine responses to ESAT6/CFP10 in infants born of mothers with or without LTBI



ESAT6/CFP10-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent the 95% confidence intervals around the mean. The solid and dashed lines represent concentrations from children born of LTBI-positive and LTBI-negative mothers respectively.

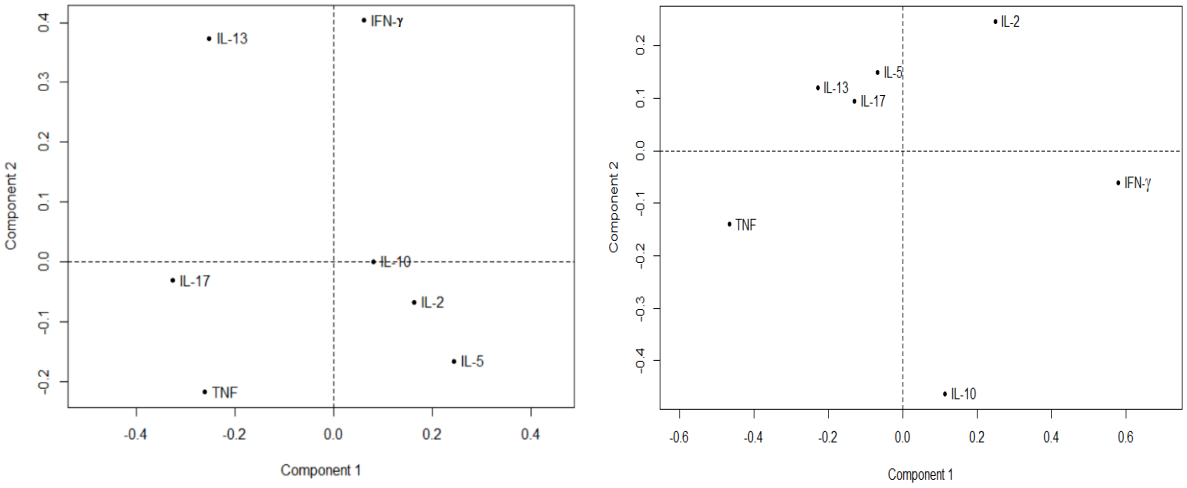
Supplementary Figure 3. Cytokine responses to PPD in infants based on cord blood sensitisation profiles



PPD-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent 95% confidence intervals around the mean. The solid and dashed lines represents concentrations from infants who were immune sensitised or not based on cord blood responses.

Supplementary Figure 4. Three-way component analysis of cytokine responses to PPD and ESAT6/CFP10, following immunisation with BCG, in Ugandan infants

A. PPD **B. ESAT6/CFP10**



A) Component loadings for cytokine responses to PPD. B) Component loadings for cytokine responses to ESAT6/CFP10.

Chapter 4. A pairwise joint modelling approach for the analysis of longitudinal immuno-epidemiological data

4.1. Introduction

This chapter illustrates an extension of statistical modelling methods for the analysis of longitudinal multiple outcome data, with application to the IBS case study (**thesis objective 1**). A pairwise joint modelling approach is discussed and its benefits over simpler statistical analysis approaches are highlighted. Missing data (**thesis objective 3**) are handled via a likelihood based approach under the MAR assumption. Results are presented in Research paper 2, entitled "Analysis of multivariate longitudinal immuno-epidemiological data using a pairwise joint modelling approach", this manuscript is under review at the BMC Immunology journal.

Patterns of post-BCG cytokine responses to mycobacterial antigens are compared between infants of mothers with or without LTBI, using a statistical modelling approach which accounts for both the correlation between measurements from an individual over time and the correlation between multiple outcomes assessed at each time point. Research paper 2 extends the univariate models of paper 1 to a situation where all outcomes can be modelled jointly, providing an opportunity for a test for the effect of LTBI on the joint evolution of all, or groups of, outcomes and the estimation of association structures of the outcomes.

4.2. Research Paper 2: Analysis of multivariate longitudinal immuno-epidemiological data using a pairwise joint modelling approach

Analysis of multivariate longitudinal immuno-epidemiological data using a pairwise joint modelling approach

Lawrence Lubyayi[1,2*], Patrice A Mawa[2,3], Stephen Cose[2,4], Alison M Elliott[2,4], Jonathan Levin[1], Emily L Webb[5]

[1]Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa

[2]Immunomodulation and Vaccines Programme, Medical Research Council/Uganda Virus Research Institute and London School of Hygiene and Tropical Medicine Uganda Research Unit Entebbe, Uganda

[3]Uganda Virus Research Institute, Entebbe, Uganda

[4]Department of Clinical Research, London School of Hygiene and Tropical Medicine, United Kingdom

[5]MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, United Kingdom

***Corresponding author:** Lawrence Lubyayi, MRC/UVRI and LSHTM Uganda Research Unit, Plot 51-59 Nakiwogo Road, P. O. Box 49 Entebbe Uganda. Email: lawrence.lubyayi@mrcuganda.org, lawrencelby@gmail.com

ABSTRACT

Background: Immuno-epidemiologists are often faced with multivariate outcomes, measured repeatedly over time. Such data are characterised by complex inter- and intra-outcome relationships which must be accounted for during analysis. Scientific questions of interest might include determining the joint effect of a treatment on the evolution of all outcomes together, or grouping outcomes that change in the same way. Modelling the different outcomes separately may not be appropriate because it ignores the underlying relationships between outcomes. In such situations, a joint modelling strategy is necessary. This paper describes a pairwise joint modelling approach and discusses its benefits over more simple statistical analysis approaches, with application to data from a study of the response to BCG vaccination in the first year of life, conducted in Entebbe, Uganda.

Methods: The study aimed to determine the effect of maternal latent *Mycobacterium tuberculosis* infection (LTBI) on infant immune response (TNF, IFN-γ, IL-13, IL-10, IL-5, IL-17A and IL-2 responses to PPD), following immunisation with BCG. A simple analysis ignoring the correlation structure of multivariate longitudinal data is first shown. Univariate linear mixed models are then used to describe longitudinal profiles of each outcome, and are then combined into a multivariate mixed model, specifying a joint distribution for the random effects to account for correlations between the multiple outcomes. Due to model fitting complexity because of the high number of outcomes, a pairwise joint modelling approach, where all possible pairs of bivariate mixed models are fitted, is used to obtain parameter estimates.

Results: Both univariate and pairwise longitudinal analysis approaches are consistent in finding that LTBI had no impact on the evolution of cytokine responses to PPD. Estimates from the pairwise joint modelling approach were more precise. Major advantages of the pairwise approach include the

opportunity to test for the effect of LTBI on the joint evolution of all, or groups of, outcomes and the ability to estimate association structures of the outcomes.

Conclusions: The pairwise joint modelling approach reduces the complexity of analysis of high-dimensional multivariate repeated measures, allows for proper accounting for association structures and can improve our understanding and interpretation of longitudinal immuno-epidemiological data.

Key words: latent *Mycobacterium tuberculosis* infection, BCG vaccine, cytokine responses, linear mixed model, pairwise joint modelling

INTRODUCTION

Longitudinal studies are indispensable for the investigation of changes in outcomes over time. Through making measurements on study participants over time, longitudinal studies allow the direct study of temporal changes within individuals and the factors that influence these changes [1]. Since the study of change is fundamental to almost every discipline, there has been a steady growth in the number of studies using longitudinal designs [1]. However, many challenges arise when analysing data from longitudinal studies [2]. Naturally, the repeated measures arising from longitudinal studies are multi-dimensional and have a complex random-error structure that must be appropriately accounted for in the analysis [1]. Problems of missing data and attrition are also common in these studies; yet appropriate handling of missing data continues to pose one of the greatest challenges in their analysis [1, 3, 4]. These, and many other issues, increase the complexity of longitudinal data analysis, and this is particularly the case for immuno-epidemiological studies. Immuno-epidemiological studies investigate the influence of population immunity on the epidemiology of conditions such as infectious diseases, cancer, hypersensitivity and autoimmunity [5, 6]. Such studies are likely to have a large number of, often correlated, outcomes measured repeatedly over time.

The understanding of the importance of immunological mechanisms underlying human disease has grown enormously in the recent past [7]. However, the complexity of relationships between large numbers of immunological parameters poses the challenge of how to select the most appropriate statistical techniques to extract the maximum relevant information from complex datasets and to avoid spurious findings [7]. In many immuno-epidemiological studies, simple statistical approaches are applied even when complex patterns of inter-relationships between parameters are expected [7, 8]. A number of articles have given an overview of application of statistical techniques to immuno-epidemiological data [7-10], however these focus mainly on cross-sectional data. Published guidance on the analysis of

longitudinal immunological data is limited [11], particularly for longitudinal immunological data with multivariate outcomes.

In immuno-epidemiological studies, a number of scientific questions might be of interest, for instance to determine the effect of an intervention or exposure on the joint evolution of all outcomes together, or to categorise outcomes that change in the same way. Modelling the different outcomes separately may not be appropriate because it ignores the underlying relationships between them. In such situations, a joint modelling strategy is necessary [12]. This paper describes methods that can be applied in this context, with emphasis on the pairwise joint modelling approach and its benefits over more simple statistical analysis approaches. Approaches are demonstrated using data from the Infant BCG Study (IBS) [13] which was carried out in Entebbe, Uganda.

METHODS

The Infant BCG Study (IBS) Data

The IBS was an observational longitudinal immuno-epidemiological study, which enrolled infants born to mothers with (n=132) or without (n=150) latent *Mycobacterium tuberculosis (M.tb)* infection (LTBI) and followed them up for one year. Healthy mothers and their infants were recruited at Entebbe General Hospital between June 2014 and October 2016. Up to 7ml of cord blood was collected and blood samples were taken at selected time points throughout infancy. Infants were randomly assigned in a 1:1 ratio (stratified by LTBI status) to two sampling strategies to reduce the blood-sampling burden on individual infants. Half gave 2ml venous blood at 1, 6, and 14 weeks, the remainder at 4, 10, and 24 weeks. All gave blood (5ml) at 52 weeks. All infants were BCG immunised at birth or within the first week of life with BCG (Statens Serum Institut (SSI), Denmark). Immunological parameters measured included 7 cytokine/chemokine responses (IL-2, IL-5, IL-10, IL-13, IL-17A, TNF, and IFN-γ) to the *M.tb* purified protein derivative (PPD). These were assessed by Luminex in 6-day whole blood cultures, in cord

blood and at weeks 1, 4, 6, 10, 14, 24 and 52. The primary aim of the study was to determine the effect of prenatal exposure to LTBI on the infant immune response following immunisation with BCG [13]. Additional questions of interest were (i) the strength of association between the evolutions of cytokine responses over time, (ii) the effect of LTBI on the joint evolution of cytokine responses over time, and (iii) whether the relationship between cytokine responses differed comparing pre- and post-BCG time points. All these questions necessitate a joint modelling strategy.

Participants' baseline characteristics

Baseline characteristics of participants were summarised, by LTBI status, using percentages, means and standard deviations, and medians and interquartile ranges.

Simple analyses ignoring correlations between time points and outcomes

Simple analyses often reported for such cytokine response data include t-tests if the distributions are approximately normal and non-parametric alternatives such as Mann-Whitney tests when the distributional assumptions are relaxed. Such tests are usually conducted separately for individual cytokine responses and separately at each time point. Adopting this simple approach, our initial analyses employed Mann-Whitney tests comparing responses between infants born of mothers with or without LTBI. Bonferroni corrections were applied to adjust for multiple comparisons at each time point.

Graphical exploration of longitudinal profiles

Individual profile plots were constructed for each of the seven longitudinal outcomes to obtain insight into how infant responses evolved over time as well as to give an indication of between and within infant variability. When this kind of variability is present it provides the motivation for modelling approaches which take this into consideration, most commonly via the specification of random effects.

Average profile plots were then constructed, for each outcome, to describe the mean evolution of infant

responses, disaggregated by maternal LTBI status. These plots give an indication of the functional form of the evolution and an initial idea of whether this evolution differs by maternal LTBI status.

Analyses allowing for longitudinal data but ignoring correlations between different outcomes

Changes in log-transformed cytokine responses over time, by mother's LTBI status, were studied using univariate linear mixed models (LMM) adjusted for factors that showed baseline differences between the two groups.

The use of the linear mixed model for analysing univariate longitudinal data has been discussed extensively [3, 14-17]. The model handles continuous longitudinal data in an easy, valid and flexible manner [18] and can be used for data with an unequal number of measurements per subject [14]. The model is defined as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i\,; \quad \boldsymbol{b}_i \sim N(\boldsymbol{0}, D); \quad \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i); \quad \boldsymbol{b}_i, \boldsymbol{\varepsilon}_i \; are \; independent$$

where $\boldsymbol{Y}_i$ is the n_i dimensional response vector for the i^{th} subject, $1 \leq i \leq N$, N is the number of subjects, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are (n_i x p) and (n_i x q) dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects, $\boldsymbol{b}_i$ is a q-dimensional vector of random effects, $\boldsymbol{\varepsilon}_i$ is an n_i-dimensional vector of residual components, D is a covariance matrix of random effects and $\boldsymbol{\Sigma}_i$ is a covariance matrix of residuals [14]. Random effects represent an aggregation of all unobserved or unmeasured factors that make individuals respond differently to each other [17].

Longitudinal immuno-epidemiological data are often characterised by non-linear patterns over time. Such patterns can still be handled under the linear mixed effects models framework with power transformations of time. Fractional polynomials (FPs) are characterised by power terms which can be negative values and/or fractions. Conventional polynomials (CPs) are a special case of FPs with power terms having only integer values. FPs have been shown to have more favourable characteristics than

higher order CPs when modelling non-linear growth curves within the context of the linear mixed model [19-21]. The R package 'mfp' [22] was used to determine the best fitting FP for each outcome. Graphics and criteria such as the Akaike Information Criteria (AIC), Deviance, R^2 and adjusted R^2 were used to compare the best fitting FP to the best higher order CP (which had been chosen using graphical means).

Under the LMM framework, likelihood based tests, specifically Restricted Maximum Likelihood (REML), were used to check the need for inclusion of various serial correlation structures (simple, autoregressive, compound symmetry, unstructured, exponential, power and gaussian). Mixtures of chi-square distributions were then used to assess the need for extending the random effects structures. Finally, to discover the most parsimonious mean structure, likelihood ratio tests were employed under maximum likelihood estimation.

Analyses allowing for longitudinal data and correlations between outcomes

Patterns of correlation between the different outcomes were expected, for instance, certain cytokines are associated with particular cell types or functions (such as T-helper (Th)1, Th2 or regulatory functions). A statistical modelling approach which accounts for such correlation structures was necessary, to provide additional insight into the data. The seven univariate linear mixed models were thus combined into a full multivariate model resulting in a 7 x 7 covariance matrix for random effects (if only random intercepts are considered) and a 7 x 7 covariance matrix for error components, a total of 56 covariance parameters. This dimensionality of random effects is too large for standard software packages to handle.

The pairwise joint modelling approach was introduced as a novel procedure for fitting such random effects models without restricting the dimensionality [23]. The general idea is that all parameters in the full multivariate model can be identified from fitting all bivariate models for each pair of outcomes separately, then, afterwards, estimates are averaged to obtain one single estimate for each parameter

of the full joint model. For standard errors, in order to correctly calculate the sampling variability of the estimates from the pairwise approach, pseudo-log-likelihood estimation is applied [23, 24]. This approach was applied to our data in order to properly account for and to study the strength of association between the evolutions of cytokine responses over time. Parameter estimates from this approach were used to construct a Wald-type test statistic for the joint effect of maternal LTBI on evolution of all outcomes together. Principal components analysis (PCA) was then carried out on the 7 x 7 correlation matrix of random effects and the results were compared to PCA of cord blood cytokine responses to establish whether the clustering of responses before immunisation with BCG differed from the clustering of responses after BCG. A SAS macro [25] was adapted and used to implement the pairwise joint modelling approach.

Data analysis was conducted using SAS version 9.4 (SAS Institute, Cary NC, USA), Stata 15.0 (College Station, Texas, USA) and R version 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria). A 5% significance level was used for all analyses.

RESULTS

Participants' baseline characteristics

Between June 2014 and October 2016, the IBS enrolled infants born to mothers with (n=132) or without (n=150) LTBI and followed them up for one year. Baseline characteristics of participants are shown in Table 1 by LTBI status. LTBI-positive mothers were on average older, more likely to have lived with someone who had TB, to drink alcohol and to originate from the central region of Uganda. Infant characteristics (sex and birth weight) and other maternal characteristics were similar between the two groups.

Table 1. Baseline characteristics of study participants

	LTBI Negative (n=150)		LTBI Positive (n=132)	
Maternal characteristics				
Mean mother's age (SD) (mv 0, 3)[a]	23.65	(3.67)	25.53	(4.99)
Median number of pregnancies (IQR) (mv 1, 1)	2	(1-3)	2	(2-4)
Positive malaria test during pregnancy (mv 1, 1)	38	25.5%	29	22.1%
Ever lived with someone with TB (mv 2, 1)	3	2.0%	19	14.5%
BCG scarring (mv 1, 2)	106	71.1%	95	73.1%
Current marital status (mv 2, 4)				
Single	28	18.9%	20	15.6%
Married/living as married	120	81.1%	108	84.4%
Drink alcohol (mv 3, 4)	18	12.2%	28	21.9%
Mother's tribe grouping (mv 5, 3)				
Central	56	38.6%	72	55.8%
Other	89	61.4%	57	44.2%
Father's tribe grouping (mv 3, 2)				
Central	58	39.5%	75	57.7%
Other	89	60.5%	55	42.3%
Infant characteristics				
Sex of the baby, male	77	51.3%	77	58.3%
Mean birth weight in kg (SD)	3.24	(0.43)	3.21	(0.40)

Data are mean (SD), median (IQR), or n (%). SD=standard deviation, IQR=interquartile range, mv=missing values.

[a]Figures in parentheses indicate missing values in the LTBI-Negative and LTBI-Positive groups, respectively.

Due to the study design, not all infants provided samples at all time points, sample numbers assayed at each time point are shown (Additional file 1: Table S1).

Simple analyses ignoring correlations between time points and outcomes

Table 2 shows unadjusted p-values from the Mann-Whitney test for the comparisons of cytokine responses between infants born to mothers with or without LTBI at each time point for all the seven outcomes. Statistically significant differences between the two groups are highlighted at week 4 (for cytokines IFN-γ, IL-17A, and IL-10), week 10 (for TNF) and at week 24 (for IFN-γ and IL-17A). When a Bonferroni correction is applied at each of these time points, none of the differences remains significant.

Table 2. P-values from Mann-Whitney tests for the comparison of cytokine responses to PPD between the two infant groups

Outcome	Cord blood	Week1	Week4	Week6	Week10	Week14	Week24	Week52
TNF	0.060	0.697	0.137	0.096	**0.032**	0.589	0.253	0.377
IFN-γ	0.891	0.691	**0.026**	0.203	0.703	0.299	**0.055**	0.121
IL-2	0.982	0.643	0.400	0.948	0.629	0.601	0.743	0.071
IL-5	0.729	0.430	0.864	0.750	0.147	0.472	0.156	0.688
IL-13	0.888	0.915	0.351	0.715	0.178	0.620	0.496	0.375
IL-17A	0.353	0.852	**0.047**	0.408	0.144	0.418	**0.040**	0.841
IL-10	0.056	0.959	**0.014**	0.187	0.171	0.293	0.765	0.538

Unadjusted p-values for the comparisons of responses between infants born of LTBI positive or negative mothers at each time point. Values in bold are significant at the 5% level (without considering the Bonferroni correction for multiple testing)

Graphical exploration of longitudinal profiles

Figure 1 shows the individual profiles for a consecutive sample of 30 infants, for each of the seven cytokine responses to PPD. It is evident from the figure that the infants have highly variable concentrations at the start, this suggests that perhaps linear mixed models with random intercepts could be an appropriate modelling approach.

Figure 2 shows the average evolutions of the seven cytokine responses over time, stratified by mothers' LTBI status. It can be seen that in general there was a sharp increase up to about 10 weeks and then a plateauing through to 52 weeks, with some curvature which should be appropriately considered during model building. These patterns of evolution point to the consideration of fractional polynomials for modelling the mean structure of the outcomes. It is also seen that cytokine responses to PPD were similar, at all time points, between the two infant groups.

Analyses allowing for longitudinal data but ignoring correlations between different outcomes

First or second order FPs, providing the best fit for each of the seven outcomes, were identified using the R function 'mfp'. Both TNF and IFN-γ responses were best represented by second order FPs with powers m1=0.5 and m2=0.5, IL-13 with powers m1=-0.5 and m2=3, IL-17A with powers m1=-2 and m2=-

2, IL-5 with powers m1=-1 and m2=3. IL-2 was best represented by a first order FP with m1=0 (equivalent to a log transformation of time) while IL-10 was best represented by a linear function of time. For all the seven outcomes, FP models had better fit criteria (lower AIC and Deviance and higher R^2) than conventional higher order polynomials (Additional file 1: Table S2). Graphical comparisons of FPs and higher order CPs show better performance for the FP models especially at the extreme ends (Additional file 1: Figures S1A and S1B).

Univariate linear mixed models were fitted for each outcome, using the identified FP functions to capture the curvature over time. Random intercepts were included in each of the seven models to account for the variability between different infants; extending the random effects structure to linear slopes of time was not necessary (based on the mixture of chi-square tests). Each model was adjusted for sex of the infant and for factors that showed baseline differences between the two groups (mother's age, household TB contact, alcohol consumption, central region origin). There was no evidence of interaction between time and maternal LTBI status for any of the seven outcomes (Figure 3).

Analyses allowing for longitudinal data and correlations between outcomes

Table 3 shows the parameter estimates and standard errors for the interaction effects of time and mother's LTBI status, for all the seven outcomes, obtained from the pairwise approach. A joint statistical test of the effect of LTBI on the evolutions of all outcomes together yielded a Wald-type test statistic with a chi-square value of 11.04, which was not significant when compared to the chi-square distribution with 7 degrees of freedom (p-value = 0.137), implying that LTBI had no effect on the evolution of all outcomes (when considered jointly).

Table 3. Estimates and standard errors from pairwise model

Outcome	Estimate	Std.Error
IFN-γ	-0.00227	0.0025
TNF	-0.00116	0.0016
IL-13	-0.00002	0.0022
IL-17A	-0.00265	0.0015
IL-2	0.00231	0.0019
IL-5	0.00102	0.0021
IL-10	-0.00264	0.0017

Parameter estimates and standard errors for the interaction effects of time and mother's LTBI status obtained from the pairwise joint modelling approach.

Figure 3 also shows the interaction effects of time and mothers' LTBI status, from the pairwise approach. The message is consistent with that from the univariate models, but with an added benefit of improved precision indicated by smaller 95% confidence intervals.

Figure 4A shows results of a PCA on the 7 x 7 correlation matrix of random intercepts (Additional file 1: Table S3). The result indicates that IL-5 and IL-17A responses evolved in a similar way, TNF, IFN-γ, IL-10 and IL-13 responses evolved similarly to each other, whereas IL-2 responses evolved uniquely from the other cytokines. Figure 4B shows the component loadings for the seven outcomes, based on cord blood responses. Comparison with figure 4A indicates that the association structure of the infant cytokine responses changed after the infants were immunised with BCG.

DISCUSSION

This paper describes a pairwise joint modelling approach and discusses its benefits over more simple statistical analysis approaches, commonly used for immuno-epidemiological data, with application to data from the Infant BCG Study. It demonstrates that certain scientific questions cannot be addressed by simple approaches. The pairwise approach is shown to be essential in situations where it is desired to determine the effect of a treatment on the joint evolution of all outcomes and when it is desired to study the relations between evolutions of longitudinal multivariate outcomes.

Simple statistical analysis approaches such as the Mann-Whitney test are often used for multivariate longitudinal immuno-epidemiological data [7, 8]. These ignore the correlation between repeated measurements over time and cannot be used to study the changes that happen between correlated outcomes over time. However, these simple approaches are quite often erroneously interpreted to show changes over time yet they ought to be interpreted as analyses at the separate time points.

Univariate linear mixed models are shown in this study. These provide an improvement over the simple approaches, for handling of longitudinal data. They handle continuous longitudinal data in an easy, valid and flexible manner [18], account for correlation between repeated measurements from individuals and can be valid depending on the nature of research question at hand. However, they analyse each outcome separately, ignoring the correlation between multiple outcomes assessed over the various time points.

The focus of this paper is on a more appropriate statistical analysis approach for multivariate longitudinal immuno-epidemiological data, that accounts for both the correlation between measurements from an individual over time and also the correlation between the multiple outcomes assessed at each time point. With this approach, the seven longitudinal outcomes from the IBS data were jointly modelled, considering a random intercept for each outcome. This led to a covariance structure with 56 parameters. Fitting a full multivariate model with maximum likelihood estimation was not possible when four or more of the seven outcomes were considered. This task was broken down into the fitting of 21 bivariate models via the pairwise approach as described [23, 24].

Our results show that the parameter estimates from the pairwise approach had better precision than those from the univariate mixed models. This could be attributed to the fact that the pairwise approach accounts for more variability (correlation of outcomes). It has been emphasized elsewhere, though, that

gains in precision compared to univariate models should not be the key motivation for choosing the pairwise approach [24].

A major advantage of the pairwise approach is the opportunity to carry out a joint statistical test for the effect of LTBI on the evolution of all cytokine response outcomes together. This overcomes the multiple testing problem inherent with multiple outcome data measured at multiple time points, and can only be done under a full multivariate modelling approach, and not under any of the other simple or univariate approaches. This supported the conclusion that there was no difference between infants born of mothers with or without LTBI.

Another benefit of the pairwise approach, in our case, lies in the ability to estimate the association structures of the longitudinal outcomes and how these relate to each other. This can improve our understanding and interpretation of longitudinal immuno-epidemiological data. In our case, the PCA of cord blood responses suggested distinct groups with IL-10 separate from Th1 or Th2 cytokine responses. However, these initial patterns were not reflected in the profile of response that developed following immunisation with BCG as indicated by the PCA on the correlation matrix of random intercepts. Biologically, this suggests that the effect of BCG immunisation was potent enough to over-ride patterns established in *utero*, however, in the absence of a non-BCG control group this remains unconfirmed.

A limitation of our study could be in the potential influence of missing data. The mixed models used have an advantage of accommodating unbalanced data structures but they assume that data are missing at random. This assumption is inherently untestable however, and sensitivity analyses under alternative assumptions are recommended.

Conclusion

The pairwise joint modelling approach for multivariate longitudinal data has utility for immuno-epidemiological data. It reduces the complexity of analysis of multivariate repeated measures of a

relatively high dimension, while still accounting for association structures, thus providing an improvement over the simple univariate approaches in common use. The proposed approach can improve our understanding and interpretation of longitudinal immuno-epidemiological data.

Additional file

Additional file 1: **Table S1**. Number of available samples at each time point for responses to PPD. **Table S2**. Comparison of Fractional polynomial and conventional higher order polynomials using various fit-statistics. **Table S3**. Correlation matrix of random effects from the pairwise model. **Figure S1A**. Observed and fitted mean functions for the cytokine responses to PPD. **Figure S1B**. Observed and fitted mean functions for the cytokine responses to PPD. (DOCX 441 KB)

Abbreviations

BCG: Bacille Calmette-Guerine; IBS: Infant BCG Study; *M.tb*: *Mycobacterium tuberculosis*; LTBI: latent *Mycobacterium tuberculosis* infection; TNF: Tumor Necrosis Factor; IFN: Interferon; IL: Interleukin; PPD: Purified Protein Derivative; Th: T-helper; FPs: Fractional polynomials; CPs: Conventional polynomials; LMM: Linear Mixed Model; AIC: Akaike Information Criteria; REML: Restricted Maximum Likelihood; PCA: Principal Components Analysis

DECLARATIONS

Ethics approval and consent to participate

Ethics approval was given by the Uganda Virus Research Institute Research Ethics Committee (reference GC/127/13/09/16 and GC/127/16/03/434), Uganda National Council for Science and Technology (reference HS 1526) and the London School of Hygiene & Tropical Medicine (reference 7104). Written informed consent was received from all mothers, for their own and their baby's participation.

Availability of data and materials

The data sets analysed in this study are available from the corresponding author on reasonable request.

Authors' contributions

LL, AME, JL and ELW conceived this study. AME, PM, ELW and SC were key investigators on the IBS. LL analysed the data with support from AME, JL and ELW. LL wrote the initial draft of the manuscript. All authors read and approved the final version of the manuscript.

Funding

The IBS was funded by a project grant from the UK Medical Research Council, grant number MR/K019708. LL is supported by a PhD fellowship through the DELTAS Africa Initiative SSACAB (Grant No. 107754). SC was supported by funds from the DELTAS initiative MUIIplus (Grant No. 107743). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the UK Government. The views expressed in this study are those of the authors and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust or the UK government. PM was supported by a Commonwealth Scholarships Commission PhD scholarship (UGCS-2012-602). ELW received funding from MRC Grant Reference MR/K012126/1; this grant and the MRC/UVRI and LSHTM Uganda Research Unit are jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union.

Acknowledgements

We thank the study participants who took part in this study, the staff of the Immunomodulation and Vaccines (I-Vac) Programme at the MRC/UVRI and LSHTM Uganda Research Unit and the midwives of the Entebbe General Hospital Maternity Department.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

References

1. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G: **Advances in longitudinal data analysis: an historical perspective**. In: *Longitudinal data analysis.* Chapman and Hall/CRC; 2008: 17-42.
2. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P: **Longitudinal studies**. *Journal of thoracic disease* 2015, **7**(11):E537.
3. Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S: **Analysis of longitudinal data**: Oxford University Press; 2002.
4. Molenberghs G, Kenward M: **Missing data in clinical studies**, vol. 61: John Wiley & Sons; 2007.
5. Hellriegel B: **Immunoepidemiology--bridging the gap between immunology and epidemiology**. *Trends in parasitology* 2001, **17**(2):102-106.
6. Krause PJ, Kavathas PB, Ruddle NH: **Immunoepidemiology**: Springer International Publishing; 2019.
7. Genser B, Cooper PJ, Yazdanbakhsh M, Barreto ML, Rodrigues LC: **A guide to modern statistical analysis of immunological data**. *BMC immunology* 2007, **8**(1):27.
8. Genser B, Fischer JE, Figueiredo CA, Alcântara-Neves N, Barreto ML, Cooper PJ, Amorim LD, Saemann MD, Weichhart T, Rodrigues LC: **Applied immuno-epidemiological research: an approach for integrating existing knowledge into the statistical analysis of multiple immune markers**. *BMC immunology* 2016, **17**(1):11.
9. McGuinness D, Bennett S, Riley E: **Statistical analysis of highly skewed immune response data**. *Journal of immunological methods* 1997, **201**(1):99-114.
10. Bennett S, Riley E: **The statistical analysis of data from immunoepidemiological studies**. *Journal of immunological methods* 1992, **146**(2):229-239.
11. Panda A, Chen S, Shaw AC, Allore HG: **Statistical approaches for analyzing immunologic data of repeated observations: a practical guide**. *Journal of immunological methods* 2013, **398**:19-26.
12. Verbeke G, Fieuws S, Molenberghs G, Davidian M: **The analysis of multivariate longitudinal data: A review**. *Statistical methods in medical research* 2014, **23**(1):42-59.
13. Mawa PA, Nkurununngi G, Egesa M, Webb EL, Smith SG, Kizindo R, Akello M, Lule SA, Muwanga M, Dockrell HM: **The impact of maternal infection with Mycobacterium tuberculosis on the**

infant response to bacille Calmette–Guérin immunization. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2015, **370**(1671):20140137.

14. Verbeke G, Molenberghs G: **Linear mixed models for longitudinal data**: Springer Science & Business Media; 2009.

15. FitzMaurice Garrett M, Laird Nan M, Ware James H: **Applied longitudinal analysis**. In.: John Wiley & Sons; 2004.

16. Weiss RE: **Modeling longitudinal data**: Springer Science & Business Media; 2005.

17. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G: **Handbooks of modern statistical methods: Longitudinal data analysis**. In.: New York: Taylor & Francis Group; 2009.

18. Burton P, Gurrin L, Sly P: **Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling**. *Statistics in medicine* 1998, **17**(11):1261-1291.

19. Long J, Ryoo J: **Using fractional polynomials to model non-linear trends in longitudinal data**. *British Journal of Mathematical and Statistical Psychology* 2010, **63**(1):177-203.

20. Tilling K, Macdonald-Wallis C, Lawlor DA, Hughes RA, Howe LD: **Modelling childhood growth using fractional polynomials and linear splines**. *Annals of Nutrition and Metabolism* 2014, **65**(2-3):129-138.

21. Royston P, Ambler G, Sauerbrei W: **The use of fractional polynomials to model continuous risk variables in epidemiology**. *International journal of epidemiology* 1999, **28**(5):964-974.

22. Benner A: **mfp: Multivariable fractional polynomials**. *R news* 2005, **5**(2):20-23.

23. Fieuws S, Verbeke G: **Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles**. *Biometrics* 2006, **62**(2):424-431.

24. Fieuws S, Verbeke G, Molenberghs G: **Random-effects models for multivariate repeated measures**. *Statistical methods in medical research* 2007, **16**(5):387-397.

25. **A SAS MACRO FOR FITTING A MULTIVARIATE LINEAR MIXED MODEL USING THE PAIRWISE APPROACH** [https://ibiostat.be/online-resources/longitudinal#MLMMpw]

Figure 1

Title: Figure 1. Individual profiles for a sample of 30 infants

Legend:

PPD-specific cytokine responses (on the log scale) for a consecutive sample of 30 infants, for each of the 7 cytokines considered. Each infant is represented by a single line.

Figure 2

Title: Figure 2. Average evolutions of cytokine responses to PPD in infants of mothers with or without LTBI

Legend:

Average PPD-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent 95% confidence intervals for the mean. The solid and dashed lines represent concentrations from children born of LTBI-positive and LTBI-negative mothers respectively.

Figure 3

Title: Figure 3. Estimates for interaction effects of time and LTBI status from univariate and pairwise models

Legend:

Estimates and 95% confidence intervals for the interaction effects of time and mother's LTBI status from univariate linear mixed models (solid circles and associated dashed whiskers) and from the pairwise joint modelling approach (solid squares and whiskers).

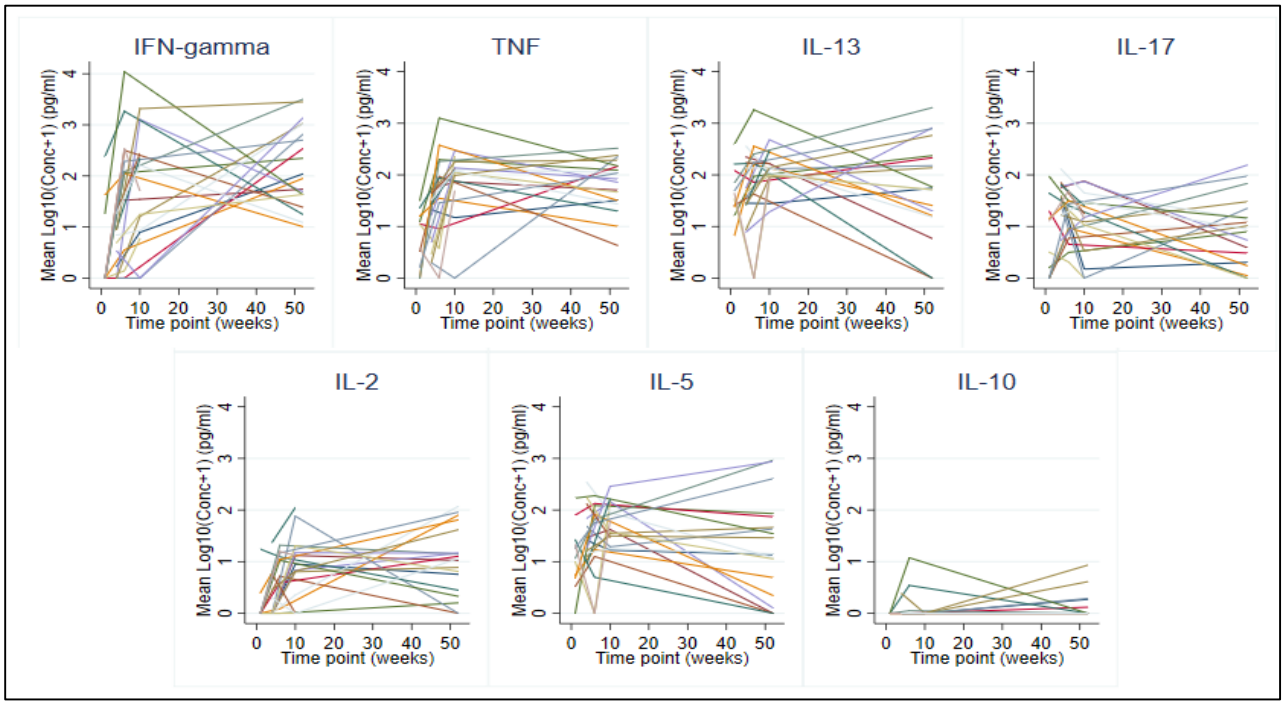
Figure 4

Title: Figure 4. Principal components analysis of the correlation matrix of random intercepts and of cord blood outcomes

Legend:

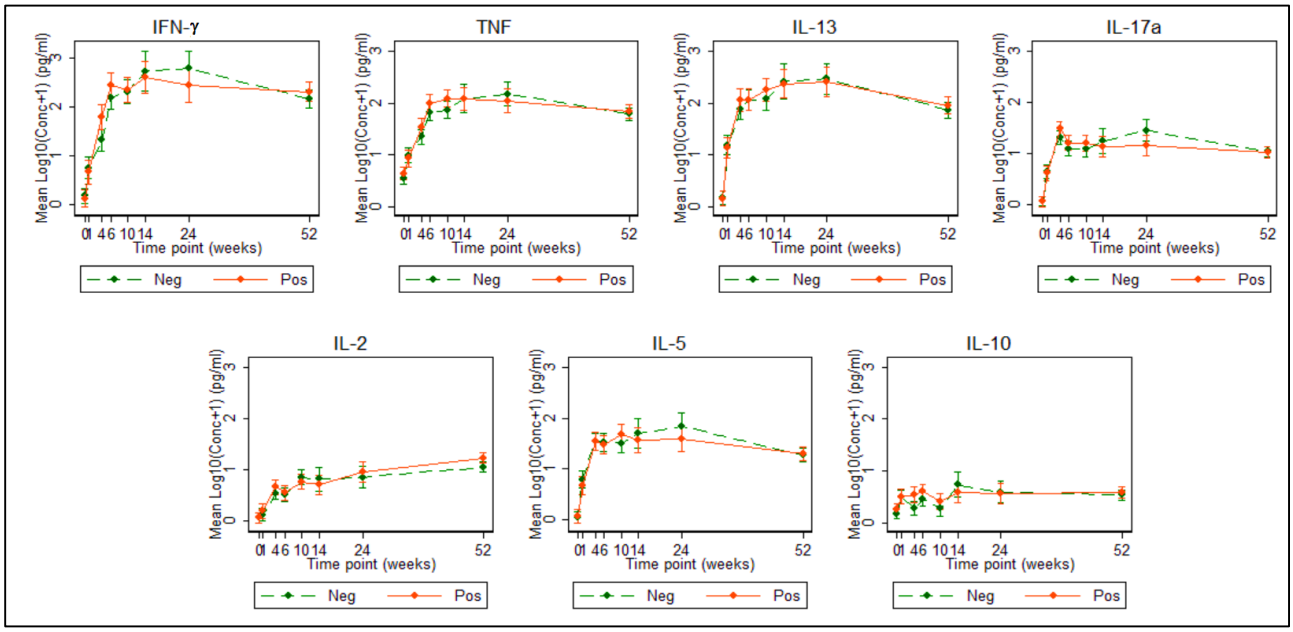
A: Component loadings for the seven outcomes based on the 7 x 7 correlation matrix of random intercepts from the pairwise approach. B: Component loadings for the seven outcomes based on cord blood responses to PPD.

Figure 1. Individual profiles for a sample of 30 infants



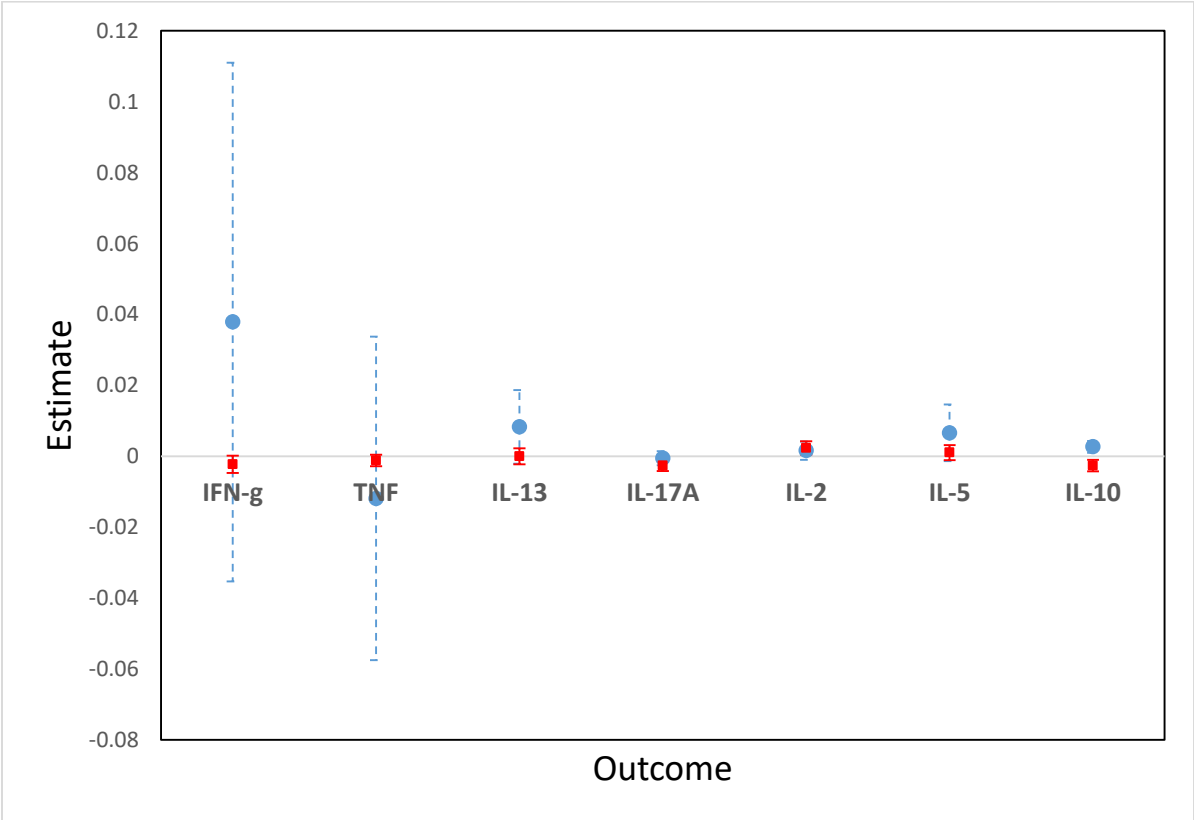
PPD-specific cytokine responses (on the log scale) for a consecutive sample of 30 infants, for each of the 7 cytokines considered. Each infant is represented by a single line.

Figure 2. Average evolutions of cytokine responses to PPD in infants of mothers with or without LTBI



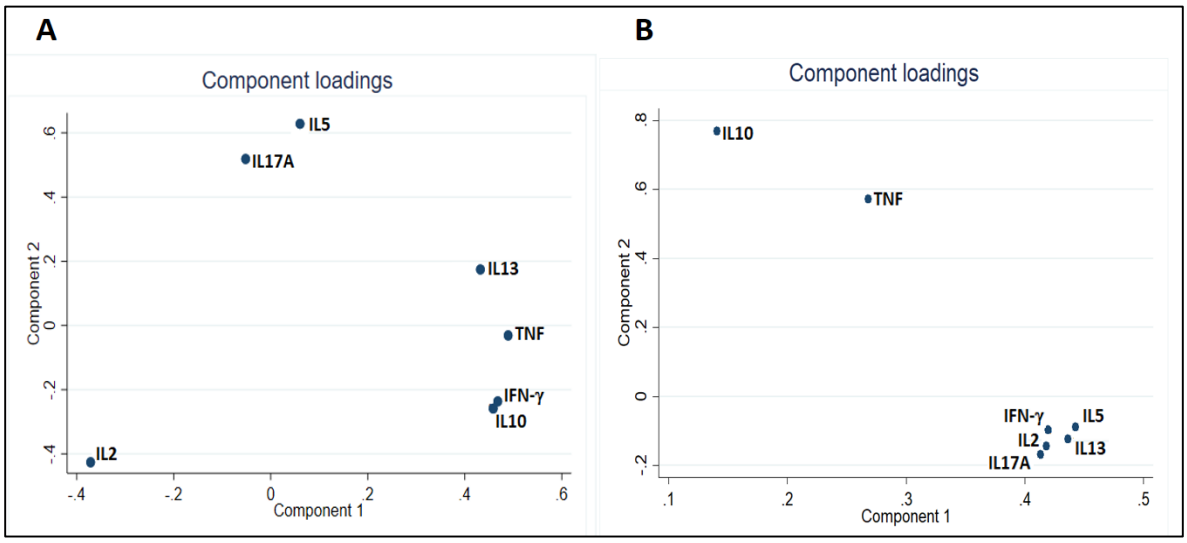
Average PPD-specific cytokine concentrations corrected for background (on the log scale). Points represent mean values and the bars represent 95% confidence intervals around the mean. The solid and dashed lines represent concentrations from children born of LTBI-positive and LTBI-negative mothers respectively.

Figure 3. Estimates for interaction effects of time and LTBI status from univariate and pairwise joint models



Estimates and 95% confidence intervals for the interaction effects of time and mother's LTBI status from univariate linear mixed models (solid circles and associated dashed whiskers) and from the pairwise joint modelling approach (solid squares and whiskers).

Figure 4. Principal components analysis of the correlation matrix of random intercepts and of cord blood outcomes



A: Component loadings for the seven outcomes based on the 7 x 7 correlation matrix of random intercepts from the pairwise approach. B: Component loadings for the seven outcomes based on cord blood responses to PPD.

4.3. Supplementary information for Research Paper 2

4.3.1. Supplementary tables for Research Paper 2

Table S1. Number of available samples at each time point for responses to PPD

Time point (weeks)	LTBI-Negative (n=150)	LTBI-Positive (n=132)	Total (n=282)
0	138	116	254
1	71	58	129
4	66	53	119
6	57	56	113
10	57	51	108
14	21	32	53
24	27	29	56
52	110	95	205

Table shows the number of samples assayed at each time point. Half of the infants were expected at time points other than weeks 0 and 52. Weeks 14 and 24 have lower numbers because they were introduced later.

Table S2. Comparison of Fractional polynomial and conventional higher order polynomials using various fit-statistics

Outcome	Model	AIC	Deviance	R^2	Adjusted R^2
IFN-gamma	Cubic	2309.9	864.83	0.2562	0.2533
	FP (0.5, 0.5)*	2303.5	859.99	0.2604	0.2585
TNF	Cubic	1513.2	316.42	0.247	0.2441
	FP (0.5, 0.5)	1508.4	315.28	0.2498	0.2478
IL-13	Cubic	1976.8	567.37	0.1463	0.143
	FP (-0.5, 3)	1961.3	557.66	0.1609	0.1588
IL-17A	Cubic	1527.7	318.48	0.0544	0.0507
	FP (-2, -2)	1476.7	299.15	0.1118	0.1095
IL-2	Cubic	1465.7	294.89	0.2146	0.2115
	FP (0)	1457.6	293.35	0.2187	0.2177
IL-5	Cubic	1844	477.81	0.1065	0.103
	FP (-1, 3)	1817.6	463.12	0.1339	0.1317
IL-10	Quadratic	1400.6	271.46	0.006027	0.003478
	Linear	1398.8	271.53	0.005768	0.004495

*FP indicates a Fractional Polynomial model with the corresponding powers of time in brackets.

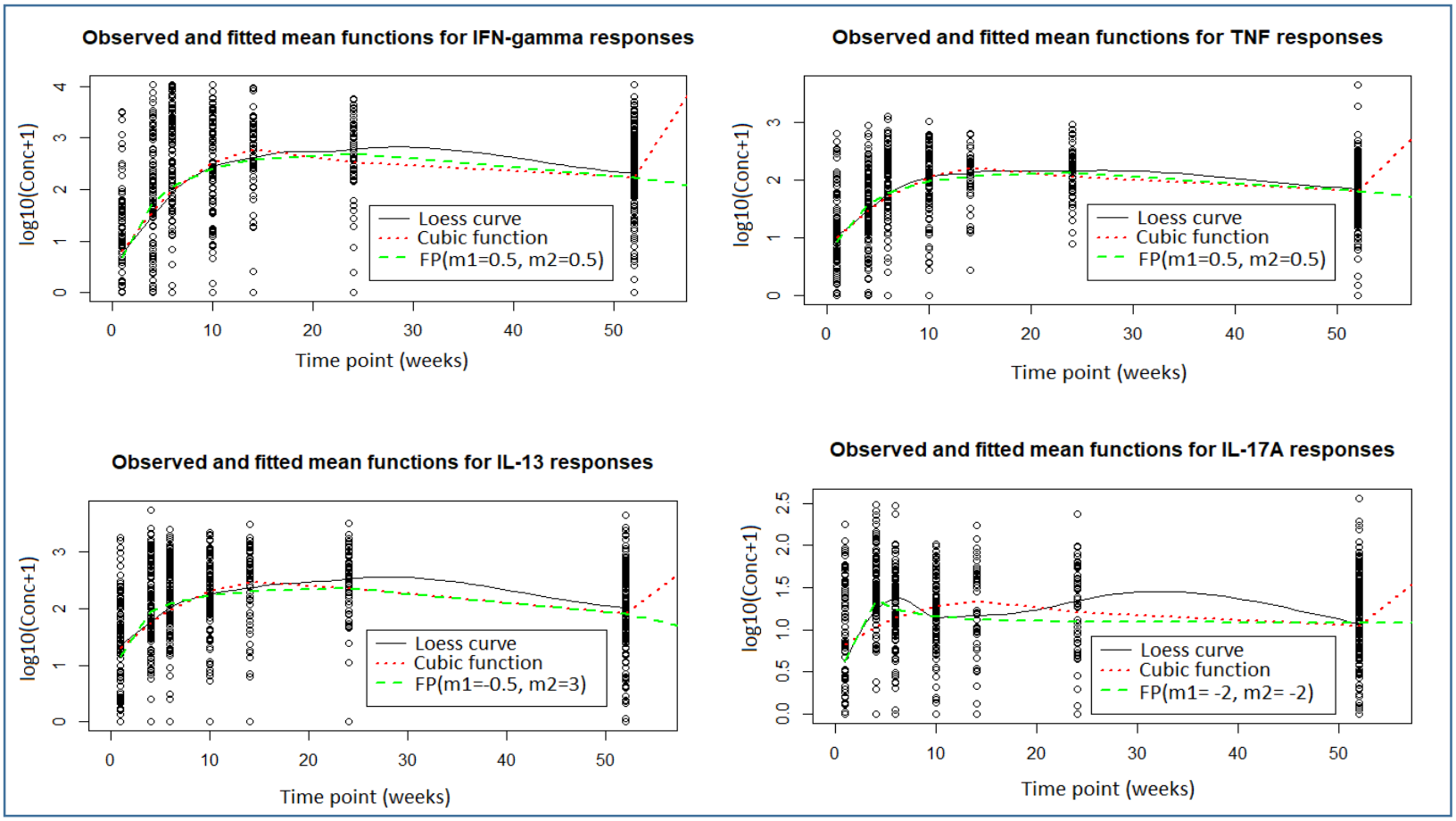
Table S3. Correlation matrix of random effects from the pairwise model

	IFN-γ	IL-10	IL-13	IL-17	IL-2	IL-5	TNF
IFN-γ	1						
IL-10	**0.81**	1					
IL-13	**0.71**	**0.70**	1				
IL-17	0.28	0.21	0.29	1			
IL-2	0.21	-0.01	0.14	0.09	1		
IL-5	0.31	0.01	**0.73**	0.42	-0.04	1	
TNF	**0.89**	**0.77**	**0.75**	0.42	-0.07	0.27	1

Correlation coefficients greater than 0.5 are in bold.

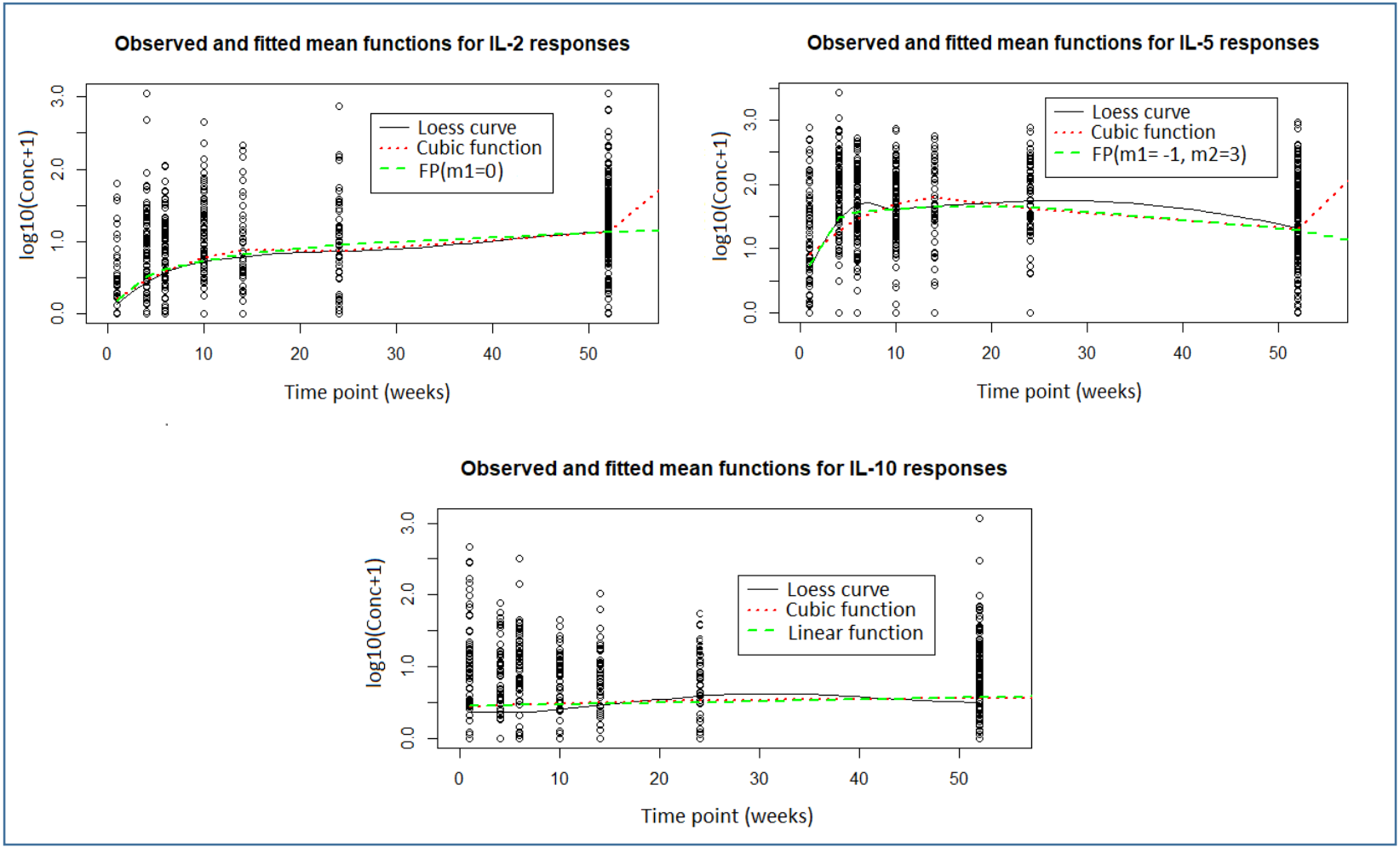
4.3.2. Supplementary figures for Research Paper 2

Figure S1A. Observed and fitted mean functions for the cytokine responses to PPD



Observed mean functions from loess smoothing are represented by black solid lines, red dotted lines represent the best fitting conventional higher order polynomial function while green dashed lines represent the best fitting fractional polynomial (FP) function of time. The figures in brackets for the FP functions represent the corresponding powers of time.

Figure S1B. Observed and fitted mean functions for the cytokine responses to PPD



Observed mean functions from loess smoothing are represented by black solid lines, red dotted lines represent the best fitting conventional higher order polynomial function while green dashed lines represent the best fitting fractional polynomial (FP) function of time. The figures in brackets for the FP functions represent the corresponding powers of time.

Chapter 5. A latent class analysis approach for investigating early childhood infection-exposure and its association with allergy-related diseases in later childhood

5.1. Introduction

This chapter illustrates the use of latent class analysis (LCA) to categorise children into underlying groups based on their early exposure to several infections, and to assess associations between these groups and allergy-related diseases in later childhood (**thesis objective 2**). Application of this statistical analysis approach is illustrated using data from the EMaBS. The LCA method also deals with missing data (**thesis objective 3**) under the MAR assumption via a likelihood based approach. Results are presented in Research paper 3, entitled "Infection-exposure in infancy is associated with reduced allergy-related disease in later childhood in a Ugandan cohort", this manuscript is under review at the eLife Journal.

LCA is used to characterise the infection-exposure of participants in the EMaBS using data on malaria, diarrhoea, upper and lower respiratory tract infections, intestinal helminths, seroprevalence of HSV, CMV and norovirus, collected during each of the first five years of life. The prevalences of ARDs (wheeze, eczema, rhinitis, SPT positivity and asIgE, each considered separately) at nine years are then compared across latent classes, adjusted for covariates. These analyses highlight a statistical analysis method for longitudinal immuno-epidemiological multiple exposure data in line with the aim of this thesis.

5.2. Research Paper 3: Infection-exposure in infancy is associated with reduced allergy-related disease in later childhood in a Ugandan cohort

Early childhood infection-exposure is associated with reduced allergy-related diseases in later childhood in a Ugandan birth cohort

Lawrence Lubyayi[1,2*], Harriet Mpairwe[2,3], Gyaviira Nkurunungi[2,4], Swaib A. Lule[5], Angela Nalwoga[2], Emily L. Webb[6†], Jonathan Levin[1†], Alison M. Elliott[2,7†]

[1]Department of Epidemiology and Biostatistics, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa

[2]Immunomodulation and Vaccines Programme, Medical Research Council/Uganda Virus Research Institute and London School of Hygiene and Tropical Medicine Uganda Research Unit Entebbe, Uganda

[3]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, United Kingdom

[4]Department of Infection Biology, London School of Hygiene and Tropical Medicine, United Kingdom

[5]Institute for Global Health, University College London, London, United Kingdom

[6]MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, United Kingdom

[7]Department of Clinical Research, London School of Hygiene and Tropical Medicine, United Kingdom

***Corresponding author:** Lawrence Lubyayi, MRC/UVRI and LSHTM Uganda Research Unit, Plot 51-59 Nakiwogo Road, PO Box 49 Entebbe Uganda. Email: lawrence.lubyayi@mrcuganda.org; lawrencelby@gmail.com

†Joint senior authorship

Abstract

Lack of early infection-exposure has been associated with increased allergy-related disease (ARD) susceptibility. In tropical Africa, little is known about which infections contribute to development of ARDs, and at which time. We used latent class analysis to characterise the early infection-exposure of participants in a Ugandan birth cohort and assessed ARDs in later childhood.

Of 2345 live-births, 2115 children (90%) had data on infections within the first year of life while 1179 (50%) had outcome data at nine years. We identified two latent classes of children based on first-year infection-exposure; class one (33% membership), characterised by higher probabilities for malaria (78%), diarrhoea (76%) and Lower Respiratory Tract Infections (22%), was associated with lower prevalence of wheeze, eczema, rhinitis and *Dermatophagoides* SPT positivity at nine years.

In conclusion, in this Ugandan birth cohort, early childhood infection-exposure, notably to malaria and diarrhoea, is associated with lower prevalence of ARDs in later childhood.

Introduction

Lack of early infection-exposure has been associated with increased allergy-related disease (ARD) susceptibility later in life, the so-called "hygiene hypothesis"[1-3] or "old friends hypothesis"[4, 5]. Studies, mainly in high-income countries (HICs), have highlighted the inverse relationship between pathogen-exposure and atopy, specifically for Hepatitis A virus[6], gut flora[7], intestinal parasites[8], mycobacteria[9], malaria[10], total burden of infections[11, 12] and farm animal sheds[13].

48 Despite increasing evidence indicating inverse relationships between pathogen-exposure
49 and atopy, other studies suggest no association[14, 15] or even increased risk of atopy following
50 a combination of early infections[16, 17]. Therefore, it remains unclear which infections, and at
51 which times, possibly contribute to reducing the risk of atopy or ARDs.

52 There is a paucity of data from countries in tropical Africa where the infectious diseases
53 burden remains high yet prevalence of ARDs remains low, despite reasonably high
54 prevalence of atopy. Similar to findings from HICs[18], we have shown previously that ARDs
55 (eczema, wheeze and rhinitis) decline with age, despite increasing atopy[19].

56 We therefore investigated the role of common early childhood infections in the
57 development of atopy and ARDs, using data from the Entebbe Mother and Baby Study
58 (EMaBS) birth cohort[20, 21]. We hypothesised that (i) there are latent classes of children with
59 similar early infection-exposure and (ii) specific profiles of early infection-exposure are
60 associated with reduced ARDs in later childhood.

61

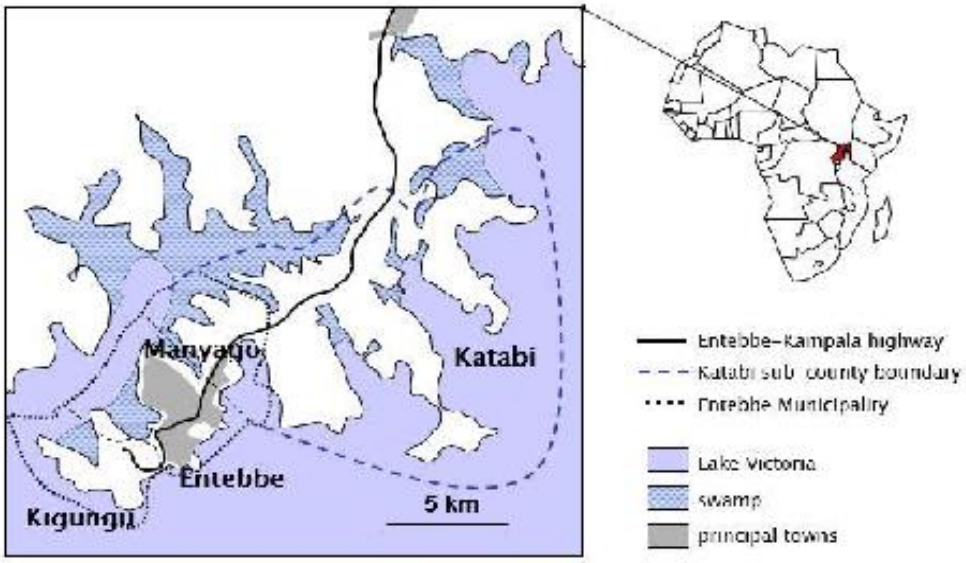
62 **Methods**

63 ***Study design and population***

64 The EMaBS has been described previously[21]. Briefly, the EMaBS, based at Entebbe General
65 Hospital, Uganda, originated as a trial of anthelminthic treatment in pregnancy and early
66 childhood, in a 2x2(x2) factorial design. Between 2003 and 2005, women in their second or
67 third trimester of pregnancy were randomised to single-dose albendazole (400mg) versus
68 placebo and single-dose praziquantel (40mg/kg) versus placebo; their children were
69 randomised to quarterly albendazole versus placebo from age 15 months to 5 years[20, 21].

70 Cohort participants continue under follow-up and had detailed data on ARDs collected at
71 age nine years. The geography of the study area has been described previously[22]; the main
72 zones of maternal residence at enrolment included Kigungu (a relatively isolated fishing
73 village, at the tip of the Entebbe peninsula), Entebbe (a commercial town), Manyago and
74 Kabale (peri-urban) and Katabi (comprised of peri-urban and rural areas) (Figure 1).

75 **Figure 1:** Map of Entebbe and Katabi showing the study area and the three main
76 geographical zones of maternal residence at enrolment.



77

78 The study area was located about 50km southwest of Kampala, Uganda's capital city.

79 ***Ethical approval***

80 The study was approved by ethics committees of the Uganda Virus Research Institute,
81 London School of Hygiene and Tropical Medicine and Uganda National Council for Science
82 and Technology. Written informed consent and assent were obtained.

Exposure variables

The main exposures for this study were children's most common clinical illnesses (malaria, diarrhoea, Upper and Lower Respiratory Tract Infections (URTI and LRTI), intestinal helminths), and seroprevalence of Herpes Simplex Virus (HSV), cytomegalovirus (CMV) and norovirus, in the first five years of life.

Malaria, diarrhoea, URTI and LRTI were assessed and recorded by a doctor when a participant presented to the study clinic with an illness. Malaria was fever (temperature ≥ 37.5°C) with parasitaemia by thick film microscopy; diarrhoea was a child's carer's definition, with stool frequency recorded[20]; URTI was the common cold while LRTI was cough, with difficulty in breathing, and fast breathing (defined by age), with or without abnormal breath sounds[21].

At scheduled annual visits, a stool sample was collected and examined for presence of helminth eggs using the Kato-Katz method[23]. Two slides were examined for each sample, read within 30 minutes for hookworm and the following day for other worms including *Schistosoma mansoni*, *Trichuris trichiura* and *Ascaris lumbricoides*[24].

Norovirus seropositivity was determined using Immunoglobulin (Ig)G responses to human norovirus virus-like particles using enzyme-linked immunosorbent assays (ELISAs), as reported previously[25]. For 779 randomly selected children, plasma samples at one year, were screened; any found negative were further tested at subsequent years until a positive response was detected[25].

HSV and CMV seropositivity were determined using IgG responses, from standard commercially available ELISAs (DiaSorin, Saluggia, Italy). All plasma samples available at two

years and five years were assayed. Children whose samples were positive at two years had their one-year samples tested. Those with negatives samples at two years were further assessed at years three and four until a positive response was detected.

Allergy-related disease outcomes at nine years

ARDs at nine years were the outcomes for this analysis. This is because there was a specific effort to collect data on ARDs, skin prick test (SPT) reactivity and allergen-specific immunoglobulin E (asIgE) during the nine-year annual visit. We used the International Study on Allergy and Asthma in Children (ISAAC) questionnaire[26] to collect data on recent (last 12 months) reported wheeze, eczema (recurrent itchy rash with typical flexural distribution) and rhinitis, classified according to responses from caregivers for their children[27].

Atopy was assessed by SPT and asIgE as previously reported[19, 27]. SPT reactivity to *Dermatophagoides* mix (*D. farinae* and *D. pteronyssinus*), *Blomia tropicalis*, German cockroach, cat, mould, grass pollen, Bermuda grass and peanut (ALK-Abelló, Laboratory Specialities (Pty) Ltd, Randburg, South Africa) was assessed using standard methods[28]. A test was classified as positive if there was a papule of average size >3mm, in the presence of saline [negative] and histamine [positive] controls. Plasma IgE to house dust mite (*Dermatophagoides pteronyssinus*) and German cockroach [*Blatella germanica*]) was measured using an in-house ELISA as previously described[29, 30]. The ELISA lower detection limit was 15.625ng/ml, but since we used 20-fold diluted plasma samples in our assay, the lower detection limit in undiluted plasma was 312.5ng/ml.

Statistical analysis

Analyses aimed to classify participants into latent classes based on infection-exposure during the first five years of life and to study the association between early infection-

exposure and ARDs at nine years. Since infection experience is an unobservable construct which can be inferred from multiple observed infections, we used the Latent Class Analysis (LCA) framework which has been shown to flexibly divide a population into mutually exclusive subgroups[31-33].

Initially we planned to use latent transition analysis (LTA), a longitudinal extension of LCA, which allows for characterisation of memberships as well as predicting transitions among memberships over time[34, 35]. LTA requires longitudinal measurement invariance of item-response probabilities, with latent classes preserving their characteristics over time[35]. Since the test for longitudinal measurement invariance did not hold, we fitted LCA models separately at each time point, as recommended[35], using infection-exposure variables as indicators. Models were fitted, successively increasing the number of classes up to a seven-class solution, separately for each of the first five years, with no a priori restrictions to find consistent classes across the different years. We used the Bayesian Information Criterion (BIC), adjusted BIC (ABIC), consistent Akaike Information Criterion (CAIC) and entropy to guide the choice of optimal number of classes[32, 36].

We conducted multiple-group LCA[32] to establish whether measurement invariance across sex holds. Multiple-group LCA allows class membership and item response probabilities to vary across groups. Formal tests for measurement invariance across sex were conducted for the selected models at each year. We incorporated the covariates maternal area of residence, and maternal and household socio-economic status (SES) at enrolment, to establish whether these impacted the probability of class membership, through multinomial logistic regression[32]. Maternal SES was determined by level of education, personal income, and occupation while household SES was based on building materials and the number of

151 rooms and items owned[37]. SES scores were split into two: low and high. We predicted ARDs
152 at nine years, from latent class membership, using a SAS macro[38] implementing a model-
153 based approach as previously described[39].

154 All models were estimated using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA) via PROC
155 LCA which handles missing data on latent class indicators under the missing at random
156 assumption[32].

157 **Results**

158 Of 2345 live-births in the EMaBS cohort, 2115 had data on illnesses during their first year of
159 life, 1945 in the second, 1884 in the third, 1856 in the fourth and 1838 in their fifth year
160 (Figure 2). Diarrhoea and LRTI were common in the first year of life but declined by the fifth
161 year, malaria increased slightly in the second year and then declined, URTI remained very
162 common throughout the first five years, helminth infections increased slightly, HSV
163 increased steadily over the five years while CMV and norovirus seropositivity increased to
164 over 85% by the second year (Figure 3).

165

166

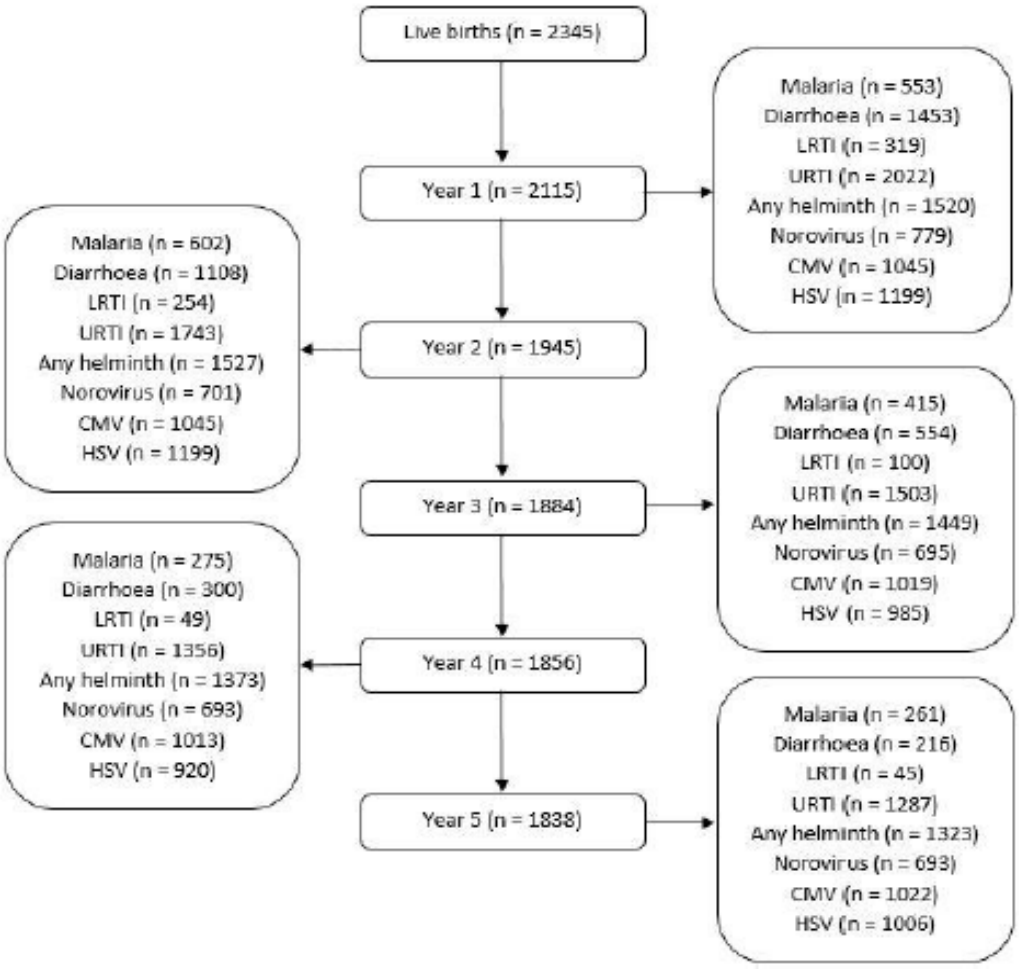
167

168

169

170

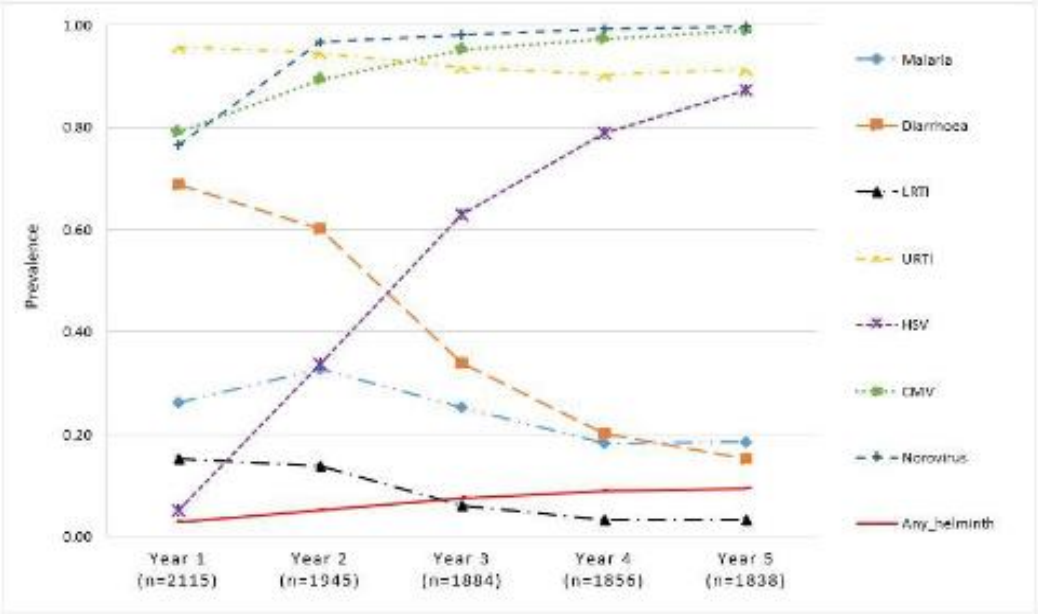
Figure 2: Flow chart showing number of the Entebbe Mother and Baby Study participants with data on illnesses/infections during each year



Numbers for each year (centre column of the flowchart) represent the children who presented at the study clinic at any point during that year. Numbers for malaria, diarrhoea, LRTI and URTI represent the children who were positively diagnosed (at least once during a given year) for each of those illnesses, on presentation at the study clinic. Numbers for any helminth, norovirus, CMV and HSV represent children samples assessed for any of those infections at the respective annual visit.

185 **Figure 3:** Proportion of children with illnesses/infections, over time, in the Entebbe Mother

186 and Baby Study cohort



187

188

189 ***Latent Class Analysis at each time point***

190 Table 1 summarises results for the LCAs at each time point. Based on BIC, CAIC and ABIC,

191 two-class solution models were favoured at all time periods. Although the entropy, which

192 measures certainty in classifying latent statuses, favoured seven-class and six-class solutions

193 for the first and second years respectively, we chose the two-class solutions for easy

194 interpretation in addition to being favoured by other statistics.

195 Multiple group LCA showed that measurement invariance by sex holds, based on the

196 likelihood ratio difference test (*P*-value = 0.379; *P*-value = 0.245; *P*-value = 0.193; *P*-value =

197 0.403 at years one, two, three and five respectively), except at year four (*P*-value = 0.002)

198 (Supplementary Table S1).

Table 1: Model fit statistics for Latent Class Analysis at each time point

	G^{2a}	AIC[b]	BIC[c]	CAIC[d]	ABIC[e]	Entropy	DF[f]
Year 1							
2-class solution	126.22	160.22	**256.88**	**273.88**	**202.87**	0.39	238
3-class solution	95.27	**147.27**	295.11	321.11	212.51	0.56	229
4-class solution	80.24	150.24	349.26	384.26	238.06	0.38	220
5-class solution	70.89	158.89	409.09	453.09	269.29	0.35	211
6-class solution	61.69	167.69	469.08	522.08	300.69	0.46	202
7-class solution	53.52	177.52	530.09	592.09	333.11	**0.57**	193
Year 2							
2-class solution	130.58	164.58	**259.32**	**276.32**	**205.31**	0.27	238
3-class solution	110.18	**162.18**	307.09	333.09	224.49	0.49	229
4-class solution	97.96	167.96	363.01	398.01	251.82	0.53	220
5-class solution	90.00	178.00	423.22	467.22	283.43	0.48	211
6-class solution	81.98	187.98	483.38	536.38	315.00	**0.66**	202
7-class solution	74.08	198.08	543.63	605.63	346.66	0.51	193
Year 3							
2-class solution	103.26	137.26	**231.46**	**248.46**	**177.45**	**0.76**	238
3-class solution	79.01	**131.01**	275.09	301.09	192.49	0.53	229
4-class solution	70.53	140.53	334.49	369.49	223.29	0.55	220
5-class solution	59.96	147.96	391.79	435.79	252.00	0.48	211
6-class solution	52.74	158.74	452.45	505.45	284.07	0.52	202
7-class solution	47.57	171.57	515.15	577.15	318.18	0.56	193
Year 4							
2-class solution	56.34	90.34	**184.28**	**201.28**	**130.27**	**0.45**	238
3-class solution	36.65	**88.65**	232.34	258.34	149.73	0.38	229
4-class solution	27.06	97.06	290.49	325.49	179.30	0.33	220
5-class solution	19.10	107.10	350.28	394.28	210.49	0.41	211
6-class solution	14.36	120.36	413.28	466.28	244.90	0.44	202
7-class solution	12.24	136.24	478.90	540.90	281.93	0.37	193
Year 5							
2-class solution	70.53	**104.53**	**198.31**	**215.31**	**144.30**	**0.72**	238
3-class solution	55.98	107.98	251.42	277.42	168.82	0.34	229
4-class solution	46.39	116.39	309.49	344.49	198.29	0.47	220
5-class solution	35.82	123.82	366.56	410.56	226.78	0.55	211
6-class solution	27.72	133.72	426.12	479.12	257.74	0.52	202
7-class solution	22.35	146.35	488.40	550.40	291.43	0.46	193

Bold indicates the best solution model for the corresponding fit index.

[a] G^2 is the likelihood ratio statistic.

[b] AIC is Akaike Information Criterion.

[c] BIC is Bayesian Information Criterion.

[d] CAIC is consistent AIC.

[e] ABIC is adjusted BIC.

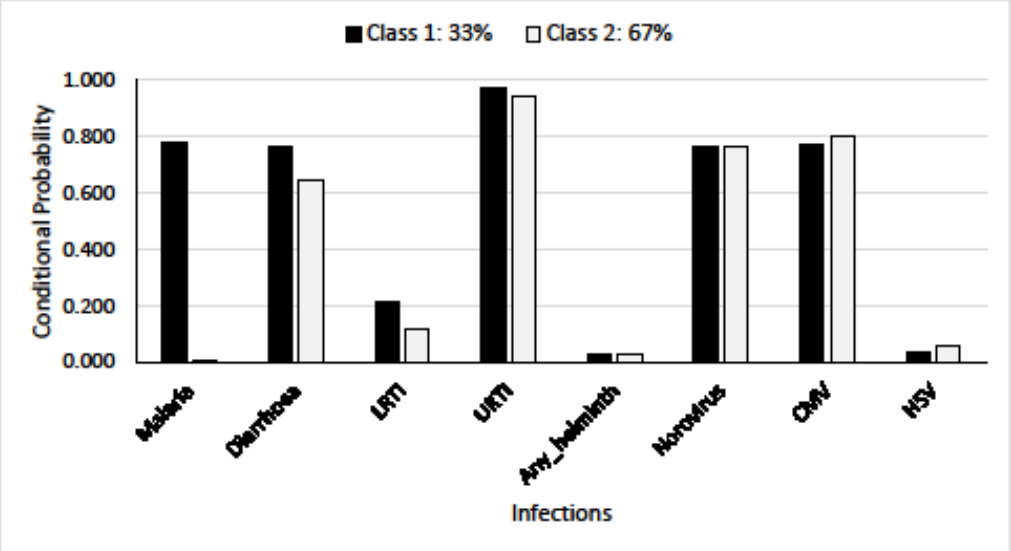
[f] DF is degree of freedom.

11

Definition of latent classes

Figure 4 shows the probability of having any of the infections, conditional on membership in either of the latent classes, during the first year of life. The probabilities of membership in latent class (LC) 1 and LC 2 are estimated as 33% and 67% respectively. LC 1 was characterised by higher probabilities for malaria, diarrhoea and LRTI (78% vs 0.5%, 76% vs 65% and 22% vs 12% respectively) compared to LC 2. Conditional probabilities for other infections were similar between the LCs.

Figure 4: Probability of having illnesses/infections conditional on latent class membership during the first year



Supplementary Figures S1, S2, S3 and S4 show class membership and conditional probabilities for infections at years two, three, four and five respectively. At all time periods, we identified two LCs one of which was characterised by higher probabilities for malaria and any of diarrhoea, LRTI, URTI or helminths. At year five, helminths were noticeably higher in

LC 1, while diarrhoea, LRTI and URTI had slightly higher probabilities in LC 2 (Supplementary Figure S4). We labelled the class with higher probabilities for various infections as LC 1 and the other as LC 2. Membership of LC 1 was 33% in year one, 41% in year two, 92% in year three, 21% and 26% for males and females (respectively) in year four and 17% in year five (Table 2).

Table 2: Number (and proportion) of children in each latent class at each time period

	LC[a] 1 Number (%)	LC 2 Number (%)
Year 1 (n = 2038)	673 (33)	1365 (67)
Year 2 (n = 1818)	745 (41)	1073 (59)
Year 3 (n = 1884)	1733 (92)	151 (8)
Year 4 Male (n = 888)	186 (21)	702 (79)
Year 4 Female (n = 844)	219 (26)	625 (74)
Year 5 (n = 1717)	292 (17)	1425 (83)

[a]LC is latent class. LC 1 represents the class with higher probabilities for various infections as compared to LC 2, however the profile of infections changes at each time period.

Effects of covariates on latent class membership

At year one, compared to children born in Entebbe town, those born in Kigungu fishing village were less likely to be in LC 1 [Odds Ratio (OR) = 0.35 (95% confidence interval (CI): 0.20, 0.59)], while those born in Katabi (peri-urban and rural areas) were more likely to be in LC 1 [OR = 2.42 (1.74, 3.35)] (Table 3).

Table 3: Odds ratio estimates for effects of covariates on membership in latent class 1 at year one

Covariate	Odds Ratio	(95% CI[a])
Maternal socio-economic status		
Higher vs lower	0.82	(0.65, 1.04)
Household's socio-economic status		
Higher vs lower	0.84	(0.66, 1.06)
Area of residence at birth		
Kigungu vs Entebbe	**0.35**	**(0.20, 0.59)**
Manyago vs Entebbe	1.17	(0.89, 1.54)
Katabi vs Entebbe	**2.42**	**(1.74, 3.35)**

[a]CI is confidence interval

242 At year two, children with higher household SES [OR = 0.61 (0.41, 0.90)], and those born in
243 Kigungu [OR = 0.24 (0.09, 0.61)] were less likely to be in LC 1 (Supplementary Table S2). At
244 year three, it was not possible to establish the effects of covariates on latent class
245 membership because most of the respondents (92%) were classified into LC 1. At year four,
246 males born in households with higher SES were less likely to be in LC 1 [OR= 0.45 (0.21,
247 0.98)] while those born in Kigungu were more likely to be in LC 1 [OR= 15.51 (5.30, 45.41)];
248 females with higher maternal SES were less likely to be in LC 1 [OR= 0.47 (0.30, 0.73)] while
249 those born in Katabi were more likely to be in LC 1 [OR= 2.90 (1.67, 5.05)] (Supplementary
250 Table S3). At year five, children with higher maternal SES were less likely to be in LC 1 [OR=
251 0.24 (0.12, 0.54)] while those born in Kigungu and Katabi were more likely to be in LC 1 [OR=
252 11.97 (4.20, 34.14) for Kigungu and OR= 3.3 (1.22, 8.95) for Katabi] (Supplementary Table
253 S4).

254 ***Prevalence of atopy and ARD outcomes at nine years***

255 Prevalence of atopy and ARDs at nine years has been reported elsewhere[19, 27]. Briefly, out of
256 1179 children for whom data on ARDs was collected at nine years, prevalence of reported
257 recent wheeze was 3.8%, eczema 4.9%, allergic rhinitis 4.6%, SPT positivity for at least one
258 allergen 25%, for *Dermatophagoides* 18%, for *Blomia* 15% and for German cockroach 12%[19].
259 Prevalence of detectable IgE to house dust mite was 29.4%, to cockroach 32.3% or to any of
260 the two 44.1%[19, 27].

261 ***Associations between latent class membership and ARD outcomes at nine years***

262 Based on covariate adjusted LCA of data collected during the first year of life, LC 1
263 membership was associated with lower probabilities of wheeze, eczema, rhinitis and SPT
264 positivity for *Dermatophagoides*; there was some suggestion of difference between the two

LCs for SPT positivity overall and for *Blomia* (with SPT positivity less common for LC 1 children) but for cockroach and the IgE outcomes there was no suggestion of a difference (Table 4).

Table 4: Probability of allergy-related disease (ARD) outcomes at 9 years by latent class membership at year one

			95% CI[a]		
ARD outcomes at 9 years	**Latent class**	**Estimate**	**Lower CI**	**Upper CI**	***P*-value**
Wheeze	LC[b] 1	0.02	0.007	0.043	**0.033**
	LC 2	0.05	0.037	0.072	
Eczema	LC 1	0.03	0.013	0.056	**0.042**
	LC 2	0.06	0.048	0.086	
Rhinitis	LC 1	0.02	0.007	0.044	**0.017**
	LC 2	0.06	0.044	0.081	
SPT-any[c]	LC 1	0.22	0.176	0.277	0.169
	LC 2	0.27	0.237	0.308	
SPT-dermatophagoides	LC 1	0.14	0.102	0.186	**0.039**
	LC 2	0.20	0.172	0.235	
SPT-cockroach	LC 1	0.11	0.078	0.150	0.796
	LC 2	0.11	0.091	0.141	
SPT-Blomia	LC 1	0.12	0.087	0.167	0.115
	LC 2	0.17	0.139	0.198	
IgE-dermatophagoides	LC 1	0.27	0.214	0.323	0.292
	LC 2	0.30	0.268	0.343	
IgE-cockroach	LC 1	0.31	0.253	0.366	0.570
	LC 2	0.33	0.291	0.367	
IgE-any[d]	LC 1	0.42	0.360	0.481	0.476
	LC 2	0.45	0.408	0.488	

[a]CI is confidence interval
[b]LC is latent class. LC 1 represents the class with higher probabilities for various infections as compared to LC 2, however the profile of infections changes at each time period.
[c]SPT-any is skin prick test reactivity to any of *Dermatophagoides* mix, *Blomia tropicalis*, German cockroach, cat, mould, grass pollen, Bermuda grass or peanut.
[d]IgE-any is allergen-specific plasma IgE (asIgE) to any of house dust mite (HDM, *Dermatophagoides pteronyssinus*) or German cockroach.
Estimates are adjusted for maternal area of residence, and maternal and household socio-economic status at enrolment.

LC 1 membership at year two was associated with a lower probability of detectable IgE to any of *Dermatophagoides* or cockroach (Supplementary Table S5). LC membership at year three was not associated with ARDs (Supplementary Table S6). LC 1 membership at year four, for females, was associated with a lower probability of SPT positivity for *B. tropicalis*

(Supplementary Table S7). At year five, LC 1 membership was associated with higher probabilities for detectable IgE to *Dermatophagoides*, cockroach and any of the two (Supplementary Table S8).

Discussion

In this study, using latent class analysis, we identified two distinguishable groups of children based on their early infection-exposure. In the first year of life, LC 1 was characterised by higher probabilities of malaria, diarrhoea and LRTI, and membership of this class was associated with lower proportions of wheeze, eczema, rhinitis and SPT positivity for *Dermatophagoides* at nine years. There was an inverse association with IgE to any of *Dermatophagoides* or cockroach at year two, generally no associations for years three and four and positive associations with IgE at year five.

A major strength of this study lies in the use of LCA, a novel person-centred analysis approach, to characterise infection-exposure - an unobservable construct inferred from multiple observed infections in a given time period. At all time periods, class membership was characterised by differences in the burden of malaria and any of diarrhoea, LRTI, URTI or helminths. Initially we aimed to use LTA, an extension of LCA, in order to characterise memberships and transitions among memberships longitudinally[34, 35], however, the requirement for longitudinal measurement invariance was not met most probably because of the substantial change in infection-exposure over the first five years of life. Consequently, conducting LCA at each time period demonstrated the strong association between infectious illnesses in infancy and ARDs in later childhood. This suggests that infancy is a critical period during which the risk of ARDs in later childhood is established.

306 The finding that LC 1 membership was associated with lower proportions of wheeze,
307 eczema, rhinitis and SPT positivity for *Dermatophagoides* is consistent with observations
308 from elsewhere[10] and with our earlier results which showed that children with more
309 frequent clinical malaria infections in their first five years of life had a lower incidence of
310 eczema in the same period[24], and supports the hygiene hypothesis that infections in early
311 life are associated with a lower risk of ARDs in later childhood[6-13]. However, results were less
312 consistent for infections in years two to five, which may be explained by the declining
313 prevalences of common infections. We cannot exclude the possibility that the results at one
314 year could be chance findings due to the multiple statistical tests conducted, although
315 patterns of association were similar for many of the ARDs.

316 We also describe the life-course of common infections in the first five years of life, in a
317 tropical African birth cohort. We show that diarrhoea, malaria and LRTI were more common
318 in infancy but gradually declined by the fifth year, URTI remained very common while
319 helminth infections increased gradually throughout the first five years, HSV increased
320 steadily over time while CMV and norovirus infections were common throughout the five
321 years. This result is especially important when studying which infections, occurring at which
322 time, contribute to the development of atopy and ARDs. We show that in as much as two
323 LCs are favoured at each time period, both class sizes and conditional probabilities for the
324 infections vary over time. This is not surprising given that the prevalences of individual
325 infections change over time.

326 Compared to children born in Entebbe town, those born in Kigungu fishing village were less
327 likely to be in LC 1 in year one, while those born in Katabi (peri-urban and rural areas) were

328 more likely to be in LC 1. These findings accord with our earlier study which showed
329 comparable geographical variations in the prevalence of malaria in these areas[22].

330 Another intriguing result was that at five years (when prevalence of helminths was highest)
331 LC 1 was associated with a higher probability of helminth infections and membership in this
332 class was associated with higher levels of aslgE responses to crude allergen extracts of dust
333 mite and cockroach, but was not associated with clinical ARDs. These findings accord with
334 our earlier study which showed positive associations between helminth infections and aslgE
335 responses but no associations with clinical ARDs[40]. It is important to note that the
336 immunoassays used in this study measured IgE sensitisation to crude allergen extracts which
337 may contain cross-reactive components conserved in several antigens; this cross-reactivity
338 may explain the higher prevalence of aslgE responses than SPT reactivity.

339 A limitation of our study could be in the potential influence of missing data. The LCA models
340 employ a maximum-likelihood routine which operates under a missing at random
341 assumption. This assumption is inherently untestable, however, we consider it reasonable
342 since the amount of missing data for each indicator variable, during each time period, was
343 relatively low.

344 **Conclusion**

345 In conclusion, we show that in this tropical African birth cohort, children can be classified
346 into two latent classes based on their early childhood infection experience and that
347 exposure to common infections, especially during the first year of life, was associated with
348 lower proportions of ARDs in later childhood.

349

Funding

This work was supported by several funding sources. The EMaBS was supported by the Wellcome Trust, UK, senior fellowships for A.M.E. [grant numbers 064693, 079110, 95778] with additional support from the UK Medical Research Council. L.L. is supported by a PhD fellowship through the DELTAS Africa Initiative SSACAB (grant number 107754). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (grant number 107754/Z/15/Z) and the UK Government. The views expressed in this study are those of the authors and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust or the UK government. E.L.W. received funding from MRC Grant Reference MR/K012126/1; this grant and the MRC/UVRI and LSHTM Uganda Research Unit are jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. S.A.L. was supported by the PANDORA-ID-NET Consortium (EDCTP Reg/Grant RIA2016E-1609) funded by the EDCTP2 programme under the Horizon 2020, the European Union's Framework Programme for Research and Innovation. H.M. was supported by the Wellcome's Institutional Strategic Support Fund (grant number 204928/Z/16/Z).

Acknowledgements

We thank the study participants of the EMaBS, the staff of the Immunomodulation and Vaccines (I-Vac) Programme at the MRC/UVRI and LSHTM Uganda Research Unit, the

midwives of the Entebbe General Hospital Maternity Department and the community field workers in Entebbe municipality and Katabi sub county.

Competing interests

The authors declare that no competing interests exist.

References

1. Strachan DP. Hay fever, hygiene, and household size. *BMJ: British Medical Journal* 1989; **299**: 1259.
2. Bloomfield S, Stanwell-Smith R, Crevel R, Pickup J. Too clean, or not too clean: the hygiene hypothesis and home hygiene. *Clinical & Experimental Allergy* 2006; **36**: 402-25.
3. Schaub B, Lauener R, von Mutius E. The many faces of the hygiene hypothesis. *Journal of allergy and clinical immunology* 2006; **117**: 969-77.
4. Rook G. 99th Dahlem conference on infection, inflammation and chronic inflammatory disorders: Darwinian medicine and the 'hygiene'or 'old friends' hypothesis. *Clinical & Experimental Immunology* 2010; **160**: 70-9.
5. Rook GA. Hygiene hypothesis and autoimmune diseases. *Clinical reviews in allergy & immunology* 2012; **42**: 5-15.
6. Strachan DP. Family size, infection and atopy: the first decade of the'hygiene hypothesis'. *Thorax* 2000; **55**: S2.
7. Bjorksten B, Naaber P, Sepp E, Mikelsaar M. The intestinal microflora in allergic Estonian and Swedish 2-year-old children. *Clinical and experimental allergy* 1999; **29**: 342-6.
8. Yazdanbakhsh M, Matricardi PM. Parasites and the hygiene hypothesis. *Clinical reviews in allergy & immunology* 2004; **26**: 15-23.
9. Shirakawa T, Enomoto T, Shimazu S-i, Hopkin JM. The inverse association between tuberculin responses and atopic disorder. *Science* 1997; **275**: 77-9.
10. Lell B, Borrmann S, Yazdanbakhsh M, Kremsner PG. Atopy and malaria. *Wiener klinische Wochenschrift* 2001; **113**: 927-9.
11. Martinez FD, Holt PG. Role of microbial burden in aetiology of allergy and asthma. *The Lancet* 1999; **354**: SII12-SII5.
12. Illi S, von Mutius E, Lau S, et al. Early childhood infectious diseases and the development of asthma up to school age: a birth cohort study. *Bmj* 2001; **322**: 390-5.
13. Loss GJ, Depner M, Hose AJ, et al. The early development of wheeze. Environmental determinants and genetic susceptibility at 17q21. *American journal of respiratory and critical care medicine* 2016; **193**: 889-97.
14. Jarvis D, Luczynska C, Chinn S, Burney P. The association of hepatitis A and Helicobacter pylori with sensitization to common allergens, asthma and hay fever in a population of young British adults. *Allergy* 2004; **59**: 1063-7.

15. Cooper PJ, Chico ME, Bland M, Griffin GE, Nutman TB. Allergic symptoms, atopy, and geohelminth infections in a rural area of Ecuador. *American journal of respiratory and critical care medicine* 2003; **168**: 313-7.
16. Seaton A, Devereux G. Diet, infection and wheezy illness: lessons from adults. *Pediatric allergy and immunology* 2000; **11**: 37-40.
17. Bager P, Westergaard T, Rostgaard K, Hjalgrim H, Melbye M. Age at childhood infections and risk of atopy. *Thorax* 2002; **57**: 379-82.
18. Bach J-F. The effect of infections on susceptibility to autoimmune and allergic diseases. *New England journal of medicine* 2002; **347**: 911-20.
19. Lule SA, Mpairwe H, Nampijja M, et al. Life-course of atopy and allergy-related disease events in tropical sub-Saharan Africa: a birth cohort study. *Pediatric Allergy and Immunology* 2017; **28**: 377-83.
20. Webb EL, Mawa PA, Ndibazza J, et al. Effect of single-dose anthelmintic treatment during pregnancy on an infant's response to immunisation and on susceptibility to infectious diseases in infancy: a randomised, double-blind, placebo-controlled trial. *Lancet* 2011; **377**: 52-62.
21. Elliott AM, Kizza M, Quigley MA, et al. The impact of helminths on the response to immunization and on the incidence of infection and disease in childhood in Uganda: design of a randomized, double-blind, placebo-controlled, factorial trial of deworming interventions delivered in pregnancy and early childhood [ISRCTN32849447]. *Clinical trials* 2007; **4**: 42-57.
22. Hillier SD, Booth M, Muhangi L, et al. Plasmodium falciparum and helminth coinfection in a semiurban population of pregnant women in Uganda. *The Journal of infectious diseases* 2008; **198**: 920-7.
23. Katz N, Chaves A, Pellegrino J. A simple device for quantitative stool thick-smear technique in schistosomiasis mansoni. *Rev Inst Med Trop Sao Paulo* 1972; **14**: 397-400.
24. Mpairwe H, Ndibazza J, Webb EL, et al. Maternal hookworm modifies risk factors for childhood eczema: results from a birth cohort in Uganda. *Pediatric Allergy and Immunology* 2014; **25**: 481-8.
25. Thorne L, Nalwoga A, Mentzer AJ, et al. The first norovirus longitudinal seroepidemiological study from Sub-Saharan Africa reveals high seroprevalence of diverse genotypes associated With host susceptibility factors. *The Journal of infectious diseases* 2018; **218**: 716-25.
26. Asher M, Keil U, Anderson H, et al. International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *European respiratory journal* 1995; **8**: 483-91.
27. Namara B, Nash S, Lule SA, et al. Effects of treating helminths during pregnancy and early childhood on risk of allergy-related outcomes: Follow-up of a randomized controlled trial. *Pediatric Allergy and Immunology* 2017; **28**: 784-92.
28. Heinzerling L, Mari A, Bergmann K-C, et al. The skin prick test–European standards. *Clinical and translational allergy* 2013; **3**: 3.
29. Mpairwe H, Webb EL, Muhangi L, et al. Anthelminthic treatment during pregnancy is associated with increased risk of infantile eczema: randomised-controlled trial results. *Pediatric Allergy and Immunology* 2011; **22**: 305-12.
30. Nkurunungi G, Kabagenyi J, Nampijja M, et al. Schistosoma mansoni-specific immune responses and allergy in Uganda. *Parasite immunology* 2018; **40**: e12506.
31. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**: 215-31.

456 32. Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS procedure for latent
457 class analysis. *Structural equation modeling: a multidisciplinary journal* 2007; **14**: 671-94.
458 33. Nylund-Gibson K, Choi AY. Ten frequently asked questions about latent class
459 analysis. *Translational Issues in Psychological Science* 2018; **4**: 440.
460 34. Velicer WF, Martin RA, Collins LM. Latent transition analysis for longitudinal data.
461 *Addiction* 1996; **91**: 197-210.
462 35. Ryoo JH, Wang C, Swearer SM, Hull M, Shi D. Longitudinal model building using
463 latent transition analysis: an example using school bullying data. *Frontiers in psychology*
464 2018; **9**: 675.
465 36. Swanson SA, Lindenberg K, Bauer S, Crosby RD. A Monte Carlo investigation of
466 factors influencing latent class analysis: An application to eating disorder research.
467 *International Journal of Eating Disorders* 2012; **45**: 677-84.
468 37. Muhangi L, Woodburn P, Omara M, et al. Associations between mild-to-moderate
469 anaemia in pregnancy and helminth, malaria and HIV infection in Entebbe, Uganda. *Trans R*
470 *Soc Trop Med Hyg* 2007; **101**: 899-907.
471 38. Dziak JJ, Bray BC, Wagner AT. LCA_Distal_BCH SAS Macro Users' Guide (Version 1.1).
472 *The Methodology Center, University Park, PA, Penn State* 2017.
473 39. Lanza ST, Tan X, Bray BC. Latent class analysis with distal outcomes: A flexible model-
474 based approach. *Structural equation modeling: a multidisciplinary journal* 2013; **20**: 1-26.
475 40. Nkurunungi G, Lubyayi L, Versteeg SA, et al. Do helminth infections underpin urban-
476 rural differences in risk factors for allergy-related outcomes? *Clinical & Experimental Allergy*
477 2019; **49**: 663-76.
478

5.3. Supplementary information for Research Paper 3

5.3.1. Supplementary tables for Research Paper 3

Supplementary Table S1: Results for measurement invariance by sex, at each time point

	MI[a]	G^{2b}	DF[c]	Diff.G^{2d}	Diff.DF[e]	*P*-value
Year 1	Yes	220.34	493	17.11	16	0.379
	No	203.23	477			
Year 2	Yes	211.09	493	19.48	16	0.245
	No	191.61	477			
Year 3	Yes	161.49	493	20.64	16	0.193
	No	140.85	477			
Year 4	Yes	126.64	493	37.75	16	0.002
	No	88.89	477			
Year 5	Yes	112.56	493	16.73	16	0.403
	No	95.83	477			

[a]MI is Measurement invariance
[b]G^2 is the likelihood ratio statistic.
[c]DF is degrees of freedom.
[d]Diff.G^2 is the difference of likelihood ratio statistics.
[e]Diff.DF is the difference of degrees of freedom.

Supplementary Table S2: Odds ratio estimates for effects of covariates on membership in latent class 1 at year two

Covariate	Odds Ratio	(95% CI[a])
Maternal socio-economic status		
Higher vs lower	1.29	(0.88, 1.90)
Household's socio-economic status		
Higher vs lower	**0.61**	**(0.41, 0.90)**
Area of residence at birth		
Kigungu vs Entebbe	**0.24**	**(0.09, 0.61)**
Manyago vs Entebbe	0.99	(0.64, 1.55)
Katabi vs Entebbe	0.64	(0.38, 1.10)

[a]CI is confidence interval

Supplementary Table S3: Odds ratio estimates for effects of covariates on membership in latent class 1 at year four

	Male		Female	
Covariate	Odds Ratio	(95% CI[a])	Odds Ratio	(95% CI)
Maternal socio-economic status				
Higher vs lower	1.03	(0.48, 2.21)	**0.47**	**(0.30, 0.73)**
Household's socio-economic status				
Higher vs lower	**0.45**	**(0.21, 0.98)**	0.80	(0.53, 1.23)
Area of residence at birth				
Kigungu vs Entebbe	**15.51**	**(5.30, 45.41)**	0.50	(0.22, 1.17)

Manyago vs Entebbe	0.72	(0.25, 2.07)	0.97	(0.58, 1.64)
Katabi vs Entebbe	2.44	(0.83, 7.17)	**2.90**	**(1.67, 5.05)**

[a]CI is confidence interval

Supplementary Table S4: Odds ratio estimates for effects of covariates on membership in latent class 1 at year five

Covariate	Odds Ratio	(95% CI[a])
Maternal socio-economic status		
Higher vs lower	**0.24**	**(0.12, 0.54)**
Household's socio-economic status		
Higher vs lower	0.64	(0.31, 1.31)
Area of residence at birth		
Kigungu vs Entebbe	**11.97**	**(4.20, 34.14)**
Manyago vs Entebbe	0.77	(0.28, 2.10)
Katabi vs Entebbe	**3.3**	**(1.22, 8.95)**

[a]CI is confidence interval

Supplementary Table S5: Probability of allergy-related disease (ARD) outcomes at 9 years by latent class membership at year two

			95% CI[a]		
ARD outcomes at 9 years	**Latent class**	**Estimate**	**Lower CI**	**Upper CI**	***P*-value**
Wheeze	LC[b] 1	0.054	0.029	0.098	0.332
	LC 2	0.029	0.013	0.062	
Eczema	LC 1	0.054	0.028	0.103	0.919
	LC 2	0.052	0.030	0.086	
Rhinitis	LC 1	0.045	0.021	0.094	0.979
	LC 2	0.046	0.026	0.080	
SPT-any[c]	LC 1	0.229	0.167	0.305	0.515
	LC 2	0.266	0.216	0.322	
SPT - dermatophagoides	LC 1	0.139	0.090	0.209	0.202
	LC 2	0.203	0.160	0.254	
SPT-cockroach	LC 1	0.064	0.031	0.127	0.092
	LC 2	0.137	0.103	0.181	
SPT-Blomia	LC 1	0.143	0.095	0.210	0.862
	LC 2	0.151	0.113	0.199	
IgE - dermatophagoides	LC 1	0.227	0.164	0.306	0.069
	LC 2	0.337	0.283	0.396	
IgE-cockroach	LC 1	0.277	0.209	0.357	0.207
	LC 2	0.354	0.299	0.414	
IgE-any[d]	LC 1	0.353	0.279	0.436	**0.024**
	LC 2	0.501	0.440	0.562	

[a]CI is confidence interval

[b]LC is latent class. LC 1 represents the class with higher probabilities for various infections as compared to LC 2, however the profile of infections changes at each time period.

[c]SPT-any is skin prick test reactivity to any of *Dermatophagoides* mix, *Blomia tropicalis*, German cockroach, cat, mould, grass pollen, Bermuda grass or peanut.

[d]IgE-any is allergen-specific plasma IgE (asIgE) to any of house dust mite (HDM, *Dermatophagoides pteronyssinus*) or German cockroach.
Estimates are adjusted for maternal area of residence, and maternal and household socio-economic status at enrolment.

Supplementary Table S6: Probability of allergy-related disease (ARD) outcomes at 9 years by latent class membership at year three

			95% CI[a]		
ARD outcomes at 9 years	**Latent class**	**Estimate**	**Lower CI**	**Upper CI**	***P*-value**
Wheeze	LC[b] 1	0.04	0.029	0.053	0.672
	LC 2	0.03	0.004	0.150	
Eczema	LC 1	0.05	0.040	0.068	0.428
	LC 2	0.02	0.003	0.164	
Rhinitis	LC 1	0.05	0.036	0.063	0.888
	LC 2	0.04	0.010	0.161	
SPT-any[c]	LC 1	0.25	0.227	0.282	0.844
	LC 2	0.24	0.140	0.382	
SPT - dermatophagoides	LC 1	0.18	0.160	0.208	0.846
	LC 2	0.17	0.089	0.304	
SPT-cockroach	LC 1	0.10	0.085	0.124	0.067
	LC 2	0.19	0.105	0.319	
SPT-Blomia	LC 1	0.15	0.127	0.172	0.356
	LC 2	0.20	0.110	0.333	
IgE - dermatophagoides	LC 1	0.29	0.265	0.323	0.591
	LC 2	0.33	0.212	0.477	
IgE-cockroach	LC 1	0.33	0.298	0.358	0.956
	LC 2	0.32	0.205	0.470	
IgE-any[d]	LC 1	0.44	0.412	0.475	0.904
	LC 2	0.45	0.316	0.596	

[a]CI is confidence interval
[b]LC is latent class. LC 1 represents the class with higher probabilities for various infections as compared to LC 2, however the profile of infections changes at each time period.
[c]SPT-any is skin prick test reactivity to any of *Dermatophagoides* mix, *Blomia tropicalis*, German cockroach, cat, mould, grass pollen, Bermuda grass or peanut.
[d]IgE-any is allergen-specific plasma IgE (asIgE) to any of house dust mite (HDM, *Dermatophagoides pteronyssinus*) or German cockroach.
Estimates are adjusted for maternal area of residence, and maternal and household socio-economic status at enrolment.

Supplementary Table S7: Probability of allergy-related disease (ARD) outcomes at 9 years by latent class membership at year four

		Male				Female			
			95% CI[a]				95% CI		
ARD outcomes at 9 years	**Latent class**	**Estimate**	**LowerCI**	**UpperCI**	***P***	**Estimate**	**LowerCI**	**UpperCI**	***P***
Wheeze	LC[b] 1	0.02	0.003	0.158	0.449	0.01	0.001	0.082	0.292
	LC 2	0.05	0.034	0.082		0.03	0.019	0.058	
Eczema	LC 1	0.02	0.002	0.161	0.381	0.05	0.021	0.122	0.950

	LC 2	0.06	0.038	0.086		0.05	0.034	0.084	
Rhinitis	LC 1	0.00	0.000	0.998	0.589	0.00	0.000	.	1.000
	LC 2	0.08	0.053	0.107		0.04	0.025	.	
SPT-any[c]	LC 1	0.21	0.120	0.349	0.364	0.19	0.120	0.286	0.301
	LC 2	0.28	0.236	0.332		0.25	0.203	0.296	
SPT - dermatophagoides	LC 1	0.15	0.072	0.276	0.394	0.12	0.069	0.212	0.240
	LC 2	0.20	0.164	0.250		0.18	0.144	0.226	
SPT-cockroach	LC 1	0.13	0.062	0.253	0.957	0.08	0.041	0.164	0.873
	LC 2	0.13	0.095	0.168		0.09	0.063	0.126	
SPT-Blomia	LC 1	0.17	0.090	0.298	0.900	0.07	0.031	0.149	**0.046**
	LC 2	0.16	0.126	0.207		0.16	0.128	0.206	
IgE - dermatophagoides	LC 1	0.41	0.282	0.548	0.108	0.31	0.220	0.420	0.506
	LC 2	0.28	0.232	0.333		0.27	0.225	0.323	
IgE-cockroach	LC 1	0.44	0.307	0.576	0.084	0.29	0.198	0.394	0.509
	LC 2	0.30	0.247	0.351		0.33	0.279	0.380	
IgE-any[d]	LC 1	0.57	0.431	0.703	0.078	0.45	0.345	0.558	0.670
	LC 2	0.42	0.365	0.475		0.42	0.368	0.475	

[a]CI is confidence interval

[b]LC is latent class. LC 1 represents the class with higher probabilities for various infections as compared to LC 2, however the profile of infections changes at each time period.

[c]SPT-any is skin prick test reactivity to any of *Dermatophagoides* mix, *Blomia tropicalis*, German cockroach, cat, mould, grass pollen, Bermuda grass or peanut.

[d]IgE-any is allergen-specific plasma IgE (asIgE) to any of house dust mite (HDM, *Dermatophagoides pteronyssinus*) or German cockroach.

Estimates are adjusted for maternal area of residence, and maternal and household socio-economic status at enrolment.

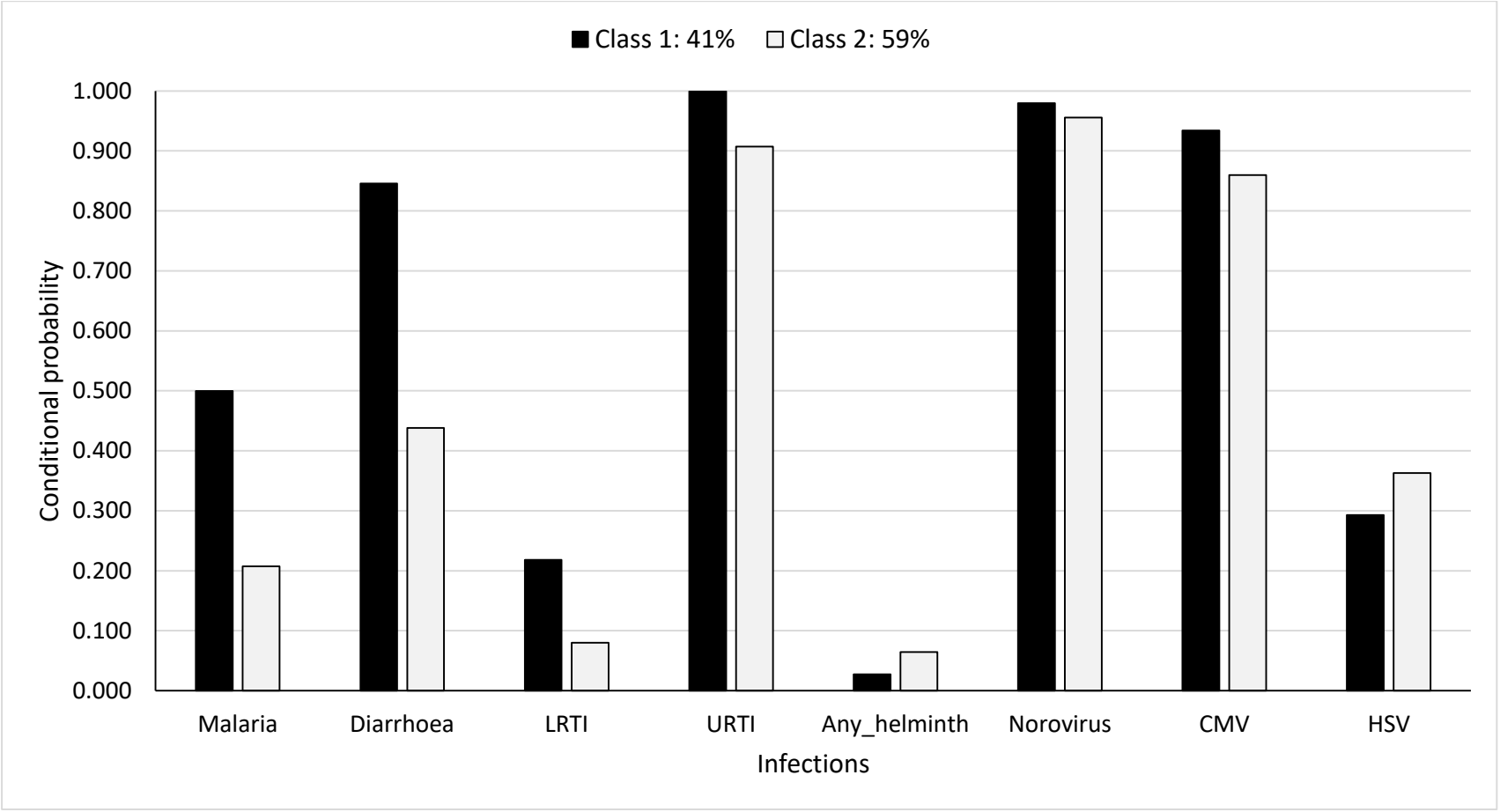
Supplementary Table S8: Probability of allergy-related disease (ARD) outcomes at 9 years by latent class membership at year five

			95% CI[a]		
ARD outcomes at 9 years	**Latent class**	**Estimate**	**Lower CI**	**Upper CI**	***P*-value**
Wheeze	LC[b] 1	0.01	0.000	0.242	0.359
	LC 2	0.05	0.033	0.064	
Eczema	LC 1	0.00	0.000		
	LC 2	0.07	0.052		
Rhinitis	LC 1	0.04	0.012	0.124	0.787
	LC 2	0.05	0.034	0.067	
SPT-any[c]	LC 1	0.19	0.113	0.302	0.215
	LC 2	0.27	0.235	0.302	
SPT - dermatophagoides	LC 1	0.11	0.055	0.216	0.140
	LC 2	0.20	0.168	0.227	
SPT-cockroach	LC 1	0.09	0.044	0.187	0.634
	LC 2	0.11	0.092	0.140	
SPT-Blomia	LC 1	0.14	0.081	0.246	0.862
	LC 2	0.15	0.128	0.183	
IgE - dermatophagoides	LC 1	0.41	0.306	0.531	**0.027**
	LC 2	0.27	0.237	0.308	
IgE-cockroach	LC 1	0.44	0.328	0.555	**0.031**
	LC 2	0.30	0.262	0.334	
IgE-any[d]	LC 1	0.59	0.468	0.698	**0.013**
	LC 2	0.41	0.372	0.449	

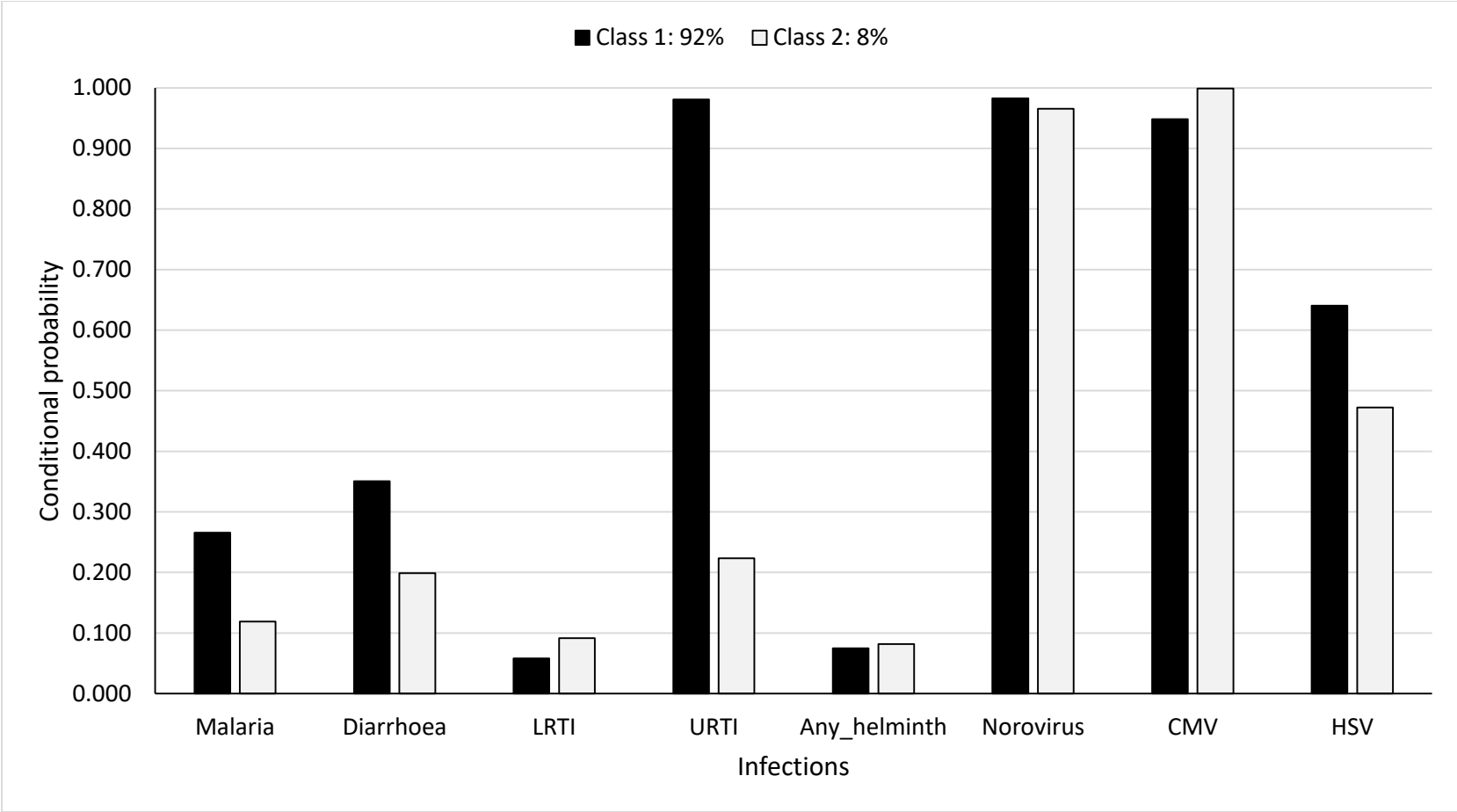
[a]CI is confidence interval
[b]LC is latent class. LC 1 represents the class with higher probabilities for various infections as compared to LC 2, however the profile of infections changes at each time period.
[c]SPT-any is skin prick test reactivity to any of *Dermatophagoides* mix, *Blomia tropicalis*, German cockroach, cat, mould, grass pollen, Bermuda grass or peanut.
[d]IgE-any is allergen-specific plasma IgE (asIgE) to any of house dust mite (HDM, *Dermatophagoides pteronyssinus*) or German cockroach.
Estimates are adjusted for maternal area of residence, and maternal and household socio-economic status at enrolment.

5.3.2. Supplementary figures for Research Paper 3

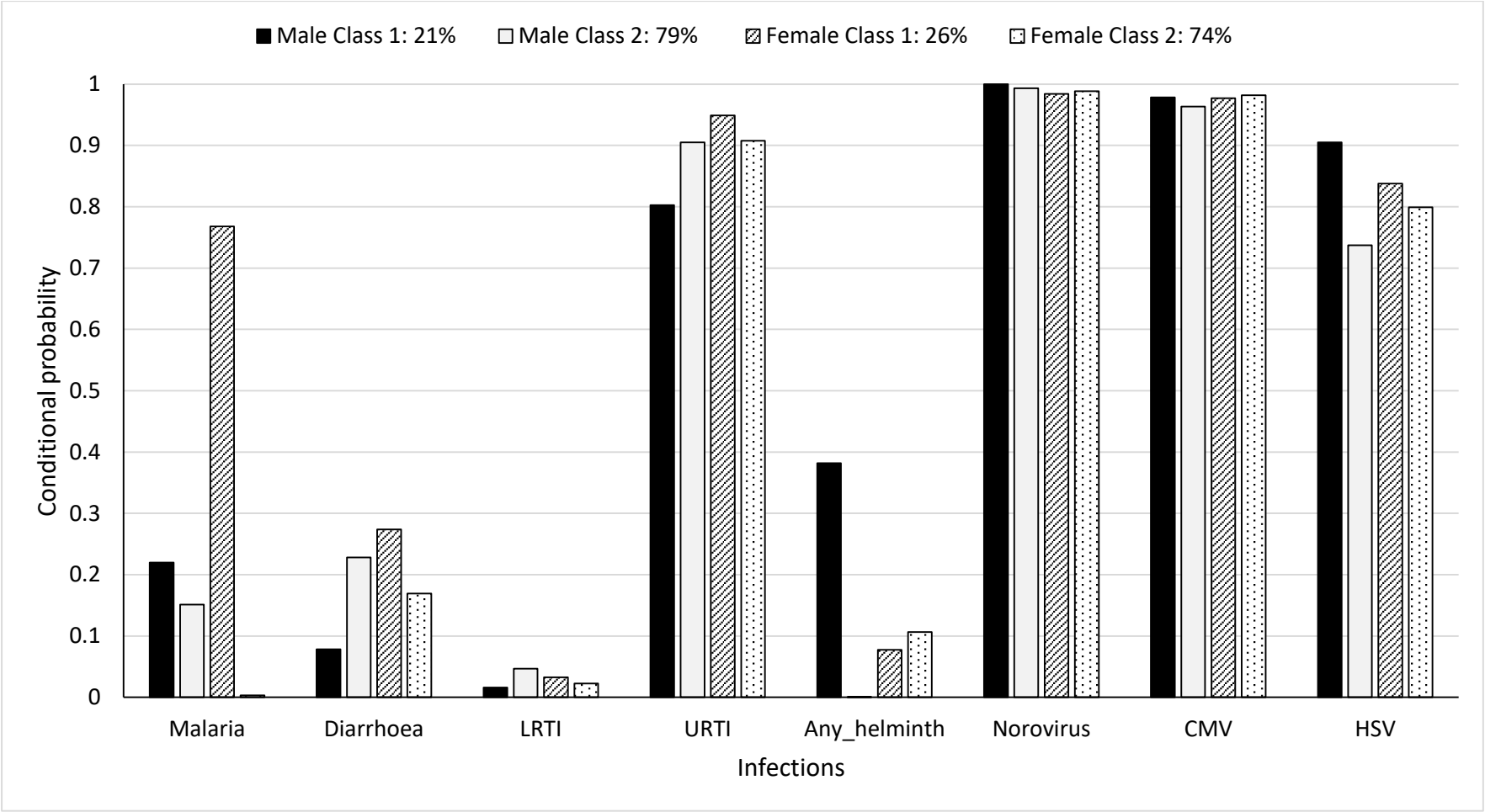
Supplementary Figure S1: Probability of having illnesses/infections conditional on latent class membership during the second year



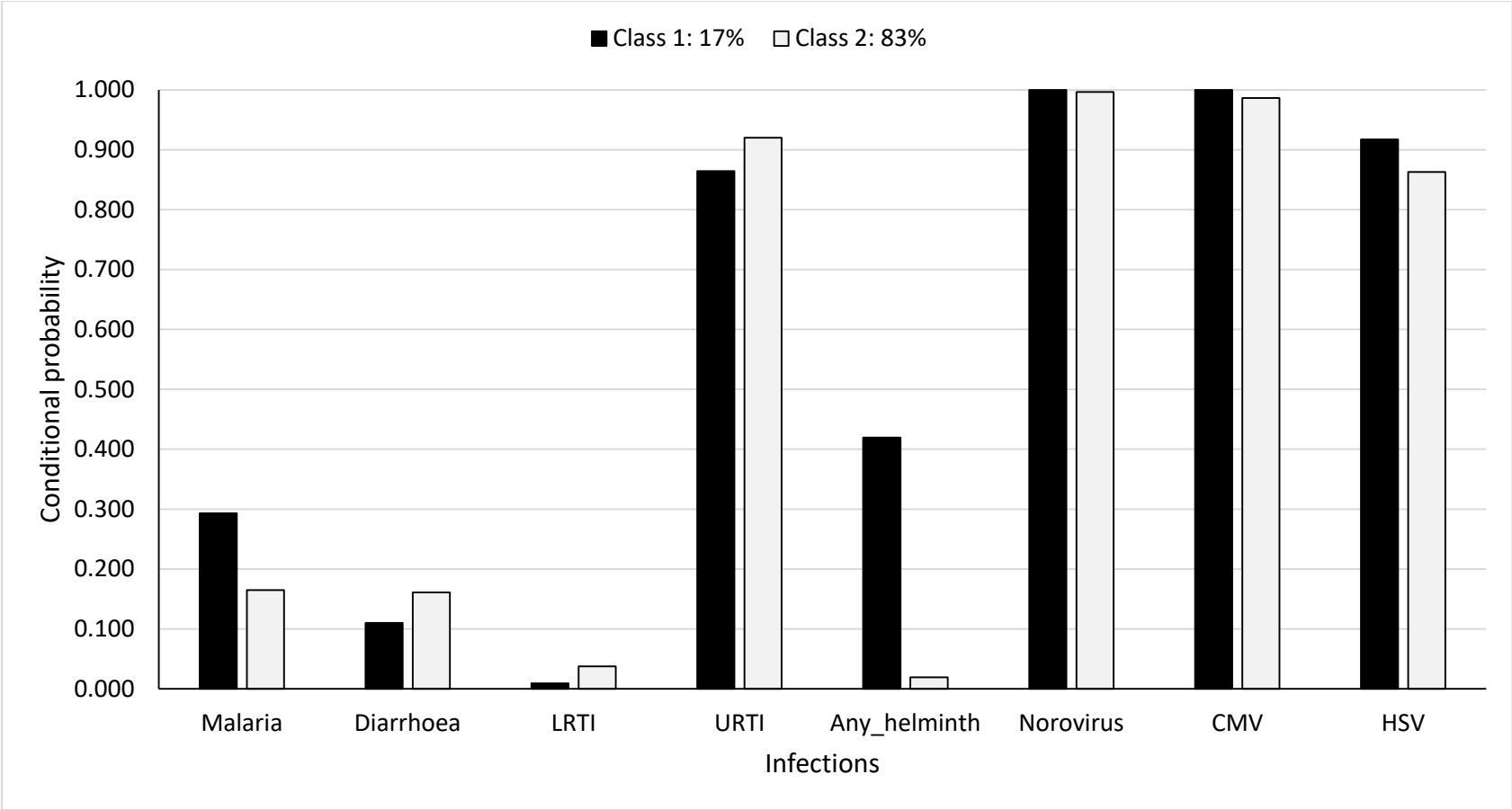
Supplementary Figure S2: Probability of having illnesses/infections conditional on latent class membership during the third year



Supplementary Figure S3: Probability of having illnesses/infections conditional on latent class membership during the fourth year



Supplementary Figure S4: Probability of having illnesses/infections conditional on latent class membership during the fifth year



Chapter 6. Summary, Discussion and Conclusions

6.1. Summary and discussion of main findings

6.1.1. Research paper 1 (Chapter 3)

Research paper 1 shows the use of longitudinal data analysis approaches (the linear mixed model and three-way component analysis models) to understand the dynamics of vaccine responses and how they differ between infants born of mothers with or without LTBI, in the Infant BCG study[98]. Differences in immune sensitisation and patterns of post-BCG cytokine and antibody responses to mycobacterial antigens were compared between infants born of mothers with or without LTBI.

This study refuted the hypothesis that maternal LTBI influences fetal sensitisation and the neonatal response to BCG immunisation. PCA in cord blood showed IL-10 and TNF-α tending to group separately from the other cytokines, but there was no evidence that maternal LTBI was associated with a differing pattern of response, or that differences in these cord blood profiles impacted the subsequent infant response to BCG. Possible explanations for observations from this study include the high proportion of infants who demonstrated cord blood cytokine responses to mycobacterial antigens, regardless of maternal LTBI status and that the BCG stimulus could have been sufficiently strong to override effects of prior *in utero* exposures[98]. Results from this study are consistent with those from a previous study in South Africa which showed that maternal HIV infection had an effect on infant immune responses in the presence of maternal *M.tb* sensitisation at birth, however, these effects were not maintained post immunisation with BCG[99].

Research paper 1 further showed that the peak of infant T cell response to PPD was sustained between ages 10 to 24 weeks, this was later than the 6 to 10 weeks period previously reported

by Soares and others[21]. This is an important finding because it informs future prime-boost strategies suggesting that they should be given much later than has previously been thought.

The policy implication from the findings of research paper 1 is that maternal latent *M.tb* infection is not the reason for reduced effectiveness of BCG in tropical countries and therefore, all infants are likely to benefit from BCG immunisation regardless of their maternal LTBI status.

6.1.2. Research paper 2 (Chapter 4)

Research paper 2 shows an extension of the univariate models discussed in Research paper 1 into a joint modelling framework for multivariate longitudinal immuno-epidemiological data, with application to the IBS case study. The paper discusses a pairwise joint modelling approach and its benefits over simpler statistical analysis approaches. This approach is considered because it accounts for both the correlation between measurements from an individual over time as well as the correlation between multiple outcomes assessed at each time point.

This study showed that the pairwise approach had utility in a situation where fitting a full multivariate model with maximum likelihood estimation had failed. Results from this study also showed that the parameter estimates from the pairwise approach had better precision than those from the univariate mixed models, although it has been advanced elsewhere that gains in precision should not be the key motivation for choice of this approach[66].

Research paper 2 also highlighted a major advantage of the pairwise approach as it permitted the opportunity to carry out a joint statistical test for the effect of LTBI on the evolution of all cytokine response outcomes together. This overcomes the multiple testing problem which is

common with multiple outcome data measured at several time points. The results from this joint statistical test supported the conclusion that there were no differences in post-BCG immune responses between infants born of mothers with or without LTBI.

This study also showed that the pairwise approach enabled estimation of association structures of longitudinal outcomes. Results show that the initial patterns from PCA of cord blood responses were not reflected in the profile of responses that developed post-BCG as indicated by the PCA on the correlation matrix of random intercepts. This improved our interpretation of the data by suggesting that probably the effect of BCG immunisation was potent enough to over-ride patterns established *in utero*.

6.1.3. Research paper 3 (Chapter 5)

Research paper 3 illustrates the use of a novel person-centred statistical analysis approach, latent class analysis (LCA), to characterise the infection-exposure of participants in the EMaBS using data on malaria, diarrhoea, upper and lower respiratory tract infections, intestinal helminths, seroprevalence of HSV, CMV and norovirus, collected during each of the first five years of life. Prevalences of ARDs (wheeze, eczema, rhinitis, SPT positivity and asIgE) at nine years were then compared across latent classes, adjusted for covariates.

This study identified two distinguishable groups of children, during each of the first five years of life, based on their infection-exposure in that year. In the first year of life, latent class (LC) 1 was characterised by higher probabilities of malaria, diarrhoea and LRTI, and membership of this class was associated with lower proportions of wheeze, eczema, rhinitis and SPT positivity for *Dermatophagoides* at age nine years. These findings support the hygiene

hypothesis which states that infections in early life are associated with a lower risk of ARDs in later childhood[100-105].

Research paper 3 further showed that associations between LC membership and ARDs were less consistent for infections in years two to five, and this could be explained by the declining prevalences of common infections. This suggested that infancy is a critical period during which the risk of ARDs in later childhood is established.

This study also showed that at five years, when the prevalence of helminths was highest, LC1 was characterised by a higher probability of helminth infections and membership in this class was associated with higher levels of asIgE responses to crude allergen extracts of dust mite and cockroach, but was not associated with clinical ARDs. These findings were similar to those from an earlier cross-sectional study, in the same geographical area, which showed positive associations between helminth infections and asIgE responses but no associations with clinical ARDs[106].

6.2. Thesis strengths

6.2.1. Statistical methods

This PhD thesis synthesised and applied recently developed robust statistical methods in novel analyses of multivariate longitudinal immuno-epidemiological data. PCA was used for dimension reduction and for investigation of relationships between cord blood cytokine responses. Three-way component analysis was used as an extension of PCA to the longitudinal situation, in order to explore relations between cytokine responses measured at multiple time points following BCG, in the IBS study. Linear mixed effects models were used to study

changes in cytokine responses over time, by mothers' LTBI status, in a flexible manner. The pairwise joint modelling approach was then applied in order to facilitate the study of association structures of the longitudinal cytokine response outcomes and the study of the joint effect of maternal LTBI on all the outcomes together.

For the EMaBS dataset, a robust statistical analysis approach, LCA, was used to construct the unobservable infection-exposure variable using prospective data on multiple observed infections. This approach has an advantage of giving a more complete picture of the underlying construct of infection-exposure as compared to dealing with each observed infection variable separately.

The statistical methods applied in this thesis allow for accommodation of various correlation structures for instance the correlation of longitudinal observations of participants and the correlation of multivariate outcomes at given time points. These methods also help to overcome the multiple testing problem associated with simpler analysis methods which deal with multivariate outcomes separately or those which can only accommodate separate analyses at each time point.

This thesis also highlights the use of multiple imputation and direct likelihood approaches for dealing with missing data. Proper handling of missing data has the advantage of yielding efficient parameter estimates, unbiased standard errors and corresponding confidence intervals[79].

6.2.2. Application of statistical methods to immuno-epidemiological data

This thesis has a major strength of illustrating the use of robust statistical analysis methods and their application to immuno-epidemiological longitudinal case studies. This PhD study

showcases two large case studies with a wealth of data, which allowed for a lot of flexibility in illustrating various statistical methods. Key immuno-epidemiological research questions for each case study had not yet been addressed, providing a unique opportunity to illustrate the use of advanced statistical methods to answer the corresponding question.

Research papers 1 and 2 describe, at length, the use of several statistical methods namely PCA, three-way analysis, linear mixed effects models and the pairwise joint models and their application to the IBS dataset. The immuno-epidemiological research question of the effect of maternal LTBI on the infant immune response following immunisation with BCG is well addressed using the robust statistical procedures as explained in chapters 3 and 4.

Research paper 3 covers an illustration of the use of LCA to address another immuno-epidemiological research topic, the association of early childhood infection-exposure with allergy-related diseases in later childhood. LCA is applied to the EMaBS case study dataset and findings are discussed in chapter 5.

6.2.3. New findings

This thesis highlights the largest study of its kind to date (research paper 1), with a new finding that there was remarkably high early exposure to mycobacterial antigens *in utero* in Uganda, but this exposure had no impact on the infant response to immunisation with BCG.

This thesis also contains new findings related to the association structures of cytokine responses pre- and post-immunisation with BCG (research paper 2). It shows that the initial sensitisation patterns (cord blood PCA) were not reflected in the profile of response post-immunisation with BCG.

Another new finding illustrated in this thesis is the identification of two latent classes of children based on their early childhood infection experience and that exposure to common infections was associated with lower proportions of ARDs in later childhood (research paper 3).

6.3. Thesis limitations

This thesis aimed to illustrate an extension of appropriate statistical modelling methods for longitudinal data with application to immunological case studies from Entebbe, Uganda. The methods highlighted in this thesis are by no means exhaustive, I selected those which I felt were most appropriate to address the scientific research questions and illustrate their application with actual case studies as described.

In this thesis, multiple imputation and direct likelihood approaches were used for dealing with missing data. These approaches assume ignorable missingness and are valid under the missing at random (MAR) mechanism. Whereas such ignorable likelihood analyses are more versatile as compared to simple methods such as complete case analysis, it is not possible to completely rule out nonignorable missingness under the missing not at random (MNAR) mechanism[107]. Analyses under the MNAR mechanism require sensitivity analyses under various assumptions, this could form a fruitful area for subsequent research.

This thesis presents the pairwise joint modelling approach as advantageous especially for the situation when fitting the full multivariate mixed model is not possible or is too time consuming. A limitation of this approach is that it could become more time consuming and harder to implement when the number of multiple outcomes is large resulting in too many

pairs to be dealt with. This challenge can be addressed by utilising faster computers with more dedicated resources to handle the complex iteration procedures.

The LCA approach described in this thesis demonstrates how multiple infections can be used to identify underlying subgroups of children based on their early life infection experience. A limitation of LCA models has been identified as reification[108], that is the treatment of something abstract as a physical thing. With this approach, it may be easy to conclude that the latent classes identified represent the actual types of individuals in the study, yet they simply provide a heuristic representation of the heterogeneity across the dimensions included in the model. A closely related issue is that the effect of sample size on the ability to identify the set of underlying classes still needs further study[108]. There were also some limitations in terms of the available data, for example the seroprevalence of norovirus should have been examined for all samples at each of the first five years for the EMaBS participants, however, this was not the case since only samples which tested negative were examined at subsequent years until a positive response was established. In addition, class membership changed over time, but this was not surprising given the declining prevalences of common infections.

6.4. Future work

This PhD thesis illustrated a number of statistical analysis approaches for dealing with longitudinal multiple outcome immuno-epidemiological data. There are other extensions which can be considered. For instance, in dealing with missing data, analyses under the MNAR mechanism can be considered, these require sensitivity analyses under various assumptions. Such models require specification of the joint distribution of the data and the missing data

mechanism. Three types of models can be classified under this specification namely selection, pattern-mixture and shared-parameter models[107, 109, 110]. Future work will need to consider extensions of such models to longitudinal immuno-epidemiological data.

This thesis considers likelihood-based frequentist methods for illustrating the analysis approaches for longitudinal multiple-outcome immuno-epidemiological data. Rizopoulos and Ghosh[111], Wang and Fan[112], and Wang and Lin[113] among others have illustrated various applications of Bayesian statistical analysis approaches to longitudinal multiple-outcome data. Such methods could be extended to immuno-epidemiological data as future work.

The statistical analysis models considered in this thesis operate under assumptions of parametric distributions. Logarithmic transformations were considered and applied to correct for departures from normality in order to adhere to the restrictive parametric assumptions. Future work could utilise non-parametric or semi-parametric approaches which are not constrained by distributional assumptions.

The EMaBS dataset had a few multiple live births from a single mother. To my knowledge, the LCA approach considered in this thesis has not been extended to clustered data of this kind. Future work will need to consider such extensions.

6.5. Conclusions

This thesis synthesised and applied statistical analysis methods that are well-established and robust in other contexts in novel analyses of multivariate longitudinal immuno-epidemiological data with case studies from Entebbe, Uganda.

The major conclusions from this PhD research are

1. Data from the IBS suggested remarkably high early sensitisation to mycobacterial antigens *in utero*, in Uganda, but there was no impact of this sensitisation on the infant immune response to immunisation with BCG. These data did not support the hypothesis that maternal LTBI results in an impaired response to BCG immunisation, in Ugandan infants. BCG vaccination at or shortly after birth is likely to be beneficial to all infants, regardless of maternal LTBI status.
2. Certain scientific questions cannot be addressed by simple statistical analysis approaches. The pairwise joint modelling approach is essential in situations where it is desired to determine the effect of a treatment on the joint evolution of all outcomes and when it is desired to study relations between evolutions of longitudinal multivariate outcomes. Consequently, the pairwise joint modelling approach improves our understanding and interpretation of longitudinal immuno-epidemiological data.
3. In a tropical African birth cohort, the EMaBS, children were classified into two latent classes based on their early childhood infection experience. Increased exposure to common infections, especially during the first year of life, was associated with lower proportions of ARDs in later childhood.

Overall, this thesis presents robust statistical methods, illustrates their application using immuno-epidemiological case studies in Entebbe, Uganda, answers key scientific research questions of interest and improves the interpretation and understanding of high-dimensional immuno-epidemiological data.

References

1. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*: CRC Press; 2008.
2. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. *Journal of Thoracic Disease* 2015; **7**: E537-E40.
3. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* 2004; **5**: 381-98.
4. McCoach DB, Kaniskan B. Using time-varying covariates in multilevel growth models. *Frontiers in psychology* 2010; **1**: 17.
5. Skup M, Zhu H, Zhang H. Multiscale Adaptive Marginal Analysis of Longitudinal Neuroimaging Data with Time-Varying Covariates. *Biometrics* 2012; **68**: 1083-92.
6. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Second ed. United States: Oxford University Press Inc., New York; 2002.
7. Molenberghs G, Kenward M. *Missing data in clinical studies*: John Wiley & Sons; 2007.
8. Hellriegel B. Immunoepidemiology--bridging the gap between immunology and epidemiology. *Trends in parasitology* 2001; **17**: 102-6.
9. Krause PJ, Kavathas PB, Ruddle NH. *Immunoepidemiology*: Springer International Publishing; 2019.
10. Genser B, Cooper PJ, Yazdanbakhsh M, Barreto ML, Rodrigues LC. A guide to modern statistical analysis of immunological data. *BMC immunology* 2007; **8**: 27.
11. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research* 2014; **23**: 42-59.
12. Genser B, Fischer JE, Figueiredo CA, et al. Applied immuno-epidemiological research: an approach for integrating existing knowledge into the statistical analysis of multiple immune markers. *BMC Immunol* 2016; **17**: 11.
13. Härtel C, Adam N, Strunk T, Temming P, Müller-Steinhardt M, Schultz C. Cytokine responses correlate differentially with age in infancy and early childhood. *Clinical & Experimental Immunology* 2005; **142**: 446-53.
14. Geiger S, Massara C, Bethony J, Soboslay P, Correa-Oliveira R. Cellular responses and cytokine production in post-treatment hookworm patients from an endemic area in Brazil. *Clinical & Experimental Immunology* 2004; **136**: 334-40.
15. Tanaka H, Demeure CE, Rubio M, Delespesse G, Sarfati M. Human monocyte–derived dendritic cells induce naive T cell differentiation into T helper cell type 2 (Th2) or Th1/Th2 effectors: Role of stimulator/responder ratio. *The Journal of experimental medicine* 2000; **192**: 405-12.
16. Rachmiel M, Bloch O, Bistritzer T, et al. TH1/TH2 cytokine balance in patients with both type 1 diabetes mellitus and asthma. *Cytokine* 2006; **34**: 170-6.
17. Tisato V, Zauli G, Gianesini S, et al. Modulation of circulating cytokine-chemokine profile in patients affected by chronic venous insufficiency undergoing surgical hemodynamic correction. *Journal of immunology research* 2014; **2014**.
18. van den Biggelaar AH, van Ree R, Rodrigues LC, et al. Decreased atopy in children infected with Schistosoma haematobium: a role for parasite-induced interleukin-10. *The Lancet* 2000; **356**: 1723-7.
19. Black GF, Weir RE, Floyd S, et al. BCG-induced increase in interferon-gamma response to mycobacterial antigens and efficacy of BCG vaccination in Malawi and the UK: two randomised controlled studies. *The Lancet* 2002; **359**: 1393-401.
20. Lekkou A, Mouzaki A, Siagris D, Ravani I, Gogos CA. Serum lipid profile, cytokine production, and clinical outcome in patients with severe sepsis. *Journal of critical care* 2014; **29**: 723-7.

21. Soares AP, Kwong Chung CK, Choice T, et al. Longitudinal changes in CD4+ T-cell memory responses induced by BCG vaccination of newborns. *The Journal of infectious diseases* 2013; **207**: 1084-94.
22. Lalor MK, Floyd S, Gorak-Stolinska P, et al. BCG vaccination induces different cytokine profiles following infant BCG vaccination in the UK and Malawi. *Journal of Infectious Diseases* 2011; **204**: 1075-85.
23. Mawa PA, Nkurununngi G, Egesa M, et al. The impact of maternal infection with Mycobacterium tuberculosis on the infant response to bacille Calmette–Guérin immunization. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2015; **370**: 20140137.
24. Nematollahi LR, Kitabchi AE, Stentz FB, et al. Proinflammatory cytokines in response to insulin-induced hypoglycemic stress in healthy subjects. *Metabolism* 2009; **58**: 443-8.
25. Bell ML, King MT, Fairclough DL. Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data: summary measures versus summary statistics. *SAGE Open* 2014; **4**: 2158244014534858.
26. Hand DJ, Crowder MJ. *Practical longitudinal data analysis*: CRC Press; 1996.
27. McGuinness D, Bennett S, Riley E. Statistical analysis of highly skewed immune response data. *Journal of immunological methods* 1997; **201**: 99-114.
28. Bennett S, Riley E. The statistical analysis of data from immunoepidemiological studies. *Journal of immunological methods* 1992; **146**: 229-39.
29. Panda A, Chen S, Shaw AC, Allore HG. Statistical approaches for analyzing immunologic data of repeated observations: a practical guide. *Journal of immunological methods* 2013; **398**: 19-26.
30. Oort FJ. Stochastic three-mode models for mean and covariance structures. *British journal of mathematical and statistical psychology* 1999; **52**: 243-72.
31. Oort FJ. Three-mode models for multivariate longitudinal data. *British journal of mathematical and statistical psychology* 2001; **54**: 49-78.
32. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* 2000; **56**: 1047-54.
33. Timmerman ME, Kiers HA. Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika* 2003; **68**: 105-21.
34. Timmerman ME, Kiers HA. Three-way component analysis with smoothness constraints. *Computational statistics & data analysis* 2002; **40**: 447-70.
35. Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* 2006; **62**: 424-31.
36. Kroonenberg PM. *Applied multiway data analysis*: John Wiley & Sons; 2008.
37. Lanza ST, Bray BC, Collins LM. An introduction to latent class and latent transition analysis. In: Schinka JA, Velicer WF, Weiner IB, editors. *Handbook of Psychology, Second Edition*. Hoboken, NJ: John Wiley & Sons Inc.; 2013. p. 691 - 716.
38. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**: 215-31.
39. Collins LM, Lanza ST. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Hoboken, NJ: John Wiley & Sons; 2010.
40. Mawa PA, Nkurununngi G, Egesa M, et al. The impact of maternal infection with Mycobacterium tuberculosis on the infant response to bacille Calmette-Guerin immunization. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2015; **370**.
41. Webb EL, Mawa PA, Ndibazza J, et al. Effect of single-dose anthelmintic treatment during pregnancy on an infant's response to immunisation and on susceptibility to infectious diseases in infancy: a randomised, double-blind, placebo-controlled trial. *Lancet (London, England)* 2011; **377**: 52-62.
42. Elliott AM, Kizza M, Quigley MA, et al. The impact of helminths on the response to immunization and on the incidence of infection and disease in childhood in Uganda: design of a

randomized, double-blind, placebo-controlled, factorial trial of deworming interventions delivered in pregnancy and early childhood [ISRCTN32849447]. *Clinical trials (London, England)* 2007; **4**: 42-57.
43. Lule SA, Mpairwe H, Nampijja M, et al. Life-course of atopy and allergy-related disease events in tropical sub-Saharan Africa: A birth cohort study. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology* 2017; **28**: 377-83.
44. Bach J-F. The effect of infections on susceptibility to autoimmune and allergic diseases. *New England journal of medicine* 2002; **347**: 911-20.
45. Strachan DP. Hay fever, hygiene, and household size. *BMJ: British Medical Journal* 1989; **299**: 1259.
46. Bloomfield S, Stanwell-Smith R, Crevel R, Pickup J. Too clean, or not too clean: the hygiene hypothesis and home hygiene. *Clinical & Experimental Allergy* 2006; **36**: 402-25.
47. Schaub B, Lauener R, von Mutius E. The many faces of the hygiene hypothesis. *Journal of allergy and clinical immunology* 2006; **117**: 969-77.
48. Entebbe Mother and Baby Study. *The Entebbe Mother and Baby Study*. 2020 [cited 2020 July 30]; Available from: http://emabs.lshtm.ac.uk
49. MRC/UVRI & LSHTM Uganda Research Unit. *About MRC/UVRI & LSHTM Uganda Research Unit*. 2020 [cited 2020 July 30]; Available from: https://www.mrcuganda.org/
50. Jolliffe IT. Principal components in regression analysis. *Principal component analysis*: Springer; 1986. p. 129-55.
51. Bro R, Smilde AK. Principal component analysis. *Analytical Methods* 2014; **6**: 2812-31.
52. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometrics and intelligent laboratory systems* 1987; **2**: 37-52.
53. Wood F, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intel Lab Syst* 1987; **2**: 37-52.
54. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279-311.
55. Carroll JD, Chang J-J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* 1970; **35**: 283-319.
56. Harshman RA. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.
57. Giordani P, Kiers HA, Del Ferraro MA. Three-way component analysis using the R package ThreeWay. *Journal of Statistical Software* 2014; **57**: 1-23.
58. Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S. *Analysis of longitudinal data*: Oxford University Press; 2002.
59. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*: Springer Science & Business Media; 2009.
60. FitzMaurice Garrett M, Laird Nan M, Ware James H. Applied longitudinal analysis. John Wiley & Sons; 2004.
61. Weiss RE. *Modeling longitudinal data*: Springer Science & Business Media; 2005.
62. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. Handbooks of modern statistical methods: Longitudinal data analysis. New York: Taylor & Francis Group; 2009.
63. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in medicine* 1998; **17**: 1261-91.
64. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982: 963-74.
65. Molenberghs G, Verbeke G. *Linear mixed models for longitudinal data*: Springer; 2000.
66. Fieuws S, Verbeke G, Molenberghs G. Random-effects models for multivariate repeated measures. *Statistical methods in medical research* 2007; **16**: 387-97.
67. Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS procedure for latent class analysis. *Structural equation modeling: a multidisciplinary journal* 2007; **14**: 671-94.

68. Nylund-Gibson K, Choi AY. Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science* 2018; **4**: 440.
69. Velicer WF, Martin RA, Collins LM. Latent transition analysis for longitudinal data. *Addiction* 1996; **91**: 197-210.
70. Ryoo JH, Wang C, Swearer SM, Hull M, Shi D. Longitudinal model building using latent transition analysis: an example using school bullying data. *Frontiers in psychology* 2018; **9**: 675.
71. Chung H, Park Y, Lanza ST. Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. *Statistics in Medicine* 2005; **24**: 2895-910.
72. Hansen WB, Graham JW. Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive medicine* 1991; **20**: 414-30.
73. Swanson SA, Lindenberg K, Bauer S, Crosby RD. A Monte Carlo investigation of factors influencing latent class analysis: An application to eating disorder research. *International Journal of Eating Disorders* 2012; **45**: 677-84.
74. Nagin DS. *Group-based modeling of development*: Harvard University Press; 2005.
75. Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* 1997; **92**: 1375-86.
76. Wang C-P, Hendricks Brown C, Bandeen-Roche K. Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association* 2005; **100**: 1054-76.
77. Lanza ST, Tan X, Bray BC. Latent class analysis with distal outcomes: A flexible model-based approach. *Structural equation modeling: a multidisciplinary journal* 2013; **20**: 1-26.
78. Dziak JJ, Bray BC, Wagner AT. LCA_Distal_BCH SAS Macro Users' Guide (Version 1.1). *The Methodology Center, University Park, PA, Penn State* 2017.
79. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine* 2011; **30**: 377-99.
80. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581-92.
81. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology* 2017; **9**: 157.
82. Rubin DB. *Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse*. Proceedings of the survey research methods section of the American Statistical Association; 1978: American Statistical Association; 1978. p. 20-34.
83. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical methods in medical research* 2007; **16**: 199-218.
84. Schafer JL. Multiple imputation: a primer. *Statistical methods in medical research* 1999; **8**: 3-15.
85. Rubin DB. Multiple imputation after 18+ years. *Journal of the American statistical Association* 1996; **91**: 473-89.
86. van Ginkel JR, Kroonenberg PM. Evaluation of multiple-imputation procedures for three-mode component models. *Journal of Statistical Computation and Simulation* 2017; **87**: 3059-81.
87. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics* 2002; **11**: 437-57.
88. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of multilevel missing data: An introduction to the R package pan. *Sage Open* 2016; **6**: 2158244016668220.
89. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis*: CRC press; 2013.
90. Bodner TE. What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal* 2008; **15**: 651-75.
91. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science* 2007; **8**: 206-13.

92. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical science* 1992; **7**: 457-72.
93. Van Ginkel JR, Kroonenberg PM. Using generalized procrustes analysis for multiple imputation in principal component analysis. *Journal of classification* 2014; **31**: 242-69.
94. Gower JC. Generalized procrustes analysis. *Psychometrika* 1975; **40**: 33-51.
95. Green BF. The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika* 1952; **17**: 429-40.
96. Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods* 2001; **6**: 330.
97. Interuniversity Institute for Biostatistics and statistical Bioinformatics. *A SAS MACRO FOR FITTING A MULTIVARIATE LINEAR MIXED MODEL USING THE PAIRWISE APPROACH*. 2006, December 13 [cited; Available from: https://ibiostat.be/online-resources/longitudinal#MLMMpw
98. Lubyayi L, Mawa PA, Nabakooza G, et al. Maternal Latent Mycobacterium tuberculosis Does Not Affect the Infant Immune Response Following BCG at Birth: An Observational Longitudinal Study in Uganda. *Frontiers in Immunology* 2020; **11**.
99. Jones CE, Hesseling AC, Tena-Coki NG, et al. The impact of HIV exposure and maternal Mycobacterium tuberculosis infection on infant immune responses to bacille Calmette-Guérin vaccination. *AIDS (London, England)* 2015; **29**: 155.
100. Strachan DP. Family size, infection and atopy: the first decade of the'hygiene hypothesis'. *Thorax* 2000; **55**: S2.
101. Yazdanbakhsh M, Matricardi PM. Parasites and the hygiene hypothesis. *Clinical reviews in allergy & immunology* 2004; **26**: 15-23.
102. Bjorksten B, Naaber P, Sepp E, Mikelsaar M. The intestinal microflora in allergic Estonian and Swedish 2-year-old children. *Clinical and experimental allergy* 1999; **29**: 342-6.
103. Martinez FD, Holt PG. Role of microbial burden in aetiology of allergy and asthma. *The Lancet* 1999; **354**: SII12-SII5.
104. Illi S, von Mutius E, Lau S, et al. Early childhood infectious diseases and the development of asthma up to school age: a birth cohort study. *Bmj* 2001; **322**: 390-5.
105. Loss GJ, Depner M, Hose AJ, et al. The early development of wheeze. Environmental determinants and genetic susceptibility at 17q21. *American journal of respiratory and critical care medicine* 2016; **193**: 889-97.
106. Nkurunungi G, Lubyayi L, Versteeg SA, et al. Do helminth infections underpin urban-rural differences in risk factors for allergy-related outcomes? *Clinical & Experimental Allergy* 2019; **49**: 663-76.
107. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test* 2009; **18**: 1-43.
108. Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science* 2013; **14**: 157-68.
109. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the american statistical association* 1995; **90**: 1112-21.
110. Molenberghs G, Michiels B, Kenward MG, Diggle PJ. Monotone missing data and pattern-mixture models. *Statistica Neerlandica* 1998; **52**: 153-61.
111. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine* 2011; **30**: 1366-80.
112. Wang W-L, Fan T-H. Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs samplers. *Journal of Multivariate Analysis* 2012; **105**: 300-10.
113. Wang W-L, Lin T-I. Bayesian analysis of multivariate t linear mixed models with missing responses at random. *Journal of Statistical Computation and Simulation* 2015; **85**: 3594-612.

Appendices

Appendix 1: Wits Human Research Ethics Committee Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R14/49 Mr L Lubyayi

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M180613

NAME: (Principal Investigator)	Mr L Lubyayi
DEPARTMENT:	School of Public Health Medical School University
PROJECT TITLE:	Statistical modeling approaches for longitudinal multiple outcome data from immuno-epidemiological studies in Entebbe, Uganda
DATE CONSIDERED:	29/06/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Professor J Levin
APPROVED BY:	Professor CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	15/08/2018

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

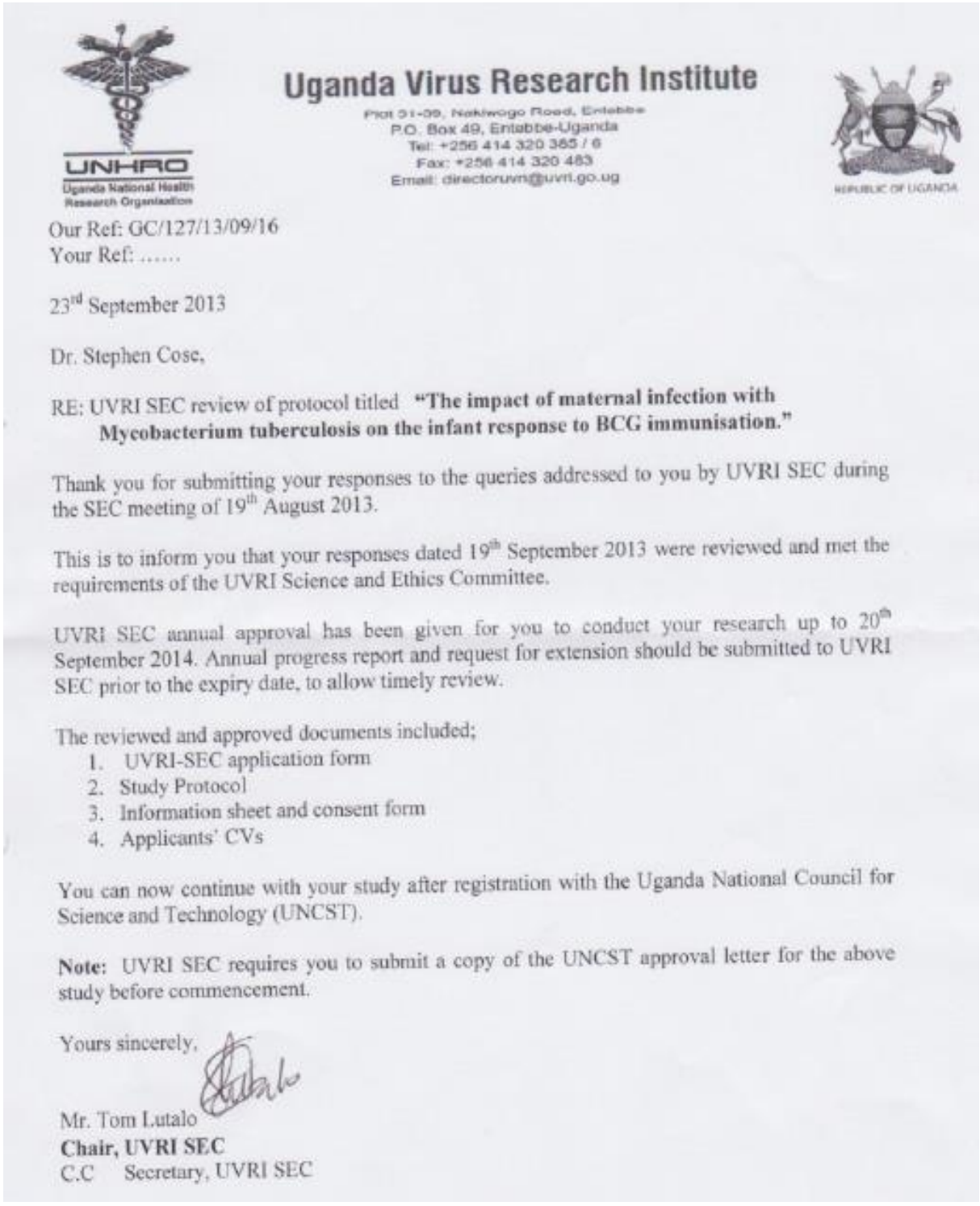
DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.
I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. I agree to submit a yearly progress report. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in June and will therefore be due in the month of June each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature — Date: 30 / 06 / 2018

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 2: The Infant BCG Study (IBS) – copies of ethical approvals



UNHRO
Uganda National Health Research Organisation

Uganda Virus Research Institute

Plot 51-59, Nakiwogo Road, Entebbe
P.O. Box 49, Entebbe-Uganda
Tel: +256 414 320 385 / 6
Fax: +256 414 320 483
Email: directoruvri@uvri.go.ug

REPUBLIC OF UGANDA

Our Ref: GC/127/13/09/16
Your Ref: .....

23rd September 2013

Dr. Stephen Cose,

RE: UVRI SEC review of protocol titled **"The impact of maternal infection with Mycobacterium tuberculosis on the infant response to BCG immunisation."**

Thank you for submitting your responses to the queries addressed to you by UVRI SEC during the SEC meeting of 19th August 2013.

This is to inform you that your responses dated 19th September 2013 were reviewed and met the requirements of the UVRI Science and Ethics Committee.

UVRI SEC annual approval has been given for you to conduct your research up to 20th September 2014. Annual progress report and request for extension should be submitted to UVRI SEC prior to the expiry date, to allow timely review.

The reviewed and approved documents included;

1. UVRI-SEC application form
2. Study Protocol
3. Information sheet and consent form
4. Applicants' CVs

You can now continue with your study after registration with the Uganda National Council for Science and Technology (UNCST).

Note: UVRI SEC requires you to submit a copy of the UNCST approval letter for the above study before commencement.

Yours sincerely,

Mr. Tom Lutalo
Chair, UVRI SEC
C.C Secretary, UVRI SEC

*Approvals for continuation of the study were sought for and obtained each year

Uganda National Council for Science and Technology

(Established by Act of Parliament of the Republic of Uganda)

17/02/2014

Our Ref: HS 1526

Dr. Stephen Cose
MRC/Uganda Virus Research Institute on AIDS
Uganda Virus Research Institute
Entebbe

Re: Research Approval: **The impact of maternal infection with Mycobacterium tuberculosis on the infant response to BCG Immunization**

I am pleased to inform you that **20/01/2014**, the Uganda National Council for Science and Technology (UNCST) approved the above referenced research project. The Approval of the research project is for the period of **20/01/2014** to **20/01/2017**.

Your research registration number with the UNCST is **HS 1526**. Please, cite this number in all your future correspondences with UNCST in respect of the above research project.

As Principal Investigator of the research project, you are responsible for fulfilling the following requirements of approval:

1. All co-investigators must be kept informed of the status of the research.
2. Changes, amendments, and addenda to the research protocol or the consent form (where applicable) must be submitted to the designated local Institutional Review Committee (IRC) or Lead Agency for re-review and approval **prior** to the activation of the changes. UNCST must be notified of the approved changes within five working days.
3. For clinical trials, all serious adverse events must be reported promptly to the designated local IRC for review with copies to the National Drug Authority.
4. Unanticipated problems involving risks to research subjects/participants or other must be reported promptly to the UNCST. New information that becomes available which could change the risk/benefit ratio must be submitted promptly for UNCST review.
5. Only approved study procedures are to be implemented. The UNCST may conduct impromptu audits of all study records.
6. A progress report must be submitted electronically to UNCST within four weeks after every 12 months. Failure to do so may result in termination of the research project.

Below is a list of documents approved with this application:

	Document Title	Language	Version	Version Date
1	Research proposal	English	1.1	31 July 2013
2	Consent Form	English, Luganda	1.1	6 Jan 2014
3	Consent for Storage	English, Luganda	1.1	6 Jan 2014

Yours sincerely,

Leah Nawegulo Omongo
for: Executive Secretary
UGANDA NATIONAL COUNCIL FOR SCIENCE AND TECHNOLOGY

cc Chair, Uganda Virus Research Institute SEC, Entebbe

LOCATION/CORRESPONDENCE

Plot 6 Kimera Road, Ntinda
P. O. Box 6884
KAMPALA, UGANDA

COMMUNICATION

TEL: (256) 414 705500
FAX: (256) 414-234579
EMAIL: info@uncst.go.ug
WEBSITE: http://www.uncst.go.ug

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

Observational / Interventions Research Ethics Committee

Stephen Cose
Lecturer
CR / ITD
LSHTM

13 January 2014

Dear Dr. Cose,

Submission Title: The impact of maternal infection with Mycobacterium tuberculosis on the infant response to BCG immunisation

LSHTM Ethics Ref: 7104

Thank you for your letter of 8 January 2014, responding to the Observational Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval for the amendment having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Information Sheet	Consent Form UG V1.0.doc	7/31/2013	1.0
Information Sheet	Information Sheet V1.1.doc	1/6/2014	1.1
Protocol / Proposal	Protocol V1.1.doc	9/19/2013	1.1

After ethical review

Any subsequent changes to the application must be submitted to the Committee via an Amendment form onthe online application website. The Principal Investigator is reminded that all studies are also required to notify the ethics committee of any serious adverse events which occur during the project via anAdverse Event form on the online application website. At the end of the study, please notify the committee via anEnd of Study form onthe online application website.

Yours sincerely,

Professor John Dil Porter
Chair

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

Appendix 3: The Entebbe Mother and Baby Study (EMaBS) – copies of ethical approvals

Tel: (256) 41 - 320631 (Director)
(256) 41 - 320385/6 (General)
Fax: (256) 41 - 320483
E-mail: arbovir@infocom.co.ug

THE REPUBLIC OF UGANDA

UGANDA VIRUS RESEARCH INSTITUTE
P. O. BOX 49, ENTEBBE (U)

Our Ref: GC/127
Your Ref:

18 May 2001

Dr. A. Elliott
TB Project
Entebbe

Dear Dr. Elliott

The UVRI Science and Ethics Committee reviewed your proposal "The Impact of helminths on the response to immunization and on susceptibility to infectious diseases in childhood in Uganda' this week. Members welcomed the revised proposal and your responses to the Committee's previous concerns.

I am pleased to state that the Committee had no further concerns with the proposal, and I can give Institutional Approval for this study. May I take this opportunity to wish you every success with this project.

Yours sincerely

Dr. S. D. K. Sempala
Director

*Approvals for continuation of the study were sought for and obtained each year

Uganda National Council For Science and Technology
(Established by Act of Parliament of the Republic of Uganda)

Your Ref:.....

Our Ref:.....MV 625..... *Date:* July 16, 2001

Dr. Alison Mary Elliott,
C/o Uganda Virus Research Institute,
P.O. Box 49,
Entebbe.

Dear Dr. Elliott,

RE: **RESEARCH PROPOSAL "THE IMPACT OF HELMINTHS ON THE RESPONSE TO IMMUNIZATION AND ON SUSCEPTIBILITY TO INFECTIOUS DISEASES IN CHILDHOOD IN UGANDA"**

The Uganda National Council for Science and Technology (UNCST) has approved the above research proposal for implementation. The approval will expire on **July 12, 2002**. If it is necessary to continue with the research beyond the expiry date, a request for continuation should be made to the Executive Secretary, UNCST.

Any problems of a serious nature related to the execution of your research project should be brought to the attention of the UNCST, and any changes should be submitted for UNCST's approval before they are implemented.

This letter serves as proof of UNCST approval and as a reminder for you to submit to UNCST timely progress reports and a final report on completion of the study.

Yours sincerely,

Julius Ecuru
for: Executive Secretary
UGANDA NATIONAL COUNCIL FOR SCIENCE AND TECHNOLOGY

LOCATION/CORRESPONDENCE
PLOT 10, KAMPALA ROAD
UGANDA HOUSE, 11TH FLOOR
P.O. BOX 6884
KAMPALA, UGANDA.

COMMUNICATION
TEL: (256) 41-250499
FAX: (256) 41-234579
E-MAIL: uncst@starcom.co.ug
WEBSITE: http://www.uncst.go.ug

LONDON SCHOOL OF HYGIENE
& TROPICAL MEDICINE

ETHICS COMMITTEE

APPROVAL FORM
Application number: 790

Name of Principal Investigator	**Dr Alison Elliott**
Department	**Infectious and Tropical Diseases**
Head of Department	**Professor Peter Smith**

Title **The impact of helminths on the response to immunisation and on susceptibility to infectious diseases in childhood in Uganda.**

Approval of this study is granted by the Committee.

Chair .....

(Professor Andrew Haines, Dean)

Date 29/7/01

Any subsequent changes to the consent form must be re-submitted to the Committee.