EXPLORING THE INTEGRATION OF TRADITIONAL AND MOLECULAR EPIDEMIOLOGICAL METHODS FOR INFECTIOUS DISEASE OUTBREAKS

CORDELIA EMMA MAITLAND COLTART

THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

UCL

2018

DECLARATION OF AUTHORSHIP

I, Cordelia Coltart, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: _____ Date: _____

ABSTRACT

Background: Understanding the transmission dynamics of infectious pathogens is critical to developing effective public health strategies. Traditionally, time consuming epidemiological methods were used, often limited by incomplete or inaccurate datasets. Novel phylogenetic techniques can determine transmission events, but have rarely been used in real-time outbreak settings to inform interventions and limit the impact of outbreaks.

Methods: I undertook a series of novel studies to explore the utility of combining phylogenetics with traditional epidemiological analysis to enhance the understanding of transmission dynamics. I investigated HIV in an endemic South African setting and Ebola in an acute outbreak in Sierra Leone. The strengths and limitations of this combined approach are explored, ethical issues investigated and recommendations made regarding the implications of this work for public health.

Results: Phylogenetics provides an exciting and synergistic tool to epidemiological analysis in outbreak investigation and control. These combined methods enable a more detailed understanding than is possible through either discipline alone. My key findings include:

- Identification of infection source: Phylogenetics gives new insight into the role of external introductions (e.g. migrators) in driving and sustaining the high incidence of HIV.
- Earlier identification of new emerging clusters: I identified a new cluster of HIV from around a mining community. This is one of the first examples of molecular methods detecting a previously unknown outbreak.
- Identification of novel mechanisms of transmission: This work suggests that children may have been infected by playing in puddles contaminated with Ebola, a previously unrecognised route of transmission.

Conclusion: The integration of these two methods facilitate sophisticated real-time techniques to maximise understanding of transmission dynamics, allowing faster and more effectively targeted interventions. Moving forwards, sequence data should be incorporated into standard outbreak investigation. This is critical at a time when infectious disease outbreaks have led to the some of the most significant global health threats of the recent past.

Table of Contents

PUBLICATIONS & PRESENTATIONS ARISING FROM THIS RESEARCH

PUBLICATIONS

1. **Coltart CEM**, Lindsey B, Ghinai I, Johnson AM, Heymann DL. 2017. The Ebola outbreak, 2013-2016: old lessons for new epidemics. *Philosophical Transactions of the Royal Society B*; 372(1721). pii: 20160297. doi: 10.1098/rstb.2016.0297.
2. Carroll MW, Matthews DA, Hiscox JA… **Coltart CEM**.. *et al*. 2015. Temporal and spatial analysis of the 2014-2015 Ebola outbreak in West Africa. *Nature*; 6;524(7563):97-101.
3. **Coltart CEM**, Johnson AM, Whitty CJM. 2015. Role of healthcare workers in early epidemic spread of Ebola: policy implications of reactive versus prophylactic vaccination strategy in outbreak prevention and control. *BMC Medicine*; 19;13:271. doi: 10.1186/s12916-015-0477-2.
4. **Coltart CEM**, Johnson AM, Heymann DL. 2017. From contamination to containment. 2017. *Natural History Magazine*; 125(8): 40-43.
5. **Coltart CEM**[*], Hoppe A[*], Parker M, *et al*. On behalf of the Ethics in Phylogenetics Working Group. Ethical considerations in HIV Phylogenetic Research. 2017. *Submitted.*

SPECIAL ISSUE CO-EDITOR

Atkins KE, Edmunds WJ, **Coltart CEM**. 2017. The 2013-2016 West African Ebola epidemic: data, decision-making and disease control. *Philosophical Transactions of the Royal Society B*;372(1721).

MAJOR PRESENTATIONS

1. Gates Foundation Meeting: Invited speaker and session chair for Measuring and Modelling Community Engagement in the Ebola outbreak. 2017. Washington DC, USA.
2. Gates Foundation/Wellcome Trust Meeting: Invited speaker and panel discussant for PANGEA - Ethics of Phylogenetics Meeting. Co-organiser of meeting. Presentation:

Lessons from phylogenetic studies in other infectious disease outbreaks. 2017. London, UK.

3. CDC meeting: Invited speaker at the West Africa Ebola Virus Disease Outbreak Surveillance Data Analysis Workshop. Presentation: The National Ebola Data Archive Project. 2015. Sierra Leone.

4. Royal Society of Tropical Medicine annual Research in Progress meeting: Winner of research in progress prize for oral presentation. Presentation: Ebola – using phylogenetics to modernize outbreak investigation. 2017. London, UK.

5. Royal College of Physicians, Quincentennial meeting: Winner of Quincentennial prize for trainee doctors. Presentation: Modernising old lessons for new epidemics. 2018. London, UK.

PUBLICATIONS AND PRESENTATIONS IN PROGRESS:

1. **Coltart CEM**, Hué S, Shahmanesh M, Tanser F, Seeley J, Gareta D, Ntuli S, Bärnighausen T, Sartorius B, de Oliveira T, Zuma T, Chimbindi N, Pillay D, Johnson AM. Ongoing HIV microepidemics in rural South Africa: the need for flexible interventions.
   Accepted for a late breaker presentation at CROI 2018. Manuscript in preparation.

2. **Coltart CEM**, Hué S, Shahmanesh M, Tanser F, Seeley J, Gareta D, Ntuli S, Bärnighausen T, Sartorius B, de Oliveira T, Zuma T, Chimbindi N, Pillay D, Johnson AM. Using molecular data to identify the transmission dynamics of HIV-1 in KwaZulu-Natal: a combined phylogenetic and empirical epidemiological analysis.
   Abstract submitted to IAS for oral presentation.

3. **Coltart CEM**, Cooper D, Hué S, Lloyd S, Cortes MC, Town K, Levine A, Cotton M, Goodfellow I, Pillay D, Johnson AM. Anatomy of a community outbreak of Ebola: The use of sequence data to enhance classical epidemiology in outbreak investigation
   Manuscript in preparation. Abstract submitted to ASTMH for oral presentation.

4. **Coltart CEM**, Zhukova A, Tostevin A, Stirrup O, Harling G, King C, Pillay D, Johnson AM, Gascuel O, Hué S. Determining patterns of migration and HIV incidence in rural KwaZulu-Natal: a novel bioinformatics approach combining molecular and epidemiological approaches.
   Manuscript in preparation.

ACKNOWLEDGEMENTS

I would like to thank the many people who have contributed to me writing this thesis and to my career. It is difficult in a few words to convey how much I appreciate their help and support.

I thank my supervisors for their support and guidance during the entire process. **Anne Johnson** has given me far more than anyone could ever expect from a supervisor. Her ability to see both the big picture and fine detail across multiple topics always impresses me. She has instilled in me a passion for epidemiology, which will continue for my entire career. She is a role model and I feel incredibly privileged to both work with her and be her last PhD student. **Deenan Pillay's** willingness to share his vast experience of both HIV and virology more generally, and his guidance on much of my work, are greatly appreciated. **Stèphane Huè** has taught me everything I know about phylogenetics. He has a remarkable ability to explain complex problems when situations seemed far too complex for me to grasp! **Chris Whitty** has been a mentor and a source of advice and encouragement, since 2007.

I thank the Wellcome Trust for sponsoring my work and the UCL Clinical PhD programme for believing in me. I have learned a huge amount and I am enormously grateful for the opportunity. I thank Tamyo Mbisa for his constructive feedback on my work, as part of my upgrade committee, and Sangita Patel and Sandy Gale for all their practical support.

I would like to thank my colleagues and friends at UCL and LSHTM, who have provided guidance, problem solving and laughter at various points during the PhD. In particular, Anna Tostevin, Guy Harling, Carina King, Kholoud Porter, Nigel Field, Pam Sonnenberg, Oliver Stirrup, Andrew Copas, Katy Town, Katie Atkins, Miles Carroll, Anna Aryee and Neelu Kumar. I would like to express my sincere appreciation to all the participants, Africa Centre team, and colleagues in Sierra Leone.

Finally, special thanks are due to my fiancè, friends and family. To Dominic, who has supported and encouraged me throughout this journey. He has been amazing and has helped me enormously. Both Dominic and my mother read each and every page of this thesis - I honestly can't thank you both enough. To my parents for giving me every opportunity in life and for their unending support. To my mother especially – who never completed her PhD, in part, because I was born - you are an inspiration to me. And my grandparents, who at 96 and 97 years old, continue to give me encouragement. And finally, my siblings for providing a sense of humour when it was needed the most! My heartfelt thanks and gratitude go to you all - THANK YOU.

LIST OF TABLES

LIST OF FIGURES

LIST OF BOXES

ABBREVIATIONS

95% CI	95% Confidence Interval
AC	Africa Centre for Population Studies (Africa Centre)
AHRI	Africa Health Research Institute (previously known as AC)
AIDS	Acquired Immunodeficiency Syndrome
CDC	Centers for Disease Control
CMO	Chief Medical Officer
DBS	Dried Blood Spot
DNA / RNA	Deoxyribonucleic Acid / Ribonucleic Acid
DOH	Department of Health
DRM	Drug Resistant Mutation
DSA	Demographic Surveillance Area
DSS	Demographic Surveillance Survey
EBOV	*Ebolavirus*
EVD; 'Ebola'	Ebola Virus Disease
GD	Genetic Distance
HCW	Healthcare Worker
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio
IHR	International Health Regulations
MOH	Ministry of Health
NGS	Next Generation Sequencing
NNS/S	Number of Nucleotide Substitutions per Site
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PrEP	Pre-exposure prophylaxis
Ro / Re	Reproductive Number / Effective Reproductive Number
SL	Sierra Leone
tMRCA	Time from Most Recent Common Ancestor
VL	Viral Load
WHO	World Health Organization
ZA	South Africa

CHAPTER 1

INTRODUCTION

1.1 OUTLINE OF THIS THESIS

The objective of this thesis is to determine the utility of combining molecular and classical epidemiological analysis in order both to understand the transmission dynamics of infectious disease outbreaks and to better inform prevention strategies. To achieve this objective, I investigated two contrasting infectious diseases: Human Immunodeficiency Virus (HIV-1) as an example of a chronic generalised infection; and Ebola Virus Disease (EVD) as an example of an acute localised infection. These have given rise to two of the most serious epidemics in the last fifty years. I studied them at their geographical epicentres: South Africa and Sierra Leone respectively. In this thesis I use these examples to draw general conclusions about the value of combining molecular and classical epidemiological approaches in the study of infectious disease outbreaks.

1.1.1 SUMMARY OF CHAPTERS

My first chapter summarises this thesis. In **Chapter 2**, I present the background to my research and highlight the reasons that this is an important area of scientific development. I begin by putting my work into context with a summary and critique of two areas: the relevant scientific literature concerning the use of molecular epidemiology in infectious disease outbreaks, and the methods that can be used to optimise the utility of combining molecular and traditional epidemiological techniques.

Chapter 3 describes the datasets I used for my analysis of HIV and Ebola, how the data were collected, the ethical consents obtained, and the methods used to link the datasets.

Chapter 4 provides a descriptive epidemiological review of the 2013-2016 Ebola outbreak in West Africa. This chapter provides an overview of the current methods used in standard outbreak investigations.

Chapter 5 explores the use of open access data to construct epidemiological transmission trees for Ebola cases from early outbreak settings for both the 2013-2016 outbreak, as well as for historical outbreaks. I use the transmission trees to undertake descriptive

hypothesis generation and hypothetical testing of different vaccination strategies for healthcare workers. Following this, I evaluate whether a long-duration vaccine for prophylactic use is more effective in limiting transmission in early epidemics as compared to reactive vaccination. This chapter demonstrates that basic epidemiological work from publicly available data can be valuable in informing public health policy and intervention strategies. At the time of undertaking this analysis I did not have access to any related empirical epidemiological data or sequence data. However, had sequence data been available for these outbreaks, it would very likely have strengthened this work.

Chapter 6 describes an isolated rural outbreak of Ebola Virus Disease (EVD) in Sierra Leone by combining clinical, epidemiological, and anthropological field data with sequence data to gain a detailed understanding of the outbreak and its transmission dynamics. I use 554 *Ebolavirus* sequences from Sierra Leone to infer a phylogenetic tree and contextualise the sequences from this outbreak within the wider epidemic. This multidisciplinary approach builds on my analysis in Chapter 5 which was limited to classical epidemiological techniques. The analysis in Chapter 6 introduced me to the techniques necessary to link epidemiological and molecular datasets. This work was undertaken with a small dataset; later chapters progress this work in applying the techniques to a larger HIV dataset.

Chapter 7 investigates the trends in HIV transmission dynamics within the Africa Health Research Institute's (Africa Centre) population surveillance site in rural Kwa-Zulu Natal, South Africa. I integrate phylogenetic analyses (n=2,179 sequences) and epidemiological data to investigate the genetic diversity of the virus and to describe the characteristics of identified clusters. A novel large cluster was identified and this chapter includes a full analysis of the evolution and characteristics of this cluster. I use more sophisticated phylogenetic techniques combined with larger epidemiological datasets to exploit more fully the potential of integrating these two disciplines. I consider the implications of this work to allow earlier detection of outbreaks and implementation of interventions.

Chapter 8 considers the association between migration and HIV acquisition. I use traditional epidemiological analysis, combined with a complementary novel bioinformatics model incorporating phylogenetics, to undertake this study. I hypothesise that the molecular approach will provide additional insights. I explore transmission events in migrators, correlates of transmission in migrators, patterns of mobility flows (both within and out of the study area), and rates of HIV transmission between different migratory

groups within the Africa Centre population. I investigate whether migration patterns can predict HIV acquisition and explore the limitations of this work.

Chapter 9 explores the ethical, legal and social issues which are raised by the increasing use of phylogenetic techniques. These issues are unique to phylogenetic studies and, therefore, have not been addressed in the existing bioethics literature. For example, phylogenetics leads to the potential to infer information about people within a linked network, potentially leading to social harms and stigmatization. Consequently, there is a need to develop an effective and sustainable model of good ethical practice in phylogenetic research. This chapter explores the novel issues arising from the use of phylogenetic techniques, and sets out recommendations to minimise the associated risks to participants while optimising the scientific benefits.

Chapter 10: I summarise the conclusions that can be drawn from my work and discuss the benefits and limitations associated with combining traditional and molecular epidemiology in analysing infectious disease outbreaks. In addition, I review how a combined epidemiological and molecular approach to infectious disease outbreak responses can be used, both to inform intervention strategies and to influence public health policy.

1.2 Context

Understanding the transmission dynamics of infectious agents is critical in developing effective public health interventions. Transmission is driven by pathogen-specific biology, by host susceptibility and biological variability, and by the socio-economic, demographic and behavioural characteristics of the population exposed to the pathogen. Traditional epidemiological approaches track the evolution of epidemics through 'time, place and person' based either on clinically- or microbiologically-confirmed observation. During outbreak investigation, transmission events are investigated using contact tracing and empirical epidemiological methods. However, there are significant limitations to this approach, including a variable ability to yield an accurate representation of the underlying transmission tree. Furthermore, collecting high quality epidemiological data is time consuming, costly, and logistically complex in an outbreak setting. Therefore, it is important to consider whether alternative methods, such as molecular studies, can address some of these limitations.

New molecular tools, such as phylogenetic analysis, have been increasingly used in research settings to understand pathogen-specific biology and variability. While these tools have the power to elucidate transmission dynamics, they have limitations including that they require diagnostic material as well as advanced computational capabilities. The former may be opportunistically available, having already been collected as part of either routine clinical diagnostic testing or an epidemiological study, limiting the cost and logistical complexity of sample collection. However, the sample processing and complex analysis often require expensive technology and high computational capacity. In addition, without linkage to epidemiological data to contextualise the findings, molecular tools have limited ability to provide insight into the fine-scale transmission dynamics of epidemics.

When the proposal for this thesis was conceived in 2014, to my knowledge, phylogenetic analysis combined with traditional epidemiological approaches had not been used in the context of generalised epidemics, at a population-level or in standard outbreak investigations. However, there were a few examples of combined use with small datasets from concentrated epidemics in high-resource settings. It is likely that the high cost of sequencing and relatively new development of the necessary complex bioinformatics meant that, at that time, these tools had not been widely adopted in traditional epidemiological studies. However, the examples mentioned above suggested the potential of combined analysis to answer critical questions that are not otherwise easily addressed, such as the characterisation of 'transmitters' and the potential to address the directionality of transmission[4,5]. Thus, the combination of molecular techniques with traditional epidemiology in the context of generalised epidemics appeared to me to offer the possibility of an enhanced understanding of transmission dynamics, which would be of significant benefit in developing surveillance methodologies and intervention strategies. This realisation led to the proposal for my doctoral work.

This is a rapidly advancing field of research and there has been much progress since the start of my work. As well as improvements in the technology and cost of molecular analysis, a catalyst for the recent advances was the West African Ebola epidemic, which was the first acute infectious disease outbreak in which near real-time sequencing technology was available. These parallel developments and publications are outlined in Chapters 2, 4 (HIV and Ebola background respectively) and 10 (Discussion).

It is now widely accepted that an interdisciplinary approach is required to optimise

methodological and technical advances[6]. I explore how best to achieve this by investigating how to integrate epidemiological and molecular methodologies in order both to enhance epidemic disease surveillance, and to allow optimisation of intervention strategies. I hope that this will help to enhance the surveillance of and response to epidemics, and that this might lead to improved global outcomes. The need for such an improvement was highlighted at the 2016 International AIDS Society conference (Durban, 2016)[7] and was a common recommendation from the four Ebola assessment panels convened in response to the 2013-2016 West African outbreak[8-11].

1.3 Aims and Objectives

My aim was to conduct research on two main themes:

1. To understand the benefit of combining traditional and molecular epidemiological analyses to enhance understanding of transmission dynamics in infectious disease outbreaks; and
2. To explore an innovative epidemiological framework to guide the integration of molecular methods into traditional epidemiological studies of epidemics. I considered how these combined approaches can best be optimised to advance real-time surveillance programmes and inform intervention strategies.

The specific objectives of this thesis are to:

1. Define transmission chains by constructing phylogenetic trees;
2. Create novel and up-to-date datasets combining clinical, epidemiological, behavioural, and phylogenetic data to explore transmission dynamics in the context of a localised Ebola epidemic and a generalized HIV epidemic;
3. Explore transmission events in-depth to define the clinical and epidemiological determinants of transmitters, high risk groups, and outcome measures. How can this inform targeted intervention strategies?;
4. Evaluate the strengths and limitations of using traditional and molecular epidemiological approaches alone in understanding transmission dynamics, and the value added by combining these two approaches;
5. Establish how the application of these methods varies between pathogens using Ebola as a contrasting example to HIV (i.e. acute localised vs. chronic generalised epidemics);
6. Assess the key ethical issues associated with undertaking phylogenetic analysis and suggest ways to minimise these risks.

CHAPTER 2

BACKGROUND

2.1 MOLECULAR EPIDEMIOLOGY – LITERATURE REVIEW

2.1.1 INTRODUCTION TO MOLECULAR EPIDEMIOLOGY

A) DEFINITION

The term 'molecular epidemiology' was first used in 1973 in relation to work on influenza virus strains[12]. Subsequently, numerous definitions have been developed encompassing work on both infectious and non-infectious diseases[1,13-15], but for the purposes of this thesis I will use the common following definition of infectious disease molecular epidemiology: *the use of molecular typing methods for infectious agents in the study of the distribution, dynamics and determinants of health and disease in human populations. It combines the disciplines of molecular biology, epidemiology and biostatistics*[1,14,16*].

While the focus of this work is on phylogenetic analyses, it is important to remember that the application of molecular epidemiology is much broader. It includes other viral strain typing methods, host genetics, host/population immunology and vector biology, all of which can also influence transmission dynamics of an outbreak.

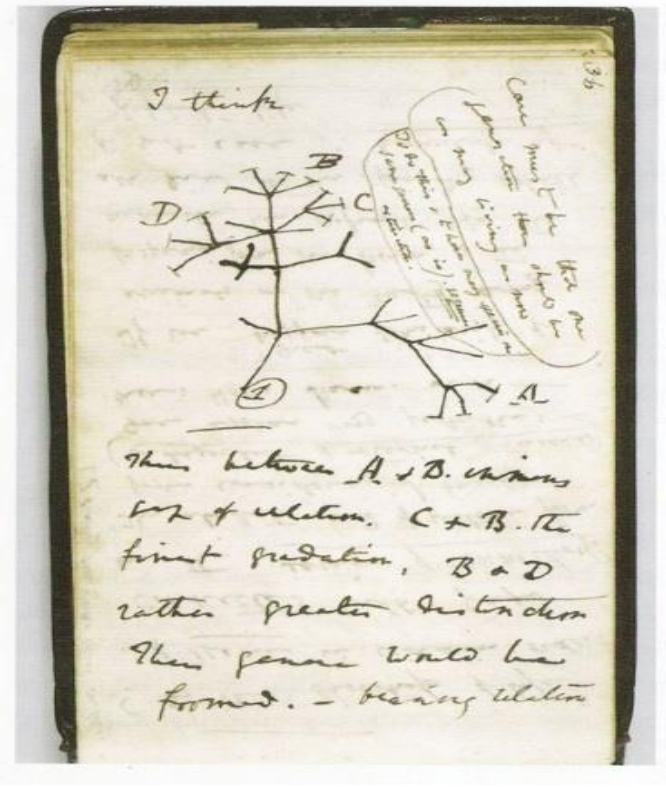
B) HISTORY OF MOLECULAR EPIDEMIOLOGY

Charles Darwin is widely considered to be the founding father of evolutionary theory, which in turn forms the basis of many of the underlying principles upon which molecular epidemiological analysis is based. The only figure to appear in '*On the Origin of Species',* published in 1859, is a branching evolutionary tree – the earliest form of a phylogenetic tree (Figure 2.1). In this seminal work, Darwin details the scientific theory that populations evolve over the course of generations through a process of natural selection. He speculates that diversity of life arose from a common descendant through a branching pattern of evolution[17].

[*] In the context of this thesis traditional epidemiology refers to empirical 'muddy boots' field epidemiology for the Ebola work and classical demographic surveillance data for the HIV work. Molecular epidemiology refers exclusively to viral sequence data and resulting phylogenetic tree analysis.

Molecular epidemiology itself did not develop until the second half of the twentieth century. One of the first formal reports of its use was in virus strain characterisation of polypeptide polymorphisms in the mid-1970s[18]. By the 1980s, the use of molecular epidemiology using restriction endonuclease analysis to study viral and microbial pathogens was widespread[15,19]. Since then, many technological advances have been made with almost continuous evolution of typing methods and sequence advancements.

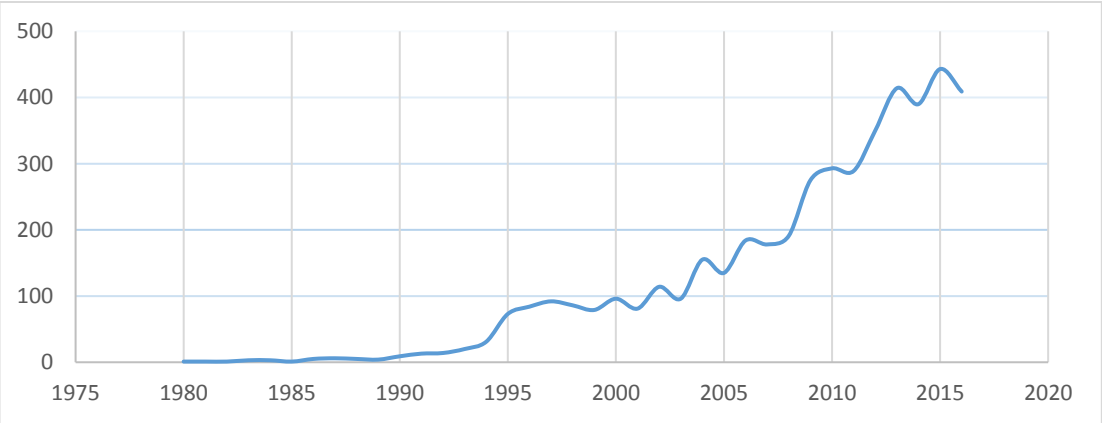
Figure 2.1: The first known evolutionary tree (Charles Darwin, 1859, On the Origin of Species)[17]



With these rapid technological advances over recent decades, the field of molecular epidemiology has grown significantly. Field and colleagues[1] highlighted this rapid increase by plotting the number of published articles about molecular epidemiology in infectious diseases from 1975-2010. An updated plot of articles published from 1975-2016 is shown in Figure 2.2. However, Field and co-workers noted a stark variation in the quality of the papers and highlighted that many do not consistently include true epidemiological applications, merely some form of molecular technique. As such, these authors produced a statement to provide a framework to standardise work and to support transparent research reporting entitled '*Strengthening the Reporting of Molecular Epidemiology for*

Infectious Diseases' (STROME-ID)[1]. The key recommendations from this work are outlined in Box 2.1.

Figure 2.2: Number of papers published per year identified with the search terms "infection" AND "molecular epidemiology" in PubMed database (n=4,974)



Box 2.1: Key conclusions from 'Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases' (STROME-ID) – Field *et al*. (2014)[1]

- When reporting molecular epidemiological results, consider:
  - Stating the numbers of participants (and samples) at each stage of the molecular sample processing;
  - Using standardised nomenclature to report information by strain type (where appropriate);
  - The need to stratify statistical models by pathogen strain;
  - Consider using dendrograms or phylogenetic trees to depict the molecular relatedness of strains;
  - For any putative transmission chain, all alternative explanations for the findings need to be investigated and the consistency between molecular and epidemiological evidence reported. Field *et al*. cite an example of a putative transmission between person A and person B in which several reasons could account for the transmission cluster observed: direct transmission (A to B or B to A), intervening cases (A to C to B), or a common source (C to A and C to B).

- Particular emphasis is given to the interpretation of molecular findings and description of limitations.
  - For example, litigation cases show that molecular data are of little value without linked epidemiological data to provide essential supportive evidence for any conclusions. Such cases show it is easier to disprove than prove transmission beyond doubt.
  - Field *et al*. highlight that authors should consider specific factors that might contribute to missing data and/or misinterpretation e.g. latent infection or disease reactivation leading to clustering with no immediately obvious epidemiological links in organisms such as *Mycobacterium tuberculosis*.

However, as this research is focused on phylogenetic analyses of viral infections, the discussion will examine techniques relevant to this form of analysis. Currently, four broad types of genetic analysis are used[1]:

1. Genetic amplification (via polymerase chain reaction (PCR)) followed by direct comparison of DNA or RNA sequence data (including whole–genome sequencing).
2. DNA fragmentation (by restriction endonuclease enzymes cutting DNA at specific recognition sequences) followed by pathogen-specific PCR amplification and the use of gel electrophoresis to compare band patterns.
3. The use of fluorescently labelled nucleotides to measure the accumulation of PCR products.
4. The use of hybridisation (short nucleic-acid recognition sequences are attached to a matrix) to compare gene expression or different genotypes of pathogen.

Viral phylogenetic analysis requires full length or partial genome sequences. Therefore, this thesis focuses on the first method listed above, that of PCR amplification of direct virus DNA/RNA.

The development of PCR-based methods has revolutionised biological science by greatly accelerating the rate at which DNA sequence data can be generated, making it routinely available[20]. PCR-based methods are technically simple, requiring only basic laboratory skills, as well as being rapid, sensitive, and specific. Furthermore, these methods are less expensive than the alternatives and the decreased cost has made these techniques more accessible and allowed rapid technological development.

Sequence data generated in this way can be used for diagnostics by specifically amplifying DNA or RNA from the pathogen of interest. Bioinformatic techniques then allow the identification of viral 'strains' or the inference of phylogenetic relationships to identify routes of disease transmission. Applying epidemiological techniques to these sequence data allows evolutionary relationships to be inferred. One of the most notable examples is the study by Ou and colleagues[21], which used comparative genetic analysis methods, including genetic distance measurements and phylogenetic tree analysis, to demonstrate the transmission of HIV-1 from an infected dentist to five patients (described below - Table 2.1).

Sequencing technology has evolved rapidly over recent decades and current methods include: direct Sanger sequencing; single genome amplifications (SGA) or cloning; and next generation sequencing (NGS). The majority of historical sequence libraries, including all sequences used in this thesis, and publicy available sequences submitted to Los Alamos and GeneBank databases, have been generated using Sanger sequencing of a single gene region. Sanger sequencing results in a single consensus sequence, which describes the most frequent base at each position of the gene. This is what is needed for classical phylogenetic analysis. It is also particularly useful for clinical resistance testing and subtype screening.

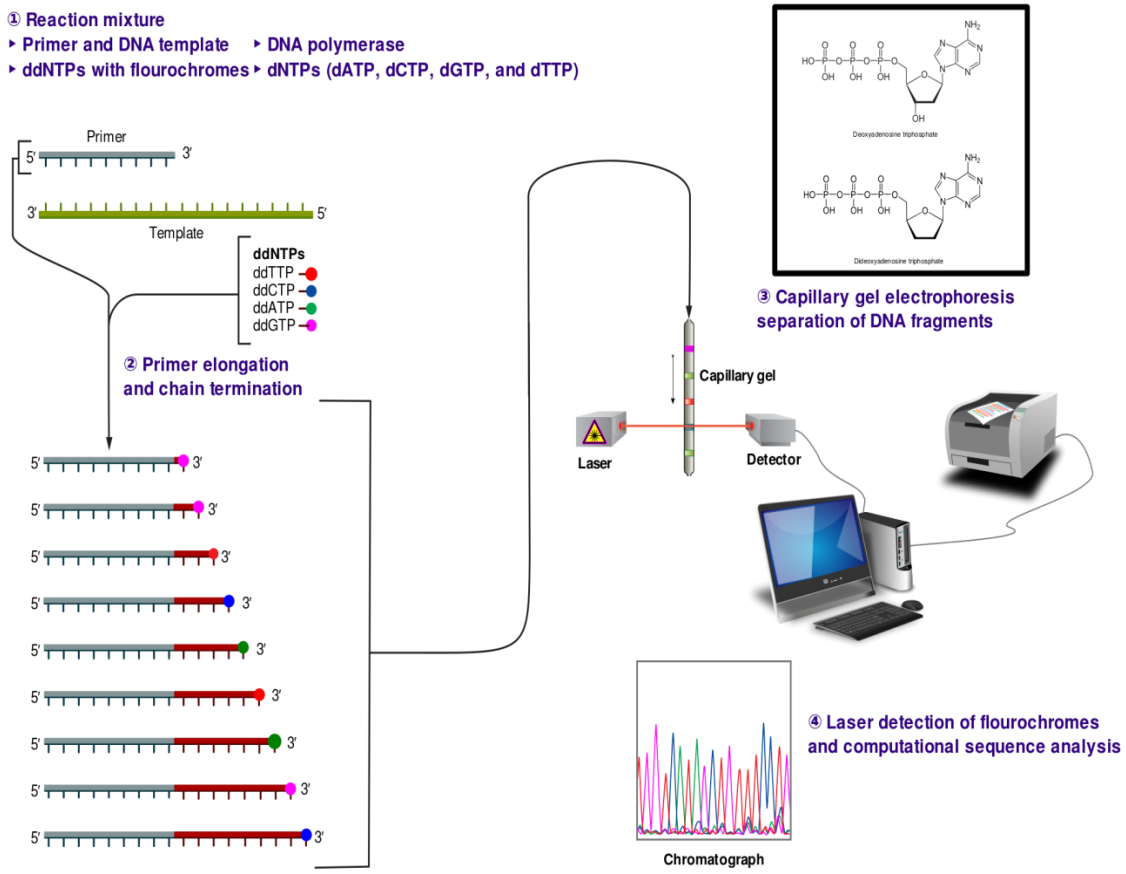
In recent years, NGS has transformed the field and is likely to become the sequencing methodology of choice moving forwards. It allows both reconstruction of full length genomes from multiple small length genetic reads, and depth of sequencing to be undertaken – i.e. multiple intra-host species to be sampled – rather than one single consensus sequence to determine quasi-species. This method is discussed further in Chapter 10, in which future developments are highlighted. However, NGS could potentially benefit phylogenetic studies in two ways: by increasing the power of the phylogenetic signal; and by allowing better ascertainment of transmission direction, on the basis that within-host viral diversity increases over time so that time of infection can be more accurately inferred. Furthermore, NGS has advantages over Sanger sequencing including speed (high throughput), decreased cost, and a smaller amount of DNA template required. Although these full length and deep sequences may be increasingly available, their utility remains limited due to the developmental costs and computing power required to produce and analyse these sequences (bioinformatics). However, given their likely future widespread use, they are worthy of inclusion here.

This thesis, however, focuses on phylogenetics in order to determine transmission dynamics. Therefore, a consensus sequence is required and the sequences used are all generated via Sanger sequencing. I focus on this method of sequencing alone henceforth. Figure 2.3 shows the steps involved from DNA template to PCR to sequence generation via Sanger sequencing methodologies.

Recent advances in genome sequencing technology mean that whole genome sequencing of a virus can be carried out in the field in less than 24 hours by means of portable

sequencing systems[22]. These systems can be as small as the size of a can of Coca-Cola and the results can be downloaded to a laptop or other mobile device. Previously, real-time/near real-time sequencing during an epidemic was not possible, as it required samples to be sent to a laboratory. This was much slower, particularly given transport delays in some locations. The 2013-2016 West African Ebola epidemic was the first time that near real-time sequencing was used in an acute emerging infection outbreak in the field. Although this method still faces many technical challenges, it provides a crucial new tool to add to epidemiological research and offers the potential both to enhance the ability to trace rapidly how a disease is transmitted, and to target interventions and resources more effectively.

Figure 2.3: The Sanger method of DNA sequencing



(1) A primer is annealed to a sequence, (2) Reagents are added: DNA polymerase, deoxynucleotide solution mix (dNTPs – dATP, dCTP, dGTP, dTTP), and a small amount of all four dideoxynucleotides (ddNTPs) labelled with fluorophores. During primer elongation, the random insertion of a ddNTP instead of a dNTP terminates synthesis of the chain because DNA polymerase cannot react with the missing hydroxyl. This produces all possible lengths of chains. (3) The products are separated on a single lane capillary gel. (4) A Laser detection system detects the labelled ddNTP and allows determination of which nucleotide predominates at each position to form a consensus sequence, by computational sequence analysis. Estevezj (2012)[23] (CC BY-SA 3.0).

However, the full potential of these advances can only be realised if sequence data can be accessed promptly for analysis. The recent outbreaks of Ebola and Zika have seen many scientists advocating early data release. This is in contrast to conventional scientific research and publishing, based on peer review, which can be a cumbersome process and makes the dissemination of data during a public-health emergency too slow to have an impact on policy. Recent moves have encouraged immediate release of data to public databases with subsequent publication of peer-reviewed analysis dependent upon the data being publicy available. Indeed, during the recent Zika crisis the World Health Organization (WHO) and international partners renewed efforts to promote rapid sharing of the latest research data to facilitate faster decision making[24].

c) Uses of molecular epidemiology

Although there are many applications of molecular epidemiology, three primary applications predominate:

1. Identification of the aetiological agents of diseases – to investigate diversity of pathogens, diagnose disease states, or for infectious disease surveillance.
2. The study of the patterns of disease transmission and linked infections – to investigate geographical distribution over time, place and person; to make predictions for further epidemic developments; and to plan intervention and prevention strategies.
3. The study of evolutionary relationships and how epidemics are evolving.

Examples of applications of molecular epidemiology to infectious diseases are outlined in Table 2.1. Other applications include forensics, identification of new species, vaccine design and monitoring for escape mutants, and improved understanding of why infections recur (reinfection vs. recurrence).

Below, I outline two examples of molecular epidemiology in infectious disease outbreak research. The first shows how this tool has been helpful in combatting an infectious outbreak (MRSA), while, in contrast, the second describes an example of molecular epidemiology being used in a manner that led to misleading conclusions being drawn, primarily due to erroneous interpretation. I set out these contrasting examples to show that it is essential to understand the limitations of this tool in order to ensure that it is used to complement and enhance our current understanding.

1. Harris and coworkers[25] described an example in which molecular methods were extremely useful in determining the source of an ongoing methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak in a Special Care Baby Unit ("SCBU") in Cambridge, UK. This outbreak persisted over six months, despite deep cleaning and extensive infection control measures. The molecular work undertaken in this study validated and expanded the findings from conventional epidemiological analysis. Conventional epidemiological analysis, including swabbing and antibiogram profiling as a proxy for strain identification, revealed 12 infants colonised with MRSA during a 6 month period. These infants were suspected of having linked infections, but a persistent outbreak could not be confirmed by conventional methods owing to time gaps between cases. The conventional epidemiological analysis led to a dead end.

Table 2.1: Example applications of molecular epidemiology

Application	Example
HIV transmission	1. **Aunt to baby HIV transmission**[26] A 76 day old baby presented unwell and subsequently tested positive for HIV, although the mother was HIV negative. The baby's aunt had breast feed the infant and was known to be HIV-positive, as was her child. Samples from the baby, aunt and cousin were used to construct a phylogenetic tree, together with 100 'local' HIV samples. These three sequences closely clustered and supported the hypothesis of surrogate transmission between the aunt and baby. 2. **Healthcare worker to patient transmission**[21] A patient, who had no known risk factors for HIV, was diagnosed with HIV. The patient had a history of an invasive procedures performed by a dentist who was known to have HIV. Healthcare worker to patient HIV transmission had not been reported prior to this case. Subsequently, six other patients were found to have been infected with HIV. Phylogenetic analysis complemented the epidemiological investigation to show that the dentist and five patients had closely related sequences, indicating that these patients became infected while receiving care from the dentist.
HIV and hepatitis transmission	**Outbreak in Libyan hospital**[27-29] In 1998, outbreaks of both HIV-1 and hepatitis C virus (HCV) infections were reported in children attending a hospital in Libya. Foreign medical staff were accused of infecting the children, and were then arrested and imprisoned. However, their legal case used phylogenetic analyses as evidence, which concluded that the viruses were already circulating in the hospital before the arrival of the foreign medical staff and therefore, they could not have been the ones to introduce the virus. The phylogenetic clusters of HIV-1 and HCV were likely to be from sub-Saharan Africa, which would be compatible with the large number of migrants in Libya. The results support nosocomial transmission and a long-standing infection control problem in the hospital.
Cholera in Haiti	**Origin of Cholera Epidemic in Haiti**[30] Although cholera has been present in Latin America since 1991, it had not been epidemic in Haiti for at least 100 years. However, shortly after the 2010 earthquake there was a severe outbreak of cholera in Haiti. Phylogenetic analysis indicated that there was a close relationship between the Haitian isolates and variant *V. cholerae* El Tor O1 strains isolated in Bangladesh in 2002 and 2008. Therefore, the Haitian epidemic was probably the result of the introduction, through human activity, of a *V. cholerae* strain from a distant geographic source, potentially from international aid workers following the earthquake.

Pursuant to this, molecular techniques were used and the studies sequenced isolates from all colonised patients in the SCBU, as well as MRSA isolates from other hospital patients, and community samples with the same antibiotic susceptibility profiles as controls. This identified 26 related cases and showed transmission occurred within the SCBU between mothers on a postnatal ward, and also in the community. The data were used to hypothesise, and subsequently to confirm, that MRSA carriage was by a staff member at the SCBU. This led to multiple introductions and caused a persistent outbreak, despite the implementation of infection control measures. Although conventional epidemiological methods failed to determine the cause of the outbreak in this example, it would be wrong to conclude that molecular methods are sufficient on their own. This study highlights that genome sequencing with linked epidemiological data holds great promise for rapid, accurate and comprehensive identification of outbreak transmission pathways, with concomitant reductions in infections, morbidity and costs[25].

2. During the 2009 Influenza pandemic in Mexico, while sequencing was used to confirm the rapid global spread of the novel H1N1 strain, molecular identification of the novel strain without adequate population based studies led to the initial erroneous belief that this was a more pathogenic strain than other seasonal influenza strains. However, this was subsequently proven to be untrue as the H1N1 strain was of similar pathogenicity to other strains. The error arose as initially only severely ill patients were sampled. This sampling bias misrepresented the population and led to misleading results[31]. Therefore, although molecular techniques hold much promise to enhance understanding, caution is required in interpreting the results, particularly as infectious disease outbreak investigations rarely have a complete/population-wide sampling frame.

2.1.2 Methodological considerations in using molecular epidemiology

a) Phylogenetic principles

1. Evolutionary assumptions and models

The principle underlying phylogenetic inference is to analyse the similarities and differences between biological organisms to infer the evolutionary history of those entities. Over time and successive generations, changes occur in the genetic code of organisms. Phylogenetic inference uses these changes to determine the genetic similarity between two organisms, assuming that the more similar the genetic sequence, the closer

in time the sequences are to having a common ancestor. For example, in the simple case of evolution in an asexual unicellular organism, each cell divides producing two cells with identical genomes except for a few genomic sites where a mutation has occurred. After continuing for 100 generations, yielding 2^{100} cells, it is possible to sample cells from the final population and theoretically infer the lineages leading up to those cells over the course of the 100 generations. If ten cells from the final population are sampled and compared to one another, two cells that share a recent common ancestor (i.e. diverged recently) will have accumulated fewer differences than a pair that diverged earlier and whose last common ancestor was many generations ago. This can be calculated by counting the number of mutations accumulated (mutations per site) – the genetic distance between two organisms.

Early work in this area hypothesised that sequence divergence accumulates at a roughly constant rate over time – this is called the 'molecular clock' hypothesis. However, this assumption is no longer accepted. The rate at which substitutions accumulate is dependent upon many factors including the underlying mutation rate, the metabolic rates of species, generation times, population sizes and selective pressures. Therefore, some substitutions occur more frequently than others – for example transitions (substitutions between purines, or between pyrimidines) are more frequent than transversions (substitutions between purines and pyrimidines)[32] – meaning that the rate at which substitutions accumulate is often not linear[33]. Furthermore, multiple substitutions may occur at the same nucleotide position, making it difficult to ascertain the true evolutionary distance[34]. Therefore, in practice, taking the end points of evolution and inferring the ancestral history is not as straightforward as might be hoped, or as the molecular clock hypothesis would suggest. There are three main challenges in inferring evolutionary phylogenetic relationships:

1. Accumulation of differences does not occur uniformly (e.g. due to random chance, differing mutation rates and selective pressures). Therefore, if a branch containing many mutations was sampled, one might mistakenly infer that samples were more distantly related from each other than they really were. In consequence, phylogenetic reconstruction methods need to account for the variation in rates of divergence between lineages.

2. In contrast, homoplasy is the occurrence of convergent and parallel evolution, leading to lineages becoming more similar to each other over time, but not due to common ancestry. If the traits are due to convergence or parallel evolution, they should not be used in analysis. For example, in HIV phylogenetics, convergent evolution occurs as a result of anti-retroviral treatment (ART) leading to selective pressures inducing drug resistant mutations. Therefore, patients on ART are more likely to have similar genotypes due to acquired mutations, rather than transmitted mutations from a common ancestor. It is common practice to remove these mutation sites from all sequences before phylogenetic analysis occurs to eliminate false results.
3. Evolution is not strictly vertical, so the notion of evolution as a branching process represented as a tree is an oversimplification that can lead to misinterpretation. For example, recombination (within sexually reproducing species) and hybridization (between species) can lead to mixing and genetic variation that is not exclusively vertical and this can be further complicated by DNA duplications, deletions etc. Thus, choosing which entities and/or which part of the genome to study is of particular importance in phylogenetic analysis.

To overcome these challenges, phylogenetic analysis incorporates an evolutionary model with statistical tests to evaluate how the evolutionary rates deviate from a uniform rate, and to convert the genetic distances that exist between sequences into measures of estimated evolutionary distance. Numerous models exist and each varies in its complexity, usually defined by three parameters: nucleotide base frequency (of the four nucleotides: A,C,G,T), base exchangeability (relative tendency of bases to be substituted for one another), and the rate of heterogeneity between sites (typically described by a gamma distribution)[34]. The Jukes Cantor model is the simplest model, including just one parameter: base substitution rate. The most complex model is the General Time Reversible Model (GTR), which includes all three parameters. Table 2.2 sets out a description of the most common models.

Table 2.2 Summary of main phylogenetic evolutionary models

Evolutionary model	Description
Jukes Cantor (JC)[35]	Assumes an equal frequency of the four bases and that all substitutions are equally likely.
Kimura's two parameter model (K2P)[32]	Assumes base frequency is equal, but that transitions occur more frequently than transversions*.
Felsenstein (F81)[36]	This model allows an unequal frequency of base substitutions.
Hasegawa, Kishino and Yano (HKY85)[37]	Allows both the base frequency and transition/transversion rate to differ.
General Time Reversible Model (GTR)[38]	Allows all substitutions to occur at a different rate and permits the user to define the initial base frequency rate of heterogeneity between sites. Also allows substitution rates to be reversible.

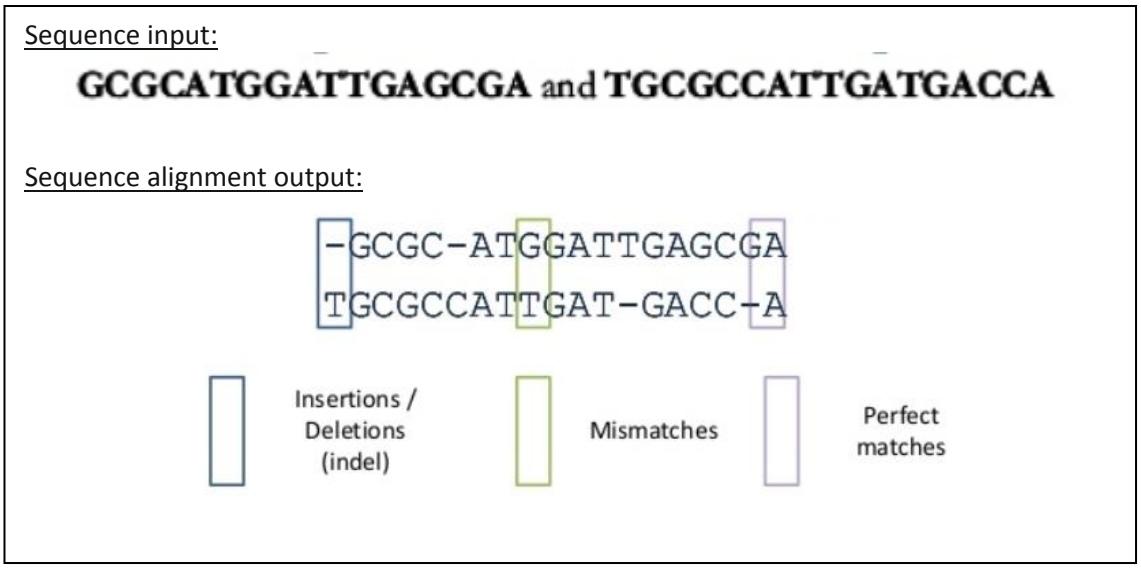
Adapted from Brown (2009)[39] *Transition/transversion (defined above)

2. Sequence alignment

Evolutionary relationships can be defined at many levels depending on the question being addressed. Thus, the building blocks used to infer the relationship by phylogenetic inference can vary from physical/morphological characteristics (e.g. the evolution of the human race) to molecular traits (using nucleotide or amino acids). The more detailed and specific the information, the higher the resolution of evolutionary history inferred. Nucleotides are used in pathogen phylogenetic analysis to provide the high resolution required to determine ancestral relationships.

In order to conduct the analysis, the sequences need to be aligned so that comparisons are made between equivalent (homologous) positions. The nucleotide sequences are aligned in-frame i.e. three codons (a triplet) coding for a specific amino acid. Thus, a data matrix is formed where rows correspond to each sequence and columns correspond to homologous positions within the sequences. When two sequences are compared the percentage similarity can be derived by calculating the number of identical nucleotides relative to the length of the sequence. The genetic differences found at specific nucleotide positions can be categorized as a match, mismatch (substitutions) or gap (insertions or deletions) when compared to a previously determined reference genome of the same organism (Figure 2.4).

Figure 2.4: Schematic of sequence alignment process



Adapted from Brown (2009)[39]

Where a reference genome is not available, a progressive approach can be used where a pair of sequences (with the highest similarity) are aligned to each other (with gaps inserted where necessary in the two sequences to optimise alignment) and this alignment is then "locked" and used as the basis to align another sequence. This continues in an iterative fashion until all of the homologous sequences have been aligned. However, the order in which the sequences are chosen to be aligned can significantly affect the results. When the sequence similarity is so low that an alignment becomes ambiguous, it is better to delete that region of the gene from the alignment.

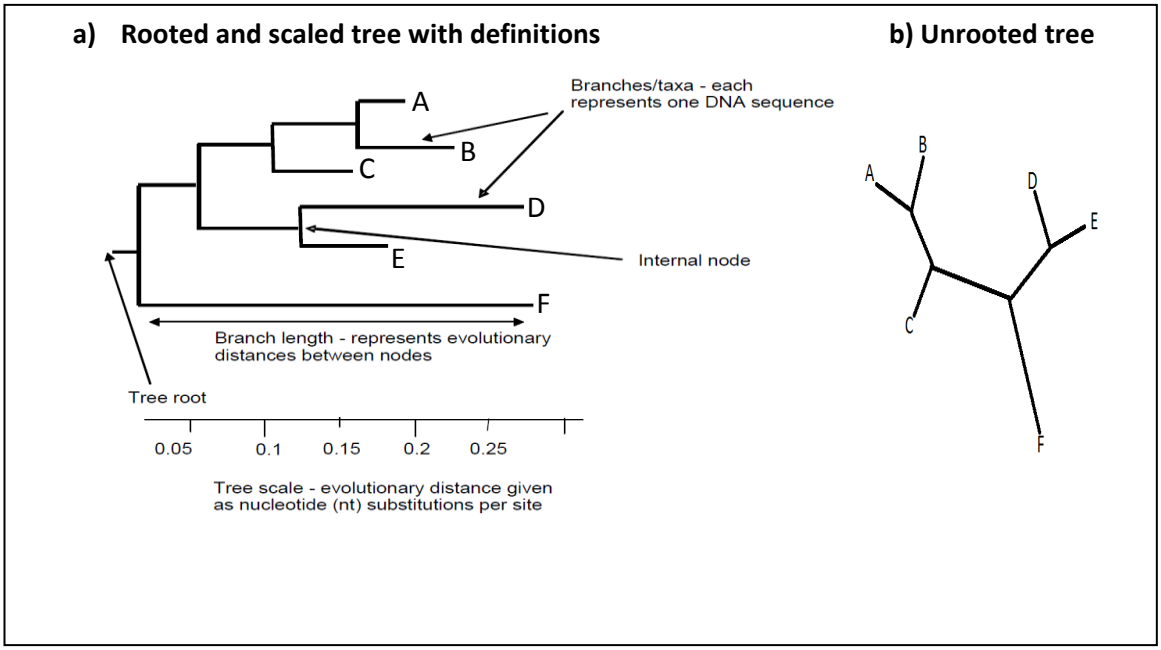
Sequence alignment can be undertaken manually using software such as MEGA[40], AliView[41] or SeaView[42]. Alternatively, automated programmes (e.g. Clustal-X[43]) can crudely align sequences by applying computational algorithms (on the frequency and extent of gaps) and are often used depending on the length of the sequence, number of comparative sequences, and similarities between sequences. This alignment can then be checked and altered manually before further analysis.

3. Methods for inferring phylogenetic trees

A phylogenetic tree is a graphical representation of the evolutionary relationship between aligned sequences (Figure 2.5). A tree consists of branches and nodes. Each branch represents one sequence (or taxa), and branches are joined together by nodes that represent theoretical ancestors. Two or more sequences that share a node are referred to

as a “cluster” or “clade”. The branching pattern, i.e. the order of the nodes and branches, is referred to as the “topology”.

Figure 2.5: Schematic diagram of phylogenetic trees defining terminology used



Adapted from Brown (2009)[39]

Trees can either be rooted or unrooted. A rooted tree has a node defined as the root from which all other nodes have originated. It provides a scale with the branch lengths indicating the genetic distance (usually calculated as substitutions per site) between the two taxa, connected by a common node. The further away a specific node is from the root of the tree, the more recently the node occurred in time. An unrooted, or unscaled, tree shows the sequences relative to one another i.e. the common ancestry of the sequences, but not the extent of genetic divergence.

There are two common techniques to root a tree:

1. Midpoint rooting: the root lies at the midpoint joining the two most dissimilar sequences (i.e. the longest distance between two terminals on the tree, and the root is placed at the midpoint of that distance). This method assumes a molecular clock-like evolution pattern across the lineages within the tree. This is often a reasonable assumption when assessing trees involving the same species, which will likely evolve at similar rates. However, this method may be misleading in instances where lineages evolve at very different rates, or there is a lot of missing data.

2. Outgroup rooting: at least one sequence is known to be outside of the rest of the study group, but not so far away that character homology becomes difficult to establish. A suitable 'outgroup' on which to root the tree may arise from:
   i) the longest branch/most distant sequence, as it is believed to be the most distantly related sequence compared to the others; or
   ii) an additional sequence is included in the dataset for this specific purpose – this should be similar to the sequences under study so as not to distort the phylogenetic tree, and might include a sequence from a very similar pathogen, or a different strain of the same pathogen.

There are many methods for creating phylogenetic trees which vary in complexity, computational demands, and their assumptions about evolutionary theory. None of the methods outlined above, regardless of complexity or other variables, guarantees that the inferred tree is the "true" phylogenetic tree. The most appropriate method will depend on the specific data used and can be categorized as "distance-based" or "character-based" methods, as explained below.

Distance-based methods calculate a measure of dissimilarity between each possible sequence pair in the alignment to produce a pairwise distance matrix. The dissimilarity score is calculated by assessing the proportion of nucleotide positions in which the two sequences differ and this is used to define the evolutionary distance between the sequences. The tree is constructed based on these distance values. The main benefit of distance-based methods is that they are well suited to analyse relatively large data sets quickly and are computationally inexpensive. However, the main criticism is that these methods only produce an overall estimate of the relative distance between sequences (no information about individual sites is given). There are two main distance-based methods: the Unweighted Pair Group Method with Arithmetic Means (UPGMA) and Neighbour-joining (NJ) methods. Table 2.3 below sets out additional details.

In character-based methods, each sequence position in the aligned sequence is considered a "character". The main advantage of these methods is that they are more rigorous than distance-based methods as they make use of each nucleotide position to exploit the maximum amount of data when comparing sequences. Furthermore, they incorporate evolutionary models into the analysis to account for non-linear rates of evolution. There

are three main character-based methods: Maximum Likelihood, Maximum Parsimony and Bayesian methods (Table 2.3).

Table 2.3: Summary of the different methods for inferring phylogenetic trees

Model	Description	Advantage	Disadvantage	Method
Distance-based				
Unweighted Pair Group Method with Arithmetic Means (UPGMA)	Tree constructed through pair-wise distance matrix. Oldest and simplest method. Rarely used now.	Quick	Assumes rate of molecular evolution is constant (i.e. linear). This assumption is no longer accepted and is a frequent criticism of this method.	Identifies the two sequences with the smallest genetic distance (GD) and halves that GD to define the branching patterns. These two sequences are combined and considered as one unit and a new distance matrix is computed between this unit and the other remaining sequences. The process is reiterated until there is only one entry in the matrix. The tree is then built on the basis of the distance matrix obtained.
Neighbour-joining (NJ)	As above, but distances adjusted by incorporating an evolutionary model. Constructs tree sequentially by finding pairs of sequence "neighbours" connected by a single interior node, and clusters by minimizing the length of all internal branches, thus minimizing the tree.	Quick. Allows application of evolutionary model. Computationally undemanding. Software programmes to build NJ trees include: ClustalW and PAUP.	Has been known to build trees incorrectly.	Uses the same method as above, but does not assume the rate of evolution is constant - a specific evolutionary model can be defined.
Character-based				
Maximum likelihood (ML)	ML algorithms search for the tree that maximizes the probability of observing the sequence/character states, given a tree topology and a model of evolution (model of nucleotide substitution).	Exhaustive. Rigorous. Allows statistical hypothesis testing. Allows incorporation of evolutionary model. Software programmes to build ML trees include: RAxML, Phylip and PAUP.	Computationally demanding. Relatively slow.	The method calculates the likelihood of all possible trees for the specified alignment, and selects the one associated with the ML. It exploits the statistical theory of likelihood probabilities. The likelihood (L) that a constructed tree represents the "true" evolutionary relationship is: L=P (D|H) P=probability that the observed sequence alignment (D) is true, given the hypothesis, H (the phylogenetic tree).

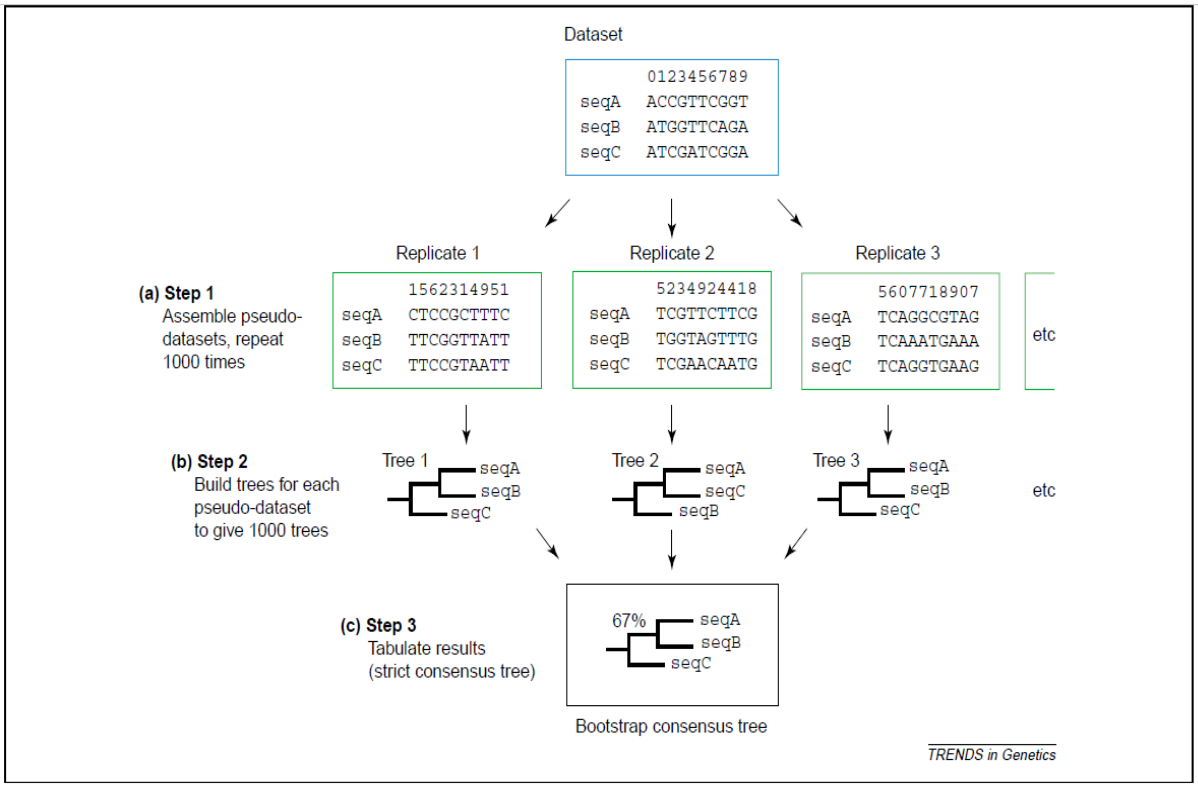
Maximum Parsimony (MP)	Aims to find the tree topology that can be explained with the smallest number of evolutionary differences (character/nucleotide differences) between sequences.	Quick.	Does not always produce single best tree. Two or more trees can be selected as the "best tree". Underestimates true divergence – assumes direct common character inheritance and no multiple substitutions at the same site. Only a small fraction of nucleotide positions used to inform tree.	For each sequence position, the MP algorithm infers the minimum number of character changes required along branches to explain the observed states at the terminal nodes. It always assumes a common character is inherited directly from a common ancestor (i.e. does not allow for multiple substitutions occurring at same site). The sum of this score for all positions is called the parsimony length of a tree. This is computed for different tree topologies and the tree that requires the minimum number of changes is selected as the MP tree.
Bayesian	Uses Bayesian theory and a user-defined evolutionary model to calculate "posterior" probabilities (the probability of a tree, given the observed data) of all possible trees for the alignment. It selects the tree with the highest probability. It searches for all plausible trees to fit the data, based on the user defined model parameters e.g. a *uniform prior* for tree topology is commonest assumption, i.e. every tree is assumed to be equally likely before looking at the data. The data are then used to provide evidence on how to update this prior belief.	Allows genetic distances obtained to be time calibrated. Allows incorporation of evolutionary model. The most frequently used software is Mr. Bayes.	Relatively time intensive, Computationally advanced.	Posterior probabilities are obtained using a sampling/simulation technique called Markov Chain Monte Carlo (MCMC). This method simulates a random set of parameters and proposes a new "state" by altering parameters. In each step, the likelihood ratio and prior ratio are calculated for the new state relative to the current state. For example, it searches at random for the tree with the highest probability "T1" and compares it to another randomly generated tree "T2". If the likelihood of T1 is lower than T2, then T2 replaces T1, if not T1 is held in the memory. The number of times a particular tree is held in memory is proportional to its likelihood, each tree being given a likelihood score on the basis of how often it was held in the MCMC memory. This allows simultaneous independent searches and provides a measure of statistical support for each tree, negating the need for other checks of tree robustness, such as bootstrapping.

Adapted from Brown (2009)[39], Vandamme (2009)[44]

4. Assessing Tree robustness

Given that it is impossible to guarantee that phylogenetic reconstructions will represent the true evolutionary relationship between sequences, a test of the reliability of an inferred tree is necessary. This test provides a measure of tree robustness, which is most frequently estimated using a technique called Bootstrapping[45]. Bootstrapping is a statistical resampling method that works via repeated, random sampling of columns of nucleotide positions. Random subsamples of the dataset are taken - each of the subsamples are the same size as the original, this being accomplished by allowing repeat sampling of sites - random sampling with replacement. A new tree is created on the basis of each subsample. This process is repeated many times (often 500-1,000 times) and the frequency with which the various parts of the tree are reproduced in each of these random subsamples is calculated. For example, if group X is found in every subsample tree, then its bootstrap support is 100%; if it is found in only two-thirds of the subsample trees, its bootstrap support is 67% (Figure 2.6). Bootstrap values are shown on the branches of the tree. As a general rule, if the bootstrap value for a given interior branch is >70-95%, then the topology at that branch is considered robust and "correct".

Figure 2.6: Schematic diagram to show Bootstrap analysis process



The dataset is randomly sampled with replacement to create multiple pseudo-datasets of the same size as the original. (a) Three examples are shown. (b) Trees are constructed from each pseudo-datasets. (c) Each tree is scored for which nodes (groupings) appear and how often. In this case, a node uniting seqA & seqB is found in two of the three replicate trees. This gives a bootstrap support for this grouping of 2/3 or 67%.
From Baldauf (2003)[46]

B) Assumptions and limitations in molecular epidemiology

The successful combination of multiple techniques and disciplines across molecular biology, epidemiology and biostatistics requires an understanding of the assumptions and limitations of each of the techniques employed. In addition, experimental design is another important facet, especially with respect to sampling strategies, such as the methods employed to deal with missing data and the appropriate use of controls to avoid misleading results.

The basic assumption underlying most molecular epidemiological studies is that pathogens with similar sequences are likely to be related such that the degree of similarity can be used to infer the time since their divergence[14]. Thus, sequence data can be used to test epidemiological hypotheses as to whether cases are linked by recent transmission events or not[47]. Furthermore, sequence data can be used in outbreak investigations to increase the precision of case definitions and to explore the characteristics and risk factors for disease transmission[1].

1. Sequencing/PCR-based limitations:

- Mutations in the primer binding sites (Figure 2.3 above) and insertions or deletions in the DNA/RNA can alter the sequence analysis and limit the validity of phylogenies inferred.
- Viral load (VL)/DNA concentrations affect the outcome of the test – the detection limits of PCR systems will vary depending upon the copy number of the target, the primer specificity, and the reaction conditions. Unless the VL is sufficiently high, it is impossible to amplify the DNA for sequence generation, making it difficult to obtain sequences from those patients with HIV on ART. One solution may be to sequence proviral DNA (integrated cellular DNA from infected cells), rather than free viral DNA.
- The reproducibility of results is very sensitive to the reaction conditions employed, and can be affected by variations in equipment and reagents between laboratories.
- The technology does not cope very well with long sequence lengths.
- The potential for contamination in either the field or laboratory will lead to erroneous results, e.g. laboratory contamination may lead to one individual's sequence being amplified and attributed to multiple other individuals, resulting in apparent clustering between these individuals.

- The DNA/RNA degrades in samples over time, called 'storage degradation'.
- Unequivocal identification of the causative agent is critical – it is crucial to ensure there is no morphologically similar, but genetically distinct, species. For example, *Mycobacterium tuberculosis* and *Mycobacterium bovis* are both part of the *Mycobacterium tuberculosis* complex of morphologically similar, but genetically different species.
- Cannot accurately detect multiple infection or intra-host viral diversity (NGS would be needed).

2. Phylogenetic assumptions and limitations

i) Genetic variability

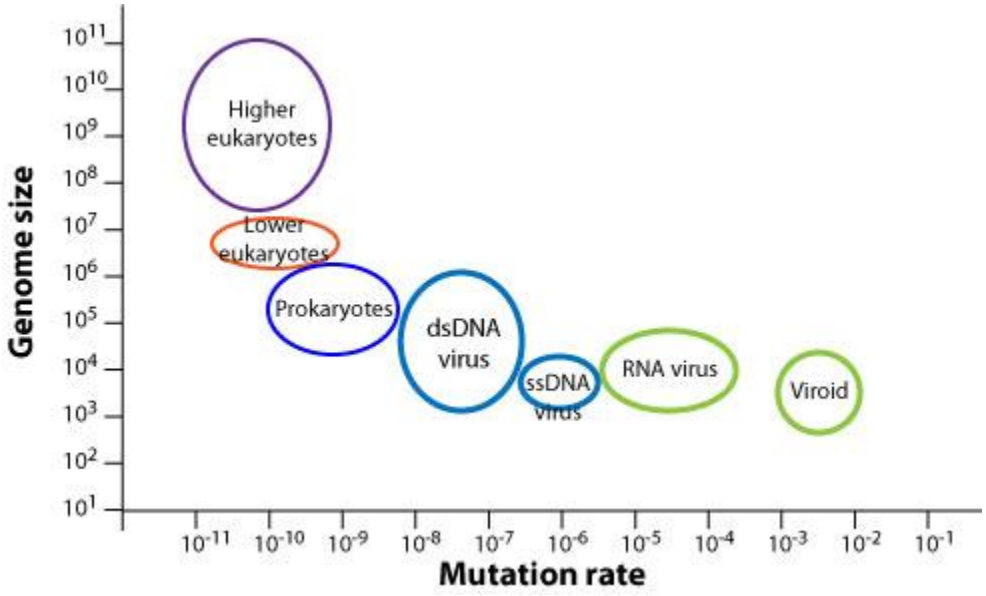
Phylogenetics exploits small differences in the genomes between organisms to determine the evolutionary relationship. Therefore, to allow a spectrum of genetic distances to be identified and used to construct a phylogeny, there must be some genetic variability in the gene under study. There is a fine line between choosing a gene that does not have enough variability (conserved) versus one that evolves so fast that it is difficult to conclude whether or not the sequences are linked.

According to evolutionary theory all organisms evolved from a common ancestor (known as homology), and all organisms have a genetic code (usually as deoxyribonucleic acid (DNA), although some viruses have a ribonucleic acid (RNA) genome). It is this genetic code that forms the basis for inferring phylogenetic relationships in pathogen genomics. Although, in theory, all organisms are amenable to phylogenetic analysis, different mechanisms by which variation occurs have led to the huge biodiversity seen today, and the difference between two genes/species can evolve to such an extent that the sequence data do not carry sufficient information to infer any relationship. Therefore, whether or not phylogenetics can confirm common ancestry (i.e. homology) will depend, among other factors, on the length of time since divergence, the type of variation acquired, and the variability of the gene.

As described above, there are many uses for phylogenetics and different uses may require different data inputs. Given that this thesis explores infectious disease outbreaks in which the relevant time period is usually short (compared to e.g. human evolution), pathogens with variable genes are necessary to study transmission dynamics via phylogenetic

analysis. Viral genomes are the fastest evolving entities in biology (Figure 2.7), primarily due to their short replication time and the large quantity of offspring released per infected cell. This results in defective replication of genomes and quasi-species. The latter is particularly true for RNA viruses. A hallmark of RNA viruses is the retro-transcription and error-prone nature of their replication, due to the absence of proofreading/repair and post-replicative error correction mechanisms that normally operate during replication of cellular DNA. Both HIV and *Ebolavirus* are single-stranded RNA viruses.

Figure 2.7: Mutation rates of different types of organisms



Reproduced from the Swiss Institute of Bioinformatics[48]

The *Ebolavirus* and HIV genomes are described in more detail later in this chapter (Figure 2.8 & 2.11). HIV phylogenetic analysis is largely restricted to *pol* sequence analysis, as it is these sequences that are readily available as a by-product of routine drug resistance mutation screening and from other studies. The *pol* gene encodes for regulatory genes of viral replication and is a relatively well-conserved region of the genome. Owing to this, there has been debate about its use in phylogenetic reconstructions, as there may be insufficient genetic variability[49]. The *gag* and *evn* genes have greater genetic variability and hence may be preferable, but are less commonly available. Huè and Clewley[50] demonstrated that phylogenetic analysis of 140 HIV *pol* sequences produced the same results as analyses using HIV *env* and *gag* sequences. In contrast, reverse transcriptase (RT) and protease genes are suboptimal for reconstructing transmission histories, because the genetic distance between RT and protease isolates from both unrelated and related

individuals may be very similar, making it difficult to differentiate relatedness[49]. Therefore, these are not good candidate genes for phylogenetic inference.

In contrast to HIV, the *Ebolavirus* genome is well conserved with less genetic variability seen between sequences. Therefore, full genome sequences are needed to detect evolutionary history rather than a single gene or genes.

ii) Definition of phylogenetic clusters

There is no consensus regarding the phylogenetic criteria that should be used to establish a putative transmission cluster in population-based studies of transmission networks. Although the criteria may vary between pathogen, gene under analysis and the reason for analysis, there is a general consensus that a cluster or putative cluster can be identified by two or more viral sequences with a low genetic distance (<1.5-4.5%), indicating that the sequences are closely related and share a most recent common ancestor. The lower the genetic distance, the more closely related are the sequences, and the more recent the transmission event (for example, a genetic distance of 1.5% identifies only recent transmissions). In addition, the bootstrap value is an additional criterion that can be used to ensure the robustness of the putative cluster.

As yet, there has been no rigorous assessment of the use of phylogenetic methods in the analysis of population level data in public health. The consequence of this is that there is no formal definition of the criteria required to ensure robustness. Instead, there is a general acceptance that robustness is achieved by using conservative cut-offs, allowing researchers to draw meaningful population level conclusions. Although there is no standard definition of conservative cut-offs, bootstrap support values >70% and a genetic distance of between 1.5-4.5%[50,51] are the most common thresholds used, and these are the criteria I use throughout this thesis, unless otherwise specified. It should be noted that many studies do not use such stringent levels.

Grabowski and Redd[52] reviewed 20 phylogenetic studies of HIV-1 transmission (a convenience sample) and assessed the variation in the criteria used to define clusters. Most studies included both genetic distance and bootstrapping in the definition. However, Grabowski and Redd noted that the rationale for the thresholds used was rarely provided, despite the fact that it may have affected the conclusions drawn. They conclude that the

thresholds used should be linked to the underlying research question. For example, very conservative cut-offs are required in forensic cases, with the primary aim of excluding a common source of infection, rather than confirming a direct transmission event[†52]. In contrast, phylogenetic analyses aimed at investigating transmission networks by identifying direct and indirect linkages within transmission chains, can use more relaxed criteria (described above), thereby increasing the chance of detecting linkages (not only very recent ones).

iii) An alternative explanation for apparent 'linked infections'

There is debate in the literature about the reliability of using phylogenetics to define transmission events. There is an argument that definitive proof of two viruses having a common origin is not possible, as viral genomes can undergo parallel or convergent evolution, resulting in very similar viruses that have no recent common ancestry. For example, as described above, drug resistant mutation sites can lead to artefactual clustering as a result of acquired resistance due to selection pressure, rather than transmitted resistance[53]. Furthermore, phylogenetics alone does not allow identification of the direction of transmission, nor rule out the possibility of missing sequences accounting for linked transmissions, for example a scenario where both individuals in a transmission cluster were infected by the same third partner, or that a third partner was the intermediary between the two. This is highlighted by Leitner *et al*.(1996), who showed that in their study one in 13 implied transmission events were erroneously constructed by phylogenetic analysis.

iv) Does phylogenetic tree topology reflect underlying network structures?

Another area of debate is how closely the phylogenetic tree topology reflects the underlying structure of networks from which the topology is generated. Comparative interpretation is complicated, largely due to a lack of statistical comparison methods. Several studies have suggested that phylogenetic topology does not correspond to underlying population structure, as determined by conventional epidemiological methods[54,55], and that this is particularly true in dynamic populations, for example in populations with high levels of migration. Therefore, the utility of phylogenetics to study

[†] Given the uncertainty associated with all phylogenetic results, it is impossible to confirm a direct transmission from phylogenetic data alone, and therefore, this information does not provide evidence 'beyond reasonable doubt'. However, it is possible to show that two sequences are definitively not linked and disprove a case.

transmission dynamics across different mobile populations (e.g. a global setting) may be more limited, when compared to the localised setting of concentrated outbreaks on which most phylogenetic studies, to date, are based.

v) Summary of limitations associated with the validity of phylogenetic reconstructions

- Concern as to whether the variability of the *pol* gene is sufficient for accurate phylogenetic reconstruction.
- Inability to verify or prove a direct link or transmission, and the ability to only infer a common origin.
- The inability to determine the direction of transmission between linked infections.
- The possibility that a linked infection is attributable to an unsampled third party.
- Inability to distinguish between true linkage and linkage secondary to parallel or convergent evolution e.g. acquired mutations.
- Ability to determine phylogenetic relationships is influenced by the overall genetic diversity of the population.

c) Molecular Bias

Bias can be defined as any systematic error leading to a difference between the estimated and true effects. Not only can molecular epidemiology suffer from the same forms of bias as traditional epidemiology (including ascertainment, participation, misclassification, and information (e.g. non-responder) bias), but it can also suffer from bias specific to molecular techniques - molecular bias. Examples of bias relevant to infectious disease molecular epidemiology studies are outlined in Table 2.4. One of the key differences between traditional and molecular epidemiology in the study of infectious diseases is the inability to obtain viral sequence data on those without the disease for molecular studies, whereas it is still possible to collect traditional epidemiological data from such a group.

Most traditional bias is caused by systematic error attributable to imperfect methodology, for example in study design, sampling strategies, classification procedures, or analysis. This may result in the true relationship between exposure and outcome being misrepresented. This is particularly likely to be true when the data used is from routine data sources. In such a scenario, the study design cannot mitigate potential bias as the sampling strategy was not designed specifically for the research in question. The molecular era has led to a significant increase in the available genetic data, allowing research questions to become

more complex without any substantial increase in the computational capabilities to deal with massive data production.

Finally, molecular epidemiology can be especially prone to error, because of the unknown and often unpredictable ways in which biochemical and molecular markers are associated with exposure and disease.

Table 2.4: Examples of bias relevant to infectious disease molecular epidemiology studies[13,56]

Type of Bias	Explanation and examples
Ascertainment bias	This occurs when there is a systematic distortion in measuring the true frequency of a disease due to the way in which the data are collected, e.g. a difference in the detection of disease in sick versus asymptomatic patients. In HIV phylogenetics, molecular ascertainment bias leads to an over-representation of those who are not on treatment, as it is only possible to sequence those with high VLs.
Participation/ Selection bias	This arises from procedures used to select subjects or samples, or from factors that influence study participation[57]. Participation/Selection bias occurs when those included in the study differ from those who are not, with respect to the outcome/s of interest. There are several types of selection bias, e.g. the 'healthy volunteer effect'. This example is relevant to molecular epidemiological studies as they are often based on populations of volunteers, for example blood donors, who are easy to access, willing to participate and cheaper to sample. Such populations tend to be self-selected and have better lifestyles, biasing the results relative to the general population.
Non-response bias	This occurs if the prevalence of disease (e.g. HIV or Ebola) is systematically different among those who consent to testing compared to those who do not. The reasons for consent or refusal to be tested can be complex, and the extent to which this causes bias in the results depends on whether the factor(s) affecting consent are related to disease acquisition. With respect to HIV, the potential for differential participation according to risk provided the rationale for implementing unlinked anonymous HIV prevalence surveys using residual blood samples without consent from the late 1980s onwards[58]. The scale of this bias is extremely difficult both to quantify and to compare across studies. With respect to this thesis, it is more likely to be an issue with respect to the HIV work.
Result bias	As technology improves over time, so do diagnostic techniques, often with improved sensitivity and specificity. This may introduce a result bias in longitudinal studies. For example, newer techniques may allow sequencing of virus at a lower copy number than previous methods. Currently, HIV genomes obtained via dried blood spot samples can only be sequenced in those patients where the VL is greater than 10,000cpm. However, proviral DNA now allows sequencing even when patients are on anti-retrovirals with an undetectable VL. So over time, VL detection will increase in those with low levels, leading to an overrepresentation of this group, with respect to historical samples.
Detection bias	The probability of identifying a disease increases if a molecular marker or genetic association of early disease is prospectively analysed in a cohort, leading to earlier detection. This in turn has the potential to lead to an apparent increase in survival time due to the earlier detection. An example of this is genetic testing for BRCA genes associated with breast cancer in high risk individuals. If BRCA is detected, individuals undergo more intensive screening and are likely to be more aware of disease, which may facilitate earlier detection of disease. Detection bias is also known as lead time bias.
Sampling Bias	Sampling bias occurs when samples are collected in such a way that some members of the intended population are less likely to be included than others. As with any study, sampling strategy can affect the results obtained as well as the likelihood of detecting

	transmission events. Infectious disease work is often based on outbreak investigation and case control studies, in contrast to the large cohort studies in non-communicable disease studies. Because it is not always possible to determine cases in outbreak settings, we seldom know the true extent of an outbreak and the spectrum of disease that it causes. This is particularly true of infections with an asymptomatic phase or in situations in which the disease manifestations are common to many other conditions, e.g. Influenza versus other viruses causing upper respiratory tract infections, leading to underrepresentation of such cases in the sample. Only sampling the sickest cases, such as in the Influenza pandemic in 2009 (described above), erroneously biased the sample towards the hypothesis that this novel strain was more pathogenic, as a population wide sampling strategy was not undertaken. Therefore, sampling strategies must be determined by the specific infection in question, particularly in relation to the timelines of infectiousness (acute versus chronic diseases) and the type of symptoms (asymptomatic cases versus broad spectrum of severity). In population level phylogenetic analysis, sampling will never be complete and, therefore, clusters will be incomplete. The reason for this is likely to be twofold. Firstly, not everyone will be sampled: even if the population had perfect 'random sampling' some people would be missing at random. Secondly, there is often some sampling bias: for example, it is well recognised at the Africa Centre that the proportion of women who both participate in surveillance and consent to dried blood spot testing for HIV is higher than in men. This is thought to be due to differential working patterns, outmigration patterns and consent rates between men and women. Most HIV phylogenetic work is based on a biased or sparse sampling of HIV transmission networks, not a representative sample of the target population. This may be because infected people are not diagnosed, are not sampled, or have differential consent rates. The consequence of not including every sequence within a population in a phylogenetic analysis is not known, but is likely to bias the results. Most population-based HIV phylogenetic studies have been conducted outside Africa where viral transmission is concentrated within high-risk populations such as men who have sex with men (MSM), commercial sex workers (CSW) and intravenous drug users (IVDU)[59]. Sequences are obtained through convenience or clinic-based recruitment and the sampling proportions of the general population are often unknown. Therefore, the networks are usually under-sampled and the proportion of sequences that phylogenetically cluster rarely exceed 30%[60].

D) OTHER CONSIDERATIONS

1. Statistical power

Another important methodological consideration is the sample size required to achieve statistical power to answer the research question. This is dependent upon the level of precision required. Several studies have shown that sample size is an important determinant of phylogenetic accuracy[61,62]. Brenner and co-workers (2007)[61] have shown that transmission events detected by phylogenetic reconstruction from patients labelled as 'recent infection' were not reproducible when the analysis was repeated with a larger and more diverse sample size, including those categorized as 'chronically infected'. Therefore, the detection of transmission events may be affected by the sample size and relative genetic diversity of the sample. It is likely that genetic diversity,

representativeness and completeness of the entire sample could affect the detection of robust clusters.

2. Confounding

Confounding occurs when the estimated effect of an exposure is distorted due to the effect of another exposure/s – the confounder. In order for this to happen, the confounder must be correlated with both the outcome and exposure of interest, and must not be on the causal pathway from exposure to outcome[57]. Due to multiple and complex gene interactions at many levels, there are likely to be many different types of confounders in molecular epidemiology. However, they are not easy to identify or adjust for, particularly as many of the interactions are unknown.

3. HIV-specific considerations

Sequence generation is only possible from a dried blood spot when the VL is >10,000 cpm. VL is the quantitative detection of HIV RNA in plasma (from free virus). Therefore, this biases the results toward those with recent infection or those with chronic infection, who are not on treatment. VL is known to peak around one month following infection, but remains elevated for approximately 10 weeks[63]. It does not remain static over the course of an individual's infection, but decreases after the initial 10 week period, before gradually increasing over subsequent years if treatment is not received. Successful sequence generation will largely be limited to scenarios of detectable VL. Although this is likely to mimic the population that is at risk of transmitting HIV infection, as high VL is associated with onward transmission, it is not representative of the HIV population as a whole.

4. Ebola-specific considerations

Ebolavirus is a category A organism due to its high infectiousness and associated mortality. Therefore, working with samples from *Ebolavirus* patients requires Level 4 biocontainment facilities, which limits the ease of sample processing and testing. However, the heating and denaturing process of RNA extraction deactivates the virus and allows sample processing in routine laboratory environments.

2.1.3 TRADITIONAL VS. MOLECULAR EPIDEMIOLOGY: ADVANTAGES AND DISADVANTAGES

There remains a lack of consensus on how best to combine traditional and molecular epidemiological techniques, and the utility of doing so. Molecular epidemiology is a relatively new technique and key questions remain regarding its utility:

- Is it informative on its own?
- If so, can we dispense with traditional epidemiology altogether?
- And is there a benefit in combining it with traditional epidemiology?

This section reviews the key differences between the two disciplines, and the advantages and disadvantages of each. It then outlines the limitations of combining both, and, finally, explores the application of a combined approach in the study of transmission dynamics of infections.

The main differences between traditional and molecular epidemiological data sources are that traditional epidemiological data includes both infected and uninfected individuals, whereas molecular datasets only include those who are infected. Epidemiological data also include information on both exposures and outcomes, whereas molecular data do not. As a result, epidemiological data can elucidate a more detailed picture of a population. In addition, as I will detail in subsequent chapters, this means that epidemiological data can improve the inferences made from molecular data. An additional difference between the two data sources is the time point in reference: epidemiological data refer to the time of data collection (current time), whereas molecular data refer to the time of transmission, which may be retrospective by many years.

Phylogenetics is a growing and exciting field and has several benefits over traditional epidemiology. For example, it can provide critical information about epidemics that is difficult to determine from traditional epidemiological studies, such as transmission of drug-resistant mutations, mixing patterns between demographic and behavioural groups and the rapidity of viral spread within populations[52]. Another key advantage of phylogenetic analysis is the ease of obtaining retrospective samples: guidelines recommend that newly diagnosed HIV infections are sequenced and tested for drug resistance mutations to inform treatment options[64]. These routine tests lead to a wealth of data and many libraries of *pol* sequences, which are appropriate for phylogenetic

reconstruction. A third benefit is that 'sequences don't lie', providing an objective data point. This is particularly important when comparing sequence data to self-reported epidemiological data in relation to sexual and other behaviours. Due to the sensitivity of the sexual behavioural data required and the stigma in some cultures associated with certain high risk sexual behaviours (e.g. polygamous sexual relations, commercial sex work and MSM), these data are difficult to capture accurately. Where these data do exist, they are often of poor quality and some data suggest that informants are not always transparent and accurate in their reporting.

In the more localised acute outbreak scenario, for example an outbreak of Ebola, obtaining high quality epidemiological and contact tracing data is time intensive, requires training to complete properly, and is, therefore, expensive. Furthermore, there may be language barriers and direct contact with those carrying a contagious infection can be hazardous. In contrast, obtaining blood or swab samples with which to undertake sequence analysis, or opportunistically using residual samples from routine diagnostic sampling, may be easier and cheaper in the case of some infectious diseases than time-consuming traditional epidemiological techniques. In contrast, the cost of data processing and analysis is lower in the case of traditional epidemiology. However, in the case of Ebola, both the sample collection and processing for molecular analysis would require high levels of expertise, personal protective equipment and category 4 laboratories, given that there were no residual sample sets upon which to rely. Therefore, if high quality, reliable and robust epidemiological data are available, this may render phylogenetics somewhat redundant.

Individually, both techniques have their limitations. However, the strengths of one technique can complement the limitations of the other. This makes combined analysis a potentially powerful tool. Research is needed on how best to combine the techniques to understand their full potential. Although the general consensus is that these tools are likely to be most powerful when combined, there remains sparse literature on methodologies to do this and a limited number of concrete examples of the benefits, particularly at the population level.

Combining the two disciplines can exploit the information contained within each dataset to an additive level. Phylogenetic analysis can become more powerful when combined with clinical and diagnostic data because this validates and increases the certainty of the

results, and also allows answers to questions that could not be provided by either technique alone. Together they may reveal critical information relevant to disease control, viral transmission dynamics, associations between socio-demographic characteristics and the time scales for epidemic evolution. For example, integrating the two has demonstrated the epidemiological utility of phylogenetic inference when supplementing surveillance data in HIV epidemics by revealing misclassification of homosexual transmission events in those who prefer to be identified as heterosexual[65]. Furthermore, it may be possible to identify likely linked infections by either technique, but only when you combine the techniques can you tackle questions such as 'who (likely) infected whom' and the 'direction of transmission'.

It is widely accepted that there is no current mechanism to determine directionality of infection, but it would be useful if this were possible. Over the last three years, work in resolving the directionality of infections has increased significantly. Developments are discussed further in Chapter 10. Molecular data from pathogens may be useful for identifying the source of infection and identifying onward transmission pairs. However, inference of who acquired infection from whom is limited by incomplete sampling, as shown in the MRSA outbreak described earlier (2.1.2 (c)). Unsampled cases may act as either a common source of infection or as an intermediary in a transmission chain for hosts infected with genetically similar pathogens. It is difficult to quantify the probability of common source or intermediate transmission events, which makes it difficult to develop statistical tests to either confirm or deny putative transmission pairs, with genetic data. Given genetic data only, it is not possible to infer the direction of transmission in a transmission pair. However, given additional information on clinical, behavioural, and demographic covariates of the infected hosts, it may be possible to define directionality. New phylogenetic and modelling approaches are being used to attempt to determine directionality of transmission and to overcome the current difficulties in determining this. Volz and Frost[66] present a method to incorporate additional information about an infectious disease epidemic –such as incidence, prevalence, and proportion of hosts sampled –to inform estimates of the probability of direct transmission events within a sample. These methods may make it possible to correct for biases stemming from incomplete sampling of the infected host population, which is common in high-morbidity community-acquired pathogens like HIV, Influenza and Dengue virus. These methods also enable forensic applications, such as source-case attribution for infectious disease

epidemics with incomplete sampling, and epidemiological applications, such as the identification of factors that increase the risk of transmission. This facilitates the development of targeted prevention methods for individuals most at risk of transmitting HIV, thereby increasing efficiency, effectiveness and impact[67-70]. Such targeted prevention methods require a finer resolution of HIV transmission dynamics than currently exists with traditional epidemiological methods. In combination, these tools can elucidate these required detailed descriptions[52].

Therefore, molecular and traditional epidemiology in combination may be able to elucidate the direction of transmission, and highlight both 'super-spreaders' (risk groups likely to generate HIV infection), and those at risk of HIV acquisition. This provides information to contribute to the identification of targeted public health responses. Going forward, the routine use of phylogenetic inference together with empirical epidemiology represents a potential major advance for use in infectious disease surveillance, monitoring of prevention strategies, and in informing public health policy, all of which have previously relied solely on clinical and epidemiological data. Thus, there appears to be a role for embedding these techniques in the standard toolkit for future policy formation with respect to infectious disease outbreaks. Studying two different pathogens in this research can help to provide a model to extrapolate these new approaches in order to study the transmission dynamics of other pathogens and settings. However, additional theoretical studies are needed to prove this added utility and to relate HIV phylogenies to network structure and transmission processes, particularly at the population level and in dynamic populations.

2.2 Background to HIV Epidemic

2.2.1 Background

The first cases of Human Immunodeficiency Virus (HIV) were cases of Acquired Immune Deficiency Syndrome (AIDS) recognised in the United States of America (USA) and Western Europe among men who have sex with men (MSM) in the early 1980s. HIV was isolated in 1983 and subsequently shown to be the cause of AIDS. Later, case reports also indicated transmission via injecting drug use (IDU), mother to child transmission, and through heterosexual sex. Over thirty years later, HIV remains a global epidemic.

Despite advances in HIV prevention, incidence continues to be high in certain areas and the overall prevalence continues to increase, in part due to successful treatment programmes resulting in people with HIV living longer. In 2013, there were approximately 2.1 million new infections and 35 million people worldwide living with HIV[71]. Sub-Saharan Africa accounts for 75% of all HIV infections and Kwa-Zulu Natal (KZN), the area in which data for this research were collected, is the worst-affected province of South Africa (prevalence 16.9%)[72]. This prevalence hides a significant heterogeneity in subpopulations, some of which have a prevalence of greater than 50%[67,68].

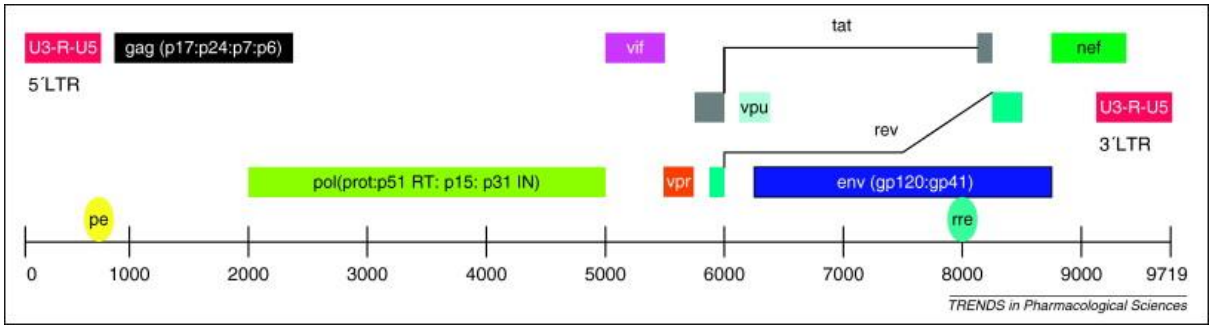
In an attempt to end the AIDS epidemic as a major public health threat by 2030, UNAIDS has set ambitious goals to expedite major reductions in HIV-related mortality and new HIV infections. These include the Fast-Track '90-90-90' targets by 2020: 90% of all people living with HIV will know their HIV status; 90% of all people with diagnosed HIV infection will receive sustained antiretroviral therapy; and 90% of all people receiving antiretroviral therapy will have viral suppression[73].

HIV is arguably the most significant epidemic of the current era. Fully understanding HIV transmission dynamics is challenging due to the complexity of the networks along which it spreads. These differ between populations, risk groups, cultures and communities. Therefore, novel methods are needed to help the study of generalized epidemics across different countries and cultures – one such approach is the combined use of molecular and traditional epidemiology.

2.2.2 HIV GENOME

HIV is a retrovirus with a genome consisting of 9,181 nucleotides of RNA divided into nine genes encoding 14 viral proteins (Figure 2.8). There are three major genes and six accessory, or auxiliary, genes. The HIV genome mutation rate is approximately $1.2x10^{-3}$ substitutions per site per year.

Figure 2.8: The HIV genome



From Sun *et al.*(2011)[74]

The three major genes encode proteins essential for viral function: *gag* encodes for internal structural proteins, *env* for transmembrane envelope proteins, and *pol* for enzymatic proteins essential for viral replication such as reverse transcriptase, protease and integrase. *Pol* is a more conserved gene than *gag* and *env* and it is the *pol* gene that forms the basis of the work in this thesis. The partial *pol* sequence used henceforth is 1,907 nucelotides long.

There are two types of HIV: HIV-1 and HIV-2[75]. HIV-1 accounts for the majority of infections globally, whilst HIV-2 is primarily localised and more prevalent in West African countries. For the remainder of this thesis all references to HIV are to HIV-1. HIV-1 can be further categorised into four groups based on genetic similarities: M (main), N (new), O (out-group) and P (possibly after the researcher who discovered this group - Plantier)[76-78]. Each group represents an independent zoonotic transmission of Simian Immunodeficiency Virus (SIV) from chimpanzees or gorillas into humans[76]. Group M is the predominant type, responsible for the majority of global infections (>90%), while groups O and N are restricted to west and central Africa[78]. Group P, which was isolated in Cameroon, is very rare[79]. Due to the substantial genetic diversity within Group M, the viruses are again further subdivided into nine subtypes (A-K), and more than 89 mosaic 'circulating recombinant forms' (CRF) and 'unique recombinant forms' (URF) also exist[76,78]. Subtype C is the most prevalent clade globally accounting for nearly 50% of infections[80]. The majority

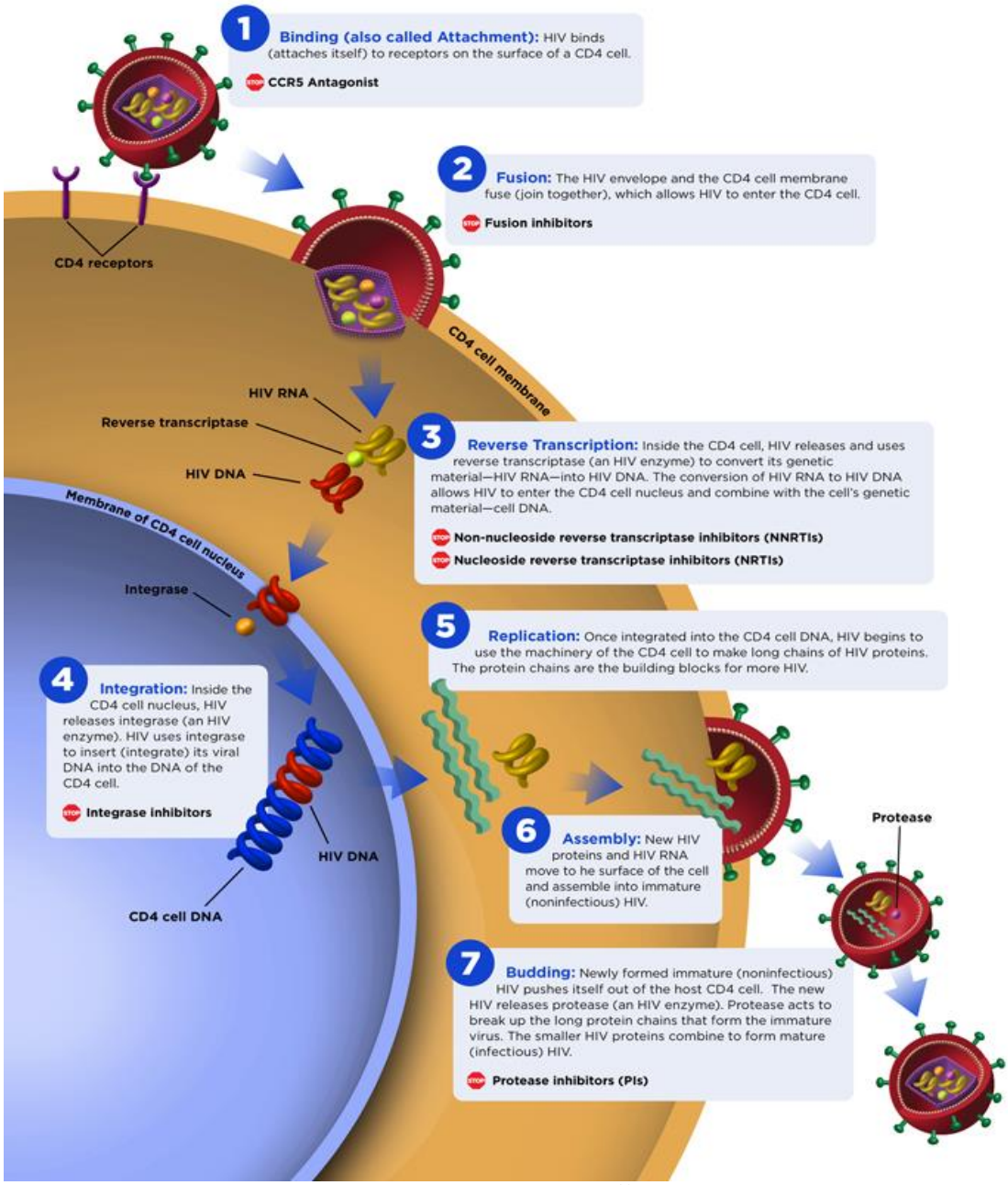
of subtype C infections are in southern Africa, where it is the predominant subtype. For this reason, HIV-1 subtype C infections are the focus of this research. Subtype C also dominates the epidemics in India, Ethiopia, southern China and is noted in east Africa, Brazil and many European countries. Subtype B (12% of infections globally) is associated with infection in Western Europe and North America, and most frequently found among MSM and IDUs[81].

2.2.3 Life cycle of HIV

The seven stages of the HIV life cycle are summarised in Figure 2.9. Upon infection, HIV binds to CD4 receptors on the host's CD4+ T-lymphocyte (CD4+) cells, which are key to host immunity. Supplementary interaction with two co-receptors allows the virus to fuse with, and enter, the host cell. The HIV-RNA is converted into HIV-DNA using the reverse transcriptase enzyme, and enters the host cell nucleus where it is integrated into the host's DNA by the integrase enzyme. This provirus can remain latent in the nucleus or be active, generating products for the creation of new virions and resulting in the transcription of viral DNA into messenger RNA, which is then translated to viral proteins. The viral RNA and viral proteins assemble at the cell membrane into a new virus, which is then released. This resulting free virus is capable of infecting another host CD4+ cell and it can be used as a measure of how much HIV is in the body by testing for 'viral load', a measure of the level of free viruses circulating in the host. Subsequently, the originally infected host cell dies. As more CD4+ cells are infected, the host's immune capacity gradually diminishes, increasing susceptibility to infections which would otherwise not cause disease – termed 'opportunistic infections'.

Continued overleaf.

Figure 2.9: The HIV life cycle



From US Department of Health and Human Services, AIDSinfo[82]

2.2.4 Clinical HIV – natural history, diagnosis and treatment

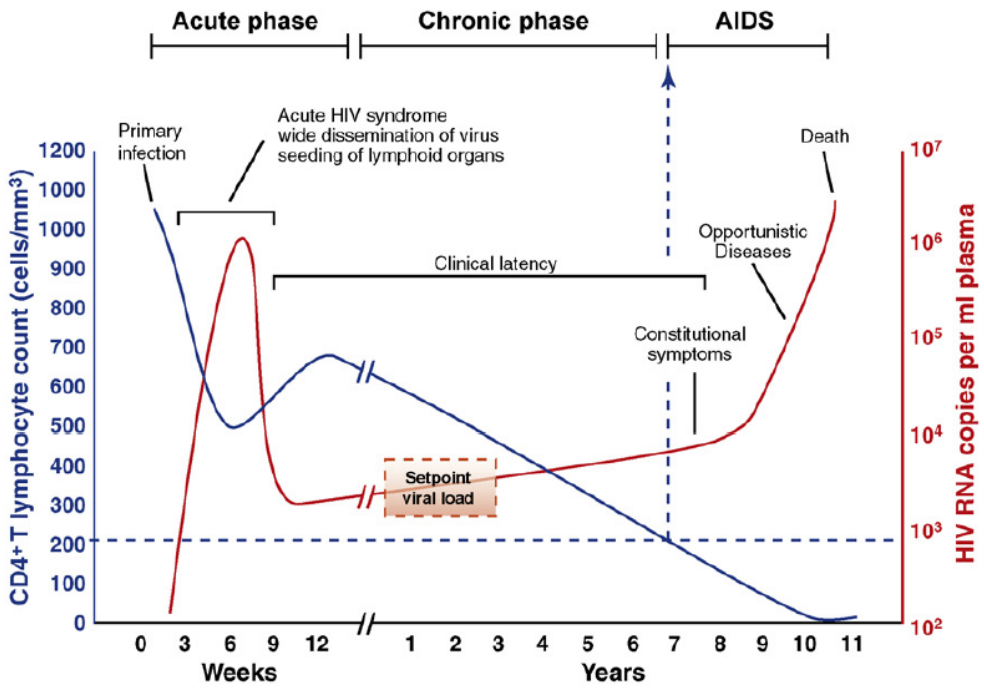
There are three main stages of HIV infection from initial infection to death in an untreated individual:

1. Recently acquired infection

Also known as primary, early or acute infection, this period covers the first few weeks following infection. Between 40-90% of patients experience some seroconversion symptoms of early HIV, such as fever, rash, sore throat, and/or lymphadenopathy. During this time, HIV levels increase in the blood reaching VL levels of up to 100 million copies/ml.

The blood levels peak around day 17, but serum levels do not peak until around day 30 following infection. However, a specific antibody response may not have developed yet and the CD4+ cell count decreases transiently as a result of cell death following viral replication. Once an antibody response has been generated by the host, HIV replication is restrained and CD4+ cell counts generally return to normal. Figure 2.10 outlines the key laboratory parameters across the different stages of infection.

Figure 2.10: Course of HIV infection



From An and Winkler (2010)[83] (reuse approved by Elsevier)

2. Asymptomatic infection

   This period begins from weeks to months (up to 3 months) after infection and can continue for around 10 years. During this period viral replication and CD4+ cell turnover is moderated at low levels by the antibody response and the patient will not usually have any clinical symptoms. The immune system gradually weakens over time.

3. Late symptomatic disease and AIDS

   This period commences once the immune system is sufficiently weakened and CD4+ cell counts reach low enough levels (<200 cells/mm^3) for the patient to develop clinical symptoms. Patients are susceptible to opportunistic infections, i.e. infections that are not pathogenic to those with normal immunity, but cause disease in the immunocompromised. The diagnosis of specific opportunistic infection (AIDS defining

opportunistic infections (OI)) defines the onset of AIDS e.g. Kaposi's Sarcoma, Pneumocystis pneumonia, or extra-pulmonary TB.

WHO often uses a clinical staging classification for HIV infection on which to base guidelines which divides infections into the following categories: stage 1 - asymptomatic, stage 2: mild symptoms, stage 3: advanced symptoms (AIDS related conditions), and stage 4: severe symptoms (AIDS defining illnesses)[84].

Investigations

The diagnosis of HIV infection is based on the detection of antibodies, antigens, or both. Many commercial kits are available and point of care rapid antibody tests have greatly increased the availability and uptake of testing, as results can be obtained within minutes. However, a major limitation to these serological tests is detection during primary infection when antibodies are absent (the initial 6-12 weeks following infection), and possible false positives in infants <18 months old, who still have maternal antibodies. In these instances, virus detection is necessary by quantification of viral RNA, proviral DNA and/or p24 antigen testing.

Once diagnosed, quantification of CD4+ cell counts (counts per mm^3 by flow cytometry) and VL is undertaken and used to monitor infection and response to ART. CD4+ cells provide an indication of how well an individual's immune system is functioning, and the VL measures the free virus in the blood (as described above). Furthermore, the higher the VL, the higher the risk of transmitting virus[85].

Treatment

Antiretroviral therapy (ART) is used to supress viral replication, thereby preventing progressive damage to the host immune system, clinical progression, and HIV-associated morbidity and mortality. However, current treatments are unable to eradicate the infection and lifelong treatment is needed. If treatment is adhered to it is very effective and life expectancy can be entirely normal, with VL suppression to undetectable levels, and rebound of CD4+ cell count to within normal parameters[86].

When to initiate ART is defined by WHO criteria, adapted by each country independently. Previously, this was based on CD4+ cell counts, plasma viraemia and clinical condition. However, in an attempt to achieve the ambitious goals set in the UNADIS Fast-Track targets for 2020 (above), guidelines have recently been adapted to treat all adults living with HIV regardless of CD4+ count, with priority given to initiate treatment in those with CD4+ counts <350 cells/mm^3 or with advanced clinical disease[87].

Treatment usually involves a combination of three or more active drugs. However, fixed dose combination tables with long half-lives have led to simplified regimes to improve adherence to therapy[64]. Some patients still experience adverse side effects. An increasing problem is the emergence of drug resistant viruses, leading to treatment failures.

The drugs inhibit key stages of the HIV lifecycle and are classified accordingly to their mechanism of action (Figure 2.9)[78]: reverse transcriptase inhibitors; fusion inhibitors; integrase inhibitors; and protease inhibitors. Reverse transcriptase inhibitors include nucleotide reverse transcriptase inhibitors (NRTIs) and non-nucleotide reverse transcriptase inhibitors (NNRTIs). Both work by inhibiting the enzyme reverse transcriptase from transcribing the viral RNA into DNA. NRTIs compete with nucleotide analogues for incorporation into DNA, and NNRTIs inhibit replication by binding to the active site of reverse transcriptase. Fusion inhibitors block HIV from fusing to the host cell's membrane. Integrase inhibitors prevent provirus from integrating into the DNA of the host cell. Protease inhibitors (PIs) bind to the active site of protease, thereby preventing the production of viral proteins for the final assembly of new virions.

Drug resistant viruses

HIV drug resistant viruses contain mutations that reduce the susceptibility of HIV to ART. Drug resistance to ART can be "acquired" or "transmitted". Acquired resistance is developed following suboptimal adherence to treatment by the patient, allowing resistance to develop, or via inducible selection pressures on the virus. Transmitted drug resistance is inherited through infection with a resistant strain. Both types of resistance are associated with poorer outcomes[86].

2.2.5 Previous HIV phylogenetics work

HIV phylogenetics work has largely been developed in high-resource settings where HIV is characterised by small, concentrated epidemics, often centred on specific risk groups e.g. MSM. In these settings routine drug resistance testing facilitates easy access to samples for sequencing. Previously, this work has centred on the study of discrete historical epidemics (origins of infection), sub-epidemics within risk or demographic groups, and also discrete transmission chains[65]. More recently, in Europe and North America, these studies have begun to combine phylogenetics with clinical, demographic, and geographic metadata[6] to study risk groups in these concentrated outbreaks (Table 2.5).

Table 2.5: Examples of phylogenetic studies incorporating epidemiological data in the study of concentrated epidemics in high-resource settings

Study type	Findings
Studies to explore transmission events in high-risk populations	These studies are frequently undertaken in discrete populations, such as MSM populations, in recently diagnosed patients. These datasets are relatively easy to access (as the data are generated routinely in testing for drug resistance transmission on diagnosis) and represent those with a high risk of onward transmission potential. The hypotheses of these studies are often to show elevated transmission potential[88,89].
Studies to assess transmission rates between different groups	Brenner *et al*. (2007)[61] studied the proportion of new transmission events in Canada from recently HIV-infected populations compared to the chronically HIV-infected population (49% vs. 27%). Given the recently HIV-infected population comprises approximately 10% of this total HIV-infected population, this suggests that those who are recently infected are disproportionately responsible for propagating ongoing infection[60]. However, Brown *et al*.(2009)[39,90] suggest that these figures may be an overestimation through failing to recognize that this infection stage is transitory. They suggests that recently HIV-infected MSM have a transmission risk of 3.04 compared to the chronically infected. Transmission rates were particularly elevated among the untreated population, with 72% of infections generated from treatment naïve MSM and 23% from MSM interrupting treatment. Overall, 69% of transmissions occurred from MSM with CD4+ counts >350 cell/mm^3 (the historical threshold for ART), therefore contributing to the policy debate on the public health benefit of universal treatment for all HIV-diagnosed individuals.

These combined methodologies have allowed phylogenetic analyses to be used more widely in HIV epidemiology. For example, they have been used:

i) To identify and describe HIV transmission pathways – to study viral linkage, characteristics associated with transmission, and risk factors for epidemic spread (molecular epidemiology)[6,91]

ii) To estimate the growth/decline of the HIV epidemic (phylodynamic tool)[92-95], including key epidemic parameters

iii) To explore the impact of migration on HIV spread and to identify hubs of transmission (phylogeography)[96].

However, population level phylogenetic analysis has rarely been carried out for a generalised epidemic in a low-resource setting, which is the focus of this research. Finally, a noteworthy mention for HIV phylogenetics during the course of this research, was the vindication of Gaëtan Dugas, often referred to as Case Zero, and widely blamed in the media for bringing HIV to the USA. In 1984, early in the HIV epidemic, CDC published a study – "*Cluster of cases of the acquired immune deficiency syndrome*"[97]. In this paper, they explore the possibility that AIDS was sexually transmitted and describe a cluster of 40 cases in 10 cities linked by sexual contact. Gaëtan Dugas was included in this study, denoted as "Patient O" to signify he came from "outside California" where the study originated (GD was a Canadian flight attendant). However, several years later, this denotation was misinterpreted by a journalist as "Patient 0" (zero) and Gaeten Dugas was named in the mainstream media. In 2016, phylogenetic analysis confirmed that there was no evidence that Gaeten Dugas was the primary case in the USA. This study provided strong evidence that HIV in the USA emergence from a pre-existing Caribbean epidemic, and was estimated to jump to the USA around 1970 (New York City), before spreading across the country[98].

2.3 Background to Ebola Epidemic

Ebolavirus causes a severe haemorrhagic fever in humans with a high case-fatality rate (25-90%)[99-101] and significant epidemic potential, as shown by the recent 2013-2016 West African outbreak. This section describes the background to Ebola Virus Disease (EVD) including virus classification, genome structure, history of EVD outbreaks, epidemiology, and clinical characteristics. Further information about the West African outbreak is described in detail in Chapter 4.

Ebolavirus is a member of the Filoviridae family. There are five strains that have been identified: *Zaire*, *Sudan*, *Bundibugyo*, *Taï Forest* and *Reston*. The first three cause the majority of disease in humans[99]. Both *Taï Fores*t and *Reston ebolavirus* cause disease in non-human primates, but recorded infections in humans are limited to one case of *Taï Forest ebolavirus* and largely asymptomatic infections with *Reston ebolavirus*[102]. The West

African Ebola outbreak was caused by the *Zaire* strain, with the specific variant responsible for the outbreak being named "*Makona*" strain, after a river flowing through the 3 main affected countries[103].

Ebolavirus is a category A Bioterrorism agent (classified by CDC). Working with live *Ebolavirus* poses safety risks and requires category 4 safety laboratories. Relatively little was known about the virus prior to the recent outbreak, but due to the potential bioterrorism risk limited work had been undertaken to begin to develop vaccines as a precaution. The unprecedented nature of this recent epidemic demonstrated the potential threat to global health, and work was expedited to greatly enhance knowledge and treatment options.

2.3.1 The *Ebolavirus* genome

The *Ebolavirus* genome (EBOV) consists of a single strand of negative sense RNA and contains 7 protein-coding genes, which when collated are 14,647 nucleotides in length (Figure 2.11)[104]. EBOV substitution rate in the *Zaire-Makona* strain has been estimated at between 0.87x10-3 and 1.43x10-3 mutations per site per year (equivalent to 16-27 mutations in each genome).This suggests the sequences diverge rapidly enough to identify distinct sub-lineages, during a prolonged epidemic[105,106].

Figure 2.11: The Ebolavirus genome

The genome is depicted in the 3'-to-5' orientation to indicate that the genomic RNA is negative sense. The proteins encoded are nucleoprotein (NP), viral protein 35 (VP35), VP40, glycoprotein (GP), soluble GP (sGP), VP30, VP24 and large protein (L). Not drawn to scale.
From Messaoudi *et al* (2015).[107]

2.3.2 Outbreaks of Ebola

At the time of writing, there have been 29 recorded outbreaks or case reports of EVD since Ebola was first identified in 1976 (Table 2.6)[99]. The majority of human disease has arisen from 15 outbreaks caused by the *Zaire* strain and eight by the *Sudan* strain. These outbreaks and cases were limited to rural communities in Sudan, Democratic Republic of Congo, Republic of Congo, Gabon and Uganda. Most of these outbreaks were small in size with just seven outbreaks involving more than 100 cases. The largest of the outbreaks

prior to 2013 occurred in Uganda in 2000-2001 with 425 cases and 224 deaths[108]. The 2013-2016 outbreak in West Africa was unprecedented in scale, being larger than all other outbreaks combined, with 28,639 reported cases and 11,316 deaths (28/02/2016)[109]. This outbreak is described in detail in Chapter 4.

Table 2.6: Previous Ebola outbreaks/infections in humans

Year	Countries	Number of outbreaks	Number of cases	Number of deaths	Viral strain
1970-79	Zaire, 1976[a] Sudan, 1976[b] United Kingdom, 1976	2 2 1	319 318 1	281 173 0	*Zaire* *Sudan* *Sudan*
1980-89	Philippines, 1989-90	1	3[c]	0	*Reston*
1990-99	USA, 1990 Gabon, 1994 Côte d'Ivoire, 1994 DRC, 1995 South Africa, 1996 Russia, 1996	1 3 1 1 1 1	4[c] 149 1 315 2 1	0 97 0 250 1 1	*Reston* *Zaire* *Taï Forest* *Zaire* *Zaire* *Zaire*
2000-09	Uganda, 2000-01 Gabon, 2001-02 Republic of Congo, 2001-02 Sudan[b], 2004 Russia, 2004 DRC, 2007 Philippines, 2008	2 1 3 1 1 2 1	574 65 235 17 1 296 6[c]	261 53 200 7 1 202 0	*Sudan/Bundibugyo* *Zaire* *Zaire* *Sudan* *Zaire* *Zaire* *Reston*
2010-13	Uganda, 2011&2012&2013 DRC, 2012	3 1	18 36	8 13	*Sudan* *Bundibugyo*

[a] Now Democratic Republic of Congo (DRC)
[b] Now South Sudan
[c] Asymptomatic infection

Table adapted from CDC website[108]

The first documented outbreak of Ebola occurred in 1976 in northern Zaire (now the Democratic Republic of Congo)[100]. It occurred in and around a mission hospital in Yambuku, adjacent to the Ebola River, after which the virus was named. In total, 318 cases were identified with a case fatality of 88%[100]. Nosocomial transmission played a significant role in this first outbreak, as it would in future Ebola outbreaks, when the re-use of contaminated needles was found to be the major factor in the initial spread of the virus.

Two months prior to the outbreak in Zaire in 1976, a similar outbreak occurred in the south of Sudan, which initially did not attract international attention[110]. This outbreak originated in a cotton factory in Nzara and was amplified by transmission in a local hospital

in Maridi, both close to the border with Zaire. Due to the close proximity, these two outbreaks were initially thought to be linked, but were later found to be caused by two separate strains of *Ebolavirus* - *Zaire* and *Sudan* strains. A total of 284 cases were identified in the Sudanese outbreak with a case fatality of 53%[108,110].

An additional strain, *Reston ebolavirus*, was first identified in the USA in monkeys imported from the Philippines in 1989[111]. The virus was transmitted to humans working with the chimpanzees, as shown by an antibody response. Although none of the humans showed classical symptoms of EVD (see below), mild fever and non-specific symptoms were reported. In addition, this strain has been identified in pigs with asymptomatic transmission to at least six humans reported in the Philippines[108].

In 1994, the *Taï Forest ebolavirus* was reported to cause an outbreak among chimpanzees in Cote d'Ivoire. A veterinarian studying the outbreak was subsequently infected following post-mortem examinations of the chimpanzees[112]. The veterinarian is the only known human to have been infected with this strain, and she survived after being hospitalized in Basel, Switzerland.

The *Bundibugyo* strain was responsible for two outbreaks in Uganda in 2007 and Democratic Republic of Congo in 2012. The Ugandan outbreak resulted in 131 cases and appears to have the lowest mortality (32%) of all viral strains that cause EVD in humans[108].

In August 2014, during the height of the West African outbreak, another outbreak of EVD was confirmed in rural Jeera County, Democratic Republic of Congo (*Zaire* strain). Sequencing data showed the two outbreaks to be unrelated[103], with the index case identified as a pregnant woman infected after contact with bush meat. The scale of this outbreak was in keeping with pre-2013 outbreaks, with 66 reported cases and 49 deaths[113].

In addition to the outbreaks described here, isolated individual cases of humans being infected have been reported, including imported cases from Gabon to Johannesburg, South Africa (*Zaire* strain) and three laboratory accidents - one in the United Kingdom (*Sudan* strain) and two in Russia (*Zaire* strain)[108].

2.3.3 Clinical Ebola – natural history, diagnosis and treatment

Outbreak emergence

EVD is a zoonotic infection - outbreaks begin in humans following a spill over event from an animal to a human. The exact trigger for these events remains unclear, although a recent publication has suggested a strong link between deforestation and the timing of Ebola outbreaks[114].

The animal reservoir for *Ebolavirus* remains uncertain, but current evidence suggests that bat species may be a natural reservoir, with other species acting as intermediate hosts[115]. The evidence to support this includes:

- Experimental inoculation of animals epidemiologically linked with outbreaks of *Ebolavirus* demonstrated viral replication in fruit and insectivorous bats without fatalities[116].
- Trapping studies have found wild bats with evidence of *Ebolavirus* "infection" and immunity (RNA fragments and Ebola specific antibodies) in areas, which have had human and non-human primate Ebola outbreaks[117].
- *Ebolavirus* has been isolated from other animals, including non-human primates and duikers (antelope). These species however appear to be unsuitable reservoirs because they have high case fatality rates during outbreaks[115,118].
- However, live *Ebolavirus* has never been isolated from a bat in nature, which makes the picture less clear[118].

Furthermore, a recent publication has suggested a strong link between deforestation and timing of Ebola outbreaks.

Disease transmission

Outbreaks of EVD are thought to originate when the index case becomes infected through contact with the blood or body fluids of an infected animal. Once the index case becomes ill or dies, the virus spreads to others, who come into direct contact with the blood and/or body fluids of the infected person. Individuals are only considered to be infectious when symptomatic, not when asymptomatic during the incubation period. The incubation period between infection and symptom onset is between 2 and 21 days, with a mean of 11.4 days[119].

Clinical Presentation

Most cases of EVD begin with the abrupt onset of fever and malaise 6 to 12 days after exposure (range 2 to 21 days). There is a broad spectrum of symptom presentation from very mild to severe haemorrhagic complications (rare). Common early symptoms include headache, vomiting, diarrhoea, myalgia, rash, and hiccups, and can then lead to haemorrhagic symptoms, uveitis, conjunctival injection and neurological symptoms (meningoencephalitis with altered consciousness, neck stiffness, and seizures). Gastrointestinal symptoms are very common and can lead to significant fluid loss and electrolyte disturbance, resulting in hypovolaemic shock or arrhythmias and sudden death.

Traditionally, haemorrhagic symptoms were thought to be the defining feature, and indeed, EVD was previously named 'Ebola haemorrhagic fever'. However, major haemorrhage is not seen in the majority of patients, although a degree of bleeding (blood in stool, petechiae, ecchymoses, oozing from venepuncture sites, and mucosal bleeding) can occur. Significant bleeding is seen in the terminal phase of some cases and in pregnancy. Most patients who survive EVD show signs of improvement during the second week of illness.

Investigations

Testing for the presence of *Ebolavirus* is by reverse-transcription polymerase chain reaction (RT-PCR). Testing to exclude other differential diagnoses of febrile illnesses should also be undertaken. This is particularly relevant for patients who are PCR-negative, but should be undertaken where possible in all cases to exclude comorbidity – malaria coinfection, for example, is not uncommon.

Other laboratory findings associated with EVD include leukopenia, thrombocytopenia, serum transaminase elevations (secondary to multifocal hepatic necrosis), decreased serum albumin, elevated amylase, electrolyte disturbances, renal and coagulation abnormalities.

Treatment options

Early diagnosis is essential in order to initiate treatment and institute strict infection control. However, there are no licensed therapies and management is largely supportive. Maintaining fluid balance with electrolyte replenishment, empirical antibiotics and anti-

diarrhoeal agents are of benefit. The use of non-steroidal anti-inflammatory drugs and vitamin K remains contentious. Different protocols exist across different settings and organisations.

Despite there being no licenced treatment options, significant progress was made during the outbreak with 15 experimental therapies trialled on humans[120]. These include blood products e.g. convalescent plasma, immunological agents e.g. monoclonal antibodies, such as ZMapp; antiviral drug therapies, e.g. Favipiravir and novel agents, such as TKM0120803[120]. However, none were proven to have significant benefit for patients. In addition, the efficacy of vaccine candidates was also studied (discussed in Chapter 4 & 5).

Potential prevention strategies

Strategies to prevent transmission of EVD during an outbreak include[121]:

- Infection control (including patient isolation) and sterilization including careful hand hygiene (hand washing with soap and water or an alcohol-based sanitizer), avoidance of contact with blood and body fluids of EVD patients(e.g. urine, faeces, saliva, sweat, urine, vomit, breast milk, semen, and vaginal fluids), and avoidance of all items in contact with infected patients (e.g. bedding, needles/medical equipment)
- Use of personal protective equipment
- Contact tracing and fever surveillance of all contacts. Future prophylactic treatment may be possible if an effective therapeutic agent is identified
- Quarantine
- Treatment if an effective agent is identified to limit infectious period
- Safe burial: avoiding rituals that require washing or handling of the body of an EVD patient
- Safe sex advice
- Engaging communities in preventing transmission
- Vaccination strategies (although no vaccine is currently licenced and this would not prevent random zoonotic emergence events resulting in outbreaks).

2.3.4 Previous Ebola phylogenetic work

Prior to the onset of the West African epidemic, the phylogenetic literature on *Ebolavirus* was very sparse. It was limited to research in a few areas, including:

i) Characterising historical outbreaks and contextualising re-emergence events[122].

ii) Comparison of *Ebolavirus* genomes with other VHF viruses, particularly Marburg virus[123].

iii) Identification of animal vectors and zoonotic emergence confirmation[115,117,124].

However, during the course of this research project, the field of *Ebolavirus* phylogenetics has expanded rapidly with numerous publications. The 2013-16 West African Ebola epidemic was the first emerging infection outbreak in which real-time sequence data were used to understand disease origin and spread. Sequencing was used to characterize the *Ebolavirus* strain responsible, determine its evolutionary rate, identify signatures of host adaption, identify and monitor diagnostic targets, and establish the characterization of the response to vaccines and treatments. However, sequence data can only be used to inform interventions, if the results are generated quickly enough to be useful. These developments are discussed in Chapter 4 and align to the work presented throughout this thesis.

2.4 Conclusions

Having identified the potential benefit of combining traditional and molecular epidemiological techniques to enhance understanding of infectious disease transmission dynamics, my work focuses on the role of each discipline, and how and when to combine the two techniques for maximum utility. As noted previously, combining the two disciplines can help to define the source of infection, onward transmission events, direction of transmission, and infection dynamics. However, both methods may give misleading answers and have their own inherent advantages and limitations. Understanding when and how to use each approach, and when to combine them, is key to this work. Ultimately, the purpose of this thesis is to consider an integrated approach incorporating both traditional and molecular epidemiological methods in order better to understand epidemic transmission dynamics. This enhanced understanding could translate into more effective infectious disease surveillance, monitoring of prevention strategies, and optimising of data to inform public health policy.

CHAPTER 3

DATA SOURCES

Clinical, laboratory and, increasingly, sequence data are often collected routinely in infectious disease outbreaks for medical and surveillance purposes. In this chapter, I explore the data sources used in this thesis for both HIV and Ebola. I describe the study sites, methods of data collection, and types of data used to create the datasets, as well as data management and ethics associated with the different datasets. I then explain the methods used to link the epidemiological, clinical and phylogenetic data to create bespoke datasets used in the work presented in this thesis.

All data presented in this thesis were collected routinely as part of the core work of the Africa Centre or as part of the emergency response to the Ebola epidemic (from various sources). I have spent time in both South Africa and Sierra Leone and have familiarised myself with the environments from which the data were collected, as well as the methods of local data collection. I personally collected field data in Sierra Leone as part of the National Ebola Archive Project (outlined in more detail below), some of which is used in this thesis. However, the rest of the data utilised is obtained from other existing data sources. I recognise it is important to engage in the centres where the data are collected to fully appreciate the key challenges and limitations encountered in both collecting and using the data. My original PhD planned to spend more time in both centres, but this time was cut short due to a serious road traffic accident. Therefore, this work focuses on the analysis, rather than generation, of routinely collected data.

3.1 HIV: Africa Centre

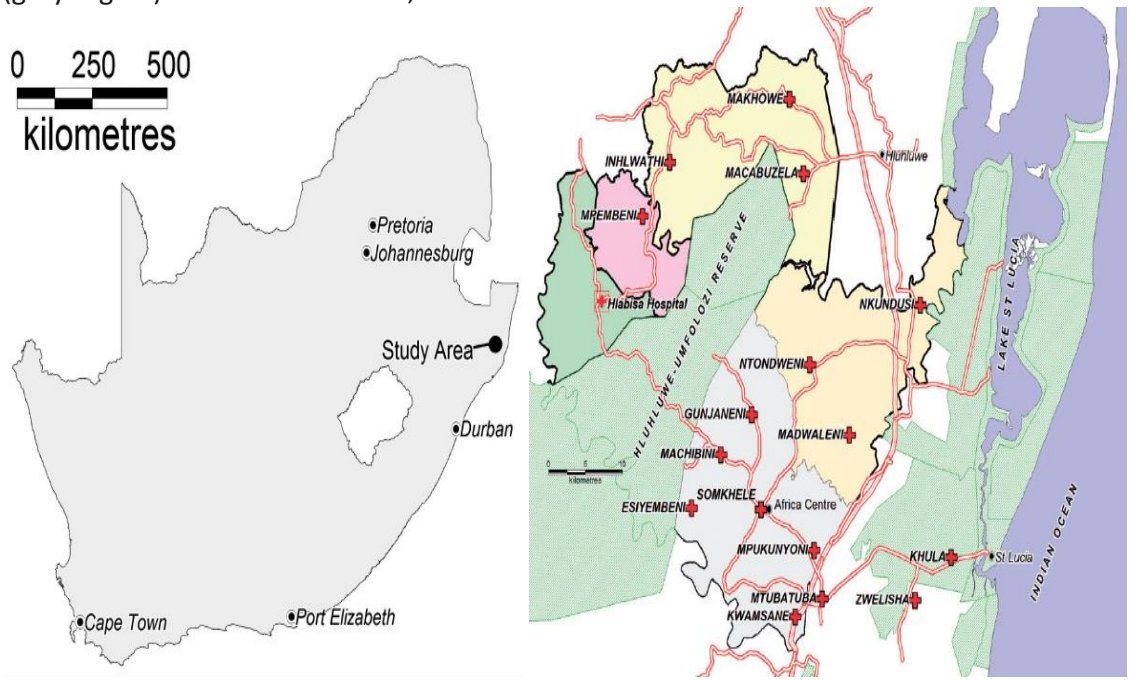
3.1.1 Study site

The Africa Centre for Health and Population studies (AC) was founded in 1997 by the University of KwaZulu-Natal and the South African Medical Research Council. It is one of the five Wellcome Trust Major Overseas Research Centres, from which it receives most of its core funding. It has recently been renamed the Africa Research Health Institute after its merger with the KwaZulu-Natal Research Institute for TB and HIV, but is referred to as AC throughout this thesis.

AC is located in northern KwaZulu-Natal (KZN), in the rural district of uMkhanyakude (Hlabisa sub-district) and near the market town of Mtubatuba[125-127]. The demographic surveillance area (DSA) covers 438km^2 (Figure 3.1A) and includes a mobile population of approximately 94,000 people (60,000 resident and 34,000 non-resident) who make up approximately 12,000 households. The DSA is mostly demarcated by clear geographical boundaries, bordered by Umfolozi Game Reserve national park to the west, Umfolozi River to the south, and the N2 national highway (between Durban and Mozambique) to the east. However, the northern border is less well defined and cuts across the Mpukunyoni tribal area (Figure 3.1B). The area is predominantly rural and consists of tribal land with scattered households, although there is a formal municipal township (KwaMsane Township) and urban/peri-urban areas bordering the main highway (KwaMsane and Indlovu Village), which are all densely populated. The tribal land is divided into areas defined by the traditional Zulu communities there, called Isigodi. The population density varies considerably across the DSA (between 20-3000/km^2). All households have been geographically mapped to within an accuracy of two metres as described below.

The area is characterised by very high levels of HIV infections: over 24% of the adult population are infected[128] and over 40% of antenatal attendees are HIV positive[129]. The HIV prevalence in this cohort peaks at over 50% in females aged 25-29 years, and 44% in men aged 30-34 years[128,130]. However, the HIV prevalence shows considerable geographical variation across the study site, between 6-36%, with evidence of localized clustering of HIV infections[130]. The HIV incidence is approximately 2.7 new infections per 100 person-years (in the adult population >15 years old), peaking at 6.6 per 100 person years in women aged 24 years[131].

Figure 3.1 A&B: Location of study area in South Africa[126,132].
Map A shows the study area in relation to the rest of South Africa. Map B shows the DSA (grey region) with its boundaries, roads and clinics.



Furthermore, this population has very high rates of unemployment (64% overall, approximately 70% in men and 60% in women[133]) and high rates of mobility (35.8% of men and 38% of women)[125,134]. The main sources of income are through employment and government grants, including pensions[135]. An open cast coal mine opened within the DSA in 2007/2008 (the Somkele mine) and this is now the largest employer in the region (employs 989 people). As part of the initial mining approval contract, the mine was required to employ local people. It is reported that 80% of employees of the mine 'live' locally, but as the AC population was not skilled in mining, roles are largely limited to administrative and transport tasks. The AC is the second largest employer in the area, employing approximately 350 people[133]. There is also a culture of temporary labour migration[125,127,134,136,137]. A more detailed exploration of migration within the Africa Centre is undertaken in Chapter 8.

The AC community structure is characterised by low marital rates (23% of men and 31% of women have ever been married), late marriage especially for men, polygamous marriages (approximately 14% of all marriages for men and 12% of all marriages for women) and multiple sexual partnerships[138]. There is also poor knowledge and disclosure of HIV status[139].

3.1.2 Data

a) Demographic, epidemiological, behavioural and clinical data

Overview of data collected

The Africa Centre Demographic Information Systems (ACDIS) commenced in 2000 to collect data on the characteristics of households and individuals living in the demographic surveillance area (DSA). To date, AC primarily uses paper-based data collection methods to undertake the demographic surveillance surveys (DSS), which are input manually into the electronic data platform – ACDIS[140]. All data are recorded in pseudo-anonymised form.

Data collected is divided into household data collection and individual data collection. The head of the household, or a proxy if the head of household is not available, is asked to provide information about the household (e.g. socio-economic factors) and an overview of all members of the household, including vital events such as births, deaths (based on verbal autopsies), marriage, and migrations. The survey rounds initially occurred twice a year, but increased to three times per year in 2005. Since then, the surveys have been standardised.

In addition, an annual interview with each individual aged 15 years and older is undertaken to collect more detailed data on health, health service utilisation, behavioural factors (including sexual behavioural risk factors), and socio-economic factors, including education and employment data. Furthermore, if participants consent, a dried blood spot (DBS) is collected at this interview on which a number of bio-measures are undertaken, including HIV sero-status, HIV viral load, HIV and TB pathogen sequencing of selected samples, and HbA1c. This population-based HIV testing commenced in 2003 and the results are linked anonymously to the other data collected. A recent memorandum of understanding with the South African Department of Health now means this dataset can be augmented with clinical and laboratory data (from the National Reference Laboratory) which will be linked to this population.

Almost all the population speaks Zulu and the data are collected by a team of local field workers who travel around the area to undertake face-to-face interviews with participants.

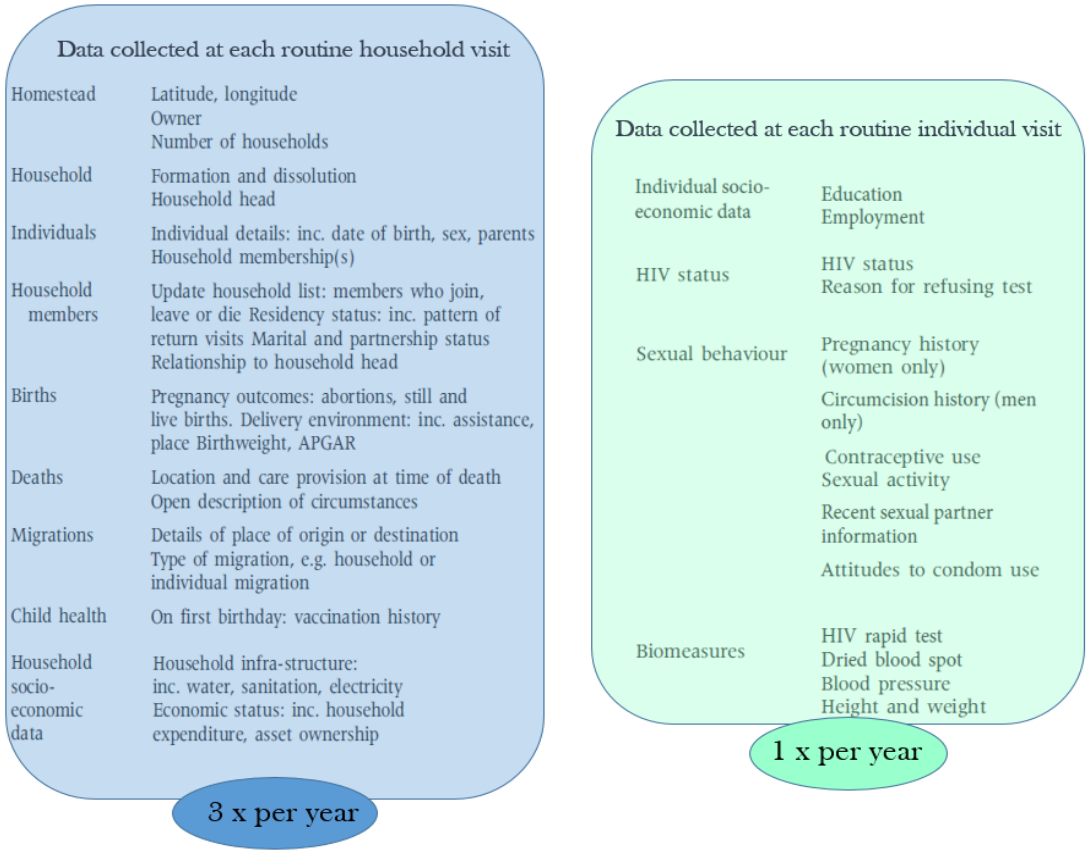
This longitudinal database provides an unrivalled and uniquely rich data source, which is likely to be one of the most complete demographic datasets anywhere in the world. It incorporates demographic, social, medical, economic, and migration data at both a household and individual level. Figure 3.2 shows the information collected with more detailed descriptions of aspects relevant to this thesis. This data source provides the necessary tools necessary to understand fully the drivers of ongoing infection and make significant scientific advancement in this field.

Inclusion criteria for AC study cohort

Individuals are included in the surveillance population if they are reported as being a member of a household in the DSA, irrespective of whether they are resident or not. The place of residence is typically considered to be the home where an individual keeps their daily belongings and spends >4 nights per week. Therefore, an individual can have membership of multiple households, but only one residency at a specific point in time.

Individual-level data, including health and sexual behaviour, is only collected on those >15 years old.

Figure 3.2: Summary of the data collected routinely at the Africa Centre[126]



Information on specific data used in the analyses in this thesis

Socio-economic data:

Household wealth is calculated based on 'The wealth index' model[141,142]. It is a composite measure of a household's cumulative living standard and calculated using easy-to-collect data on a household's ownership of selected assets, such as: televisions and bicycles; materials used for housing construction; and types of water access and sanitation facilities.

Maximum educational level is defined by the number of years completed in primary school or secondary school, or any tertiary education.

Employment status is defined as unemployed, full-time or part-time.

Health and sexual behaviour data:

Information about sexual history and behaviours over the past 12 months are asked face-to-face by fieldworkers recruited from the local community. Respondents are asked about number of sexual partners in the last 12 months and then more detailed follow-up questions about up to three of their most recent sexual partnerships, including age of partner, cohabitation, where the partner lives, condom use, circumcision status of male, regularity of relationship with partner, and concomitant partners. These questions are often not answered by respondents and sexual behaviour data are often sparsely completed in the surveillance. Furthermore, anecdotal reports suggest that participants learn to answer the questions in a way intended to minimise the time the questionnaires take, so it is uncommon to report more than one partner. For this reason, in this thesis I use the category of 'ever having reported more than one sexual partner in a 12 month period', rather than more than one partner in the last 12 months, as a proxy for risky sexual behaviour.

Residency/Migration data:

The residency status (as described above) and place of residence is recorded routinely for all household members at every surveillance visit. During each surveillance round, any change in household residency or membership is recorded, together with information about the origin or destination and the date of the move. Changes in status are referred to as migration events. Migration events are defined further in Chapter 8.

The nomenclature used by AC for the types of dwellings and living structures within the DSA are outlined below:

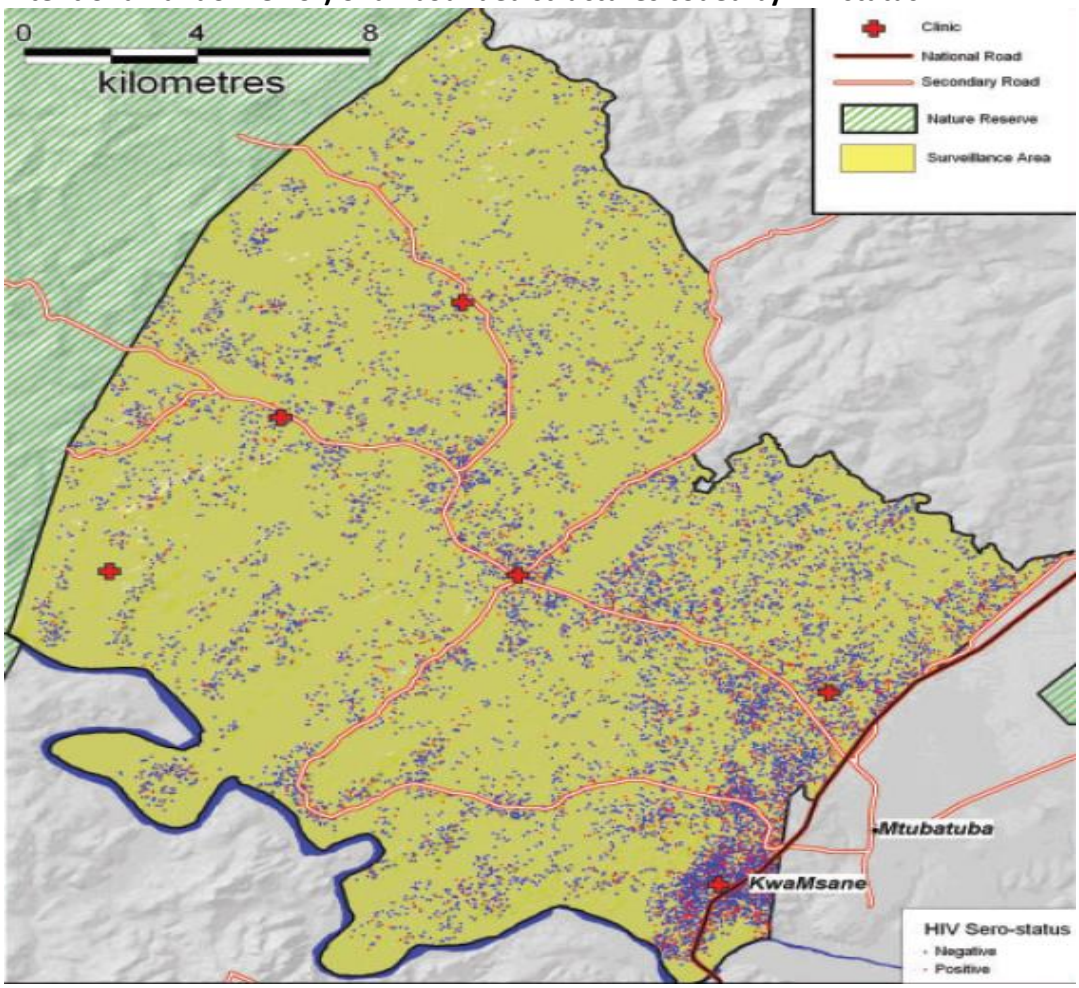
- Bounded structure: It is defined by AC as "a building, or a group of buildings, on land belonging to a single person or organisation, and used for one main purpose"[140]. It includes both residential and non-residential buildings (e.g. schools and clinics). Bounded structures are often delineated by a physical property line, e.g. a fence, or recognizable with open land between neighbouring bounded structures.
- Homestead: This is a form of bounded structure, or a component of a larger bounded structure. It is a physical dwelling space defined as "a grouping of houses or huts on one piece of land, which belongs to a single owner and which is mainly used as a place for people to live"[140]. It is typically one building for 'living' with associated outhouses for sanitation etc.
- Household: AC defines this as "a social group of one or more members [that] share in the joint household resources and know each other well enough to provide information about each other. In each household, one of the members is considered to be the head of household"[140]. This definition means that, in theory, multiple households can be contained within a single homestead. However, homestead and household are frequently used interchangeably by AC. Household membership data between 2004-2010 show that there are 1.04 households per homestead. On average there are between 6.53-7.22 individuals per household and 8.5-9.5 co-members (as individuals are members but not resident in the household)[143].

Typically, one family lives in a homestead (household), which either makes up an entire bounded structure, or else the bounded structure may be comprised of multiple homesteads, for example where extended family members live in the additional homesteads. Each bounded structure is geo-located as described below.

Geographical data

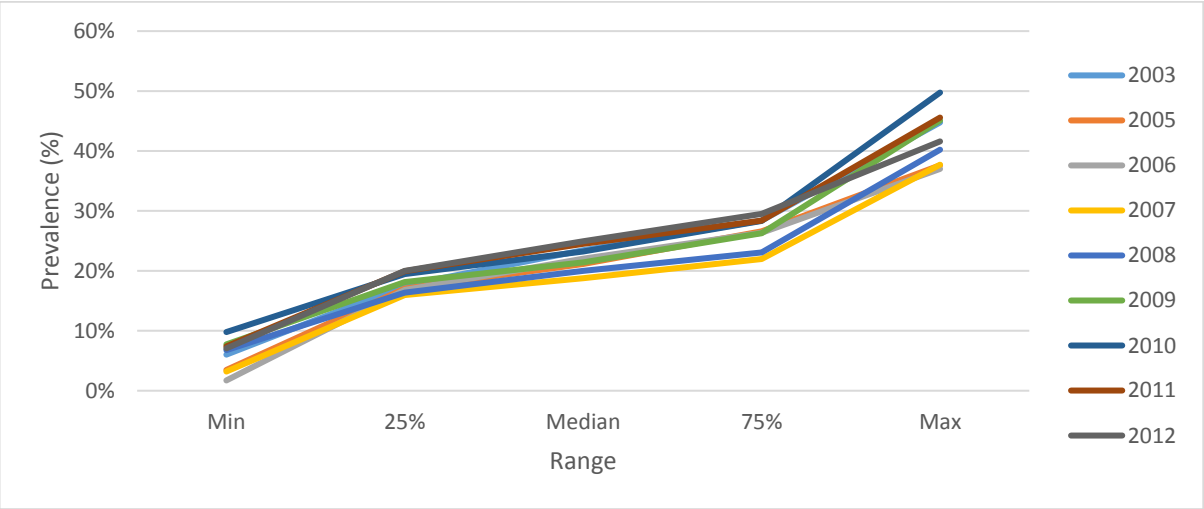
Each individual home dwelling within the AC surveillance area has been mapped using GPS coordinates to an accuracy of <2 meters[130]. This allows all individuals to be geo-located to a precise residence. Figure 3.3 shows a map with all the bounded structures located by HIV serostatus.

Figure 3.3: Map of study area showing the approximate location (incorporating an intentional random error) of all bounded structures coded by HIV status



Frank Tanser and his team at AC have used advanced spatial analytical techniques to characterise micro-geographies of incident and/or prevalent HIV infections within the AC surveillance area[130,131]. This enables the HIV prevalence at the household level to be calculated. This was done by employing a two-dimensional Gaussian kernel (radius 3km) to produce robust estimates of HIV prevalence across continuous geographical space, giving community-level estimates of HIV prevalence[130]. Clusters of infection (either higher or lower number than expected) were identified using a Kulldorff spatial scan statistic[130]. This detects the location of clusters and evaluates their statistical significance. It also adjusts for uneven geographical density of the background population, conditioned on the total number of cases observed. Therefore, for each bounded structure, an associated 'local' prevalence by year (2003-2012) was calculated. Figure 3.4 shows the changes in prevalence from 2003-2012 across all 18,500 bounded structures. I used this to categorize households into high, medium and low prevalence areas by year (determined by the interquartile range thresholds per year).

Figure 3.4: Local HIV prevalence data by households in Africa Centre surveillance area 2003-2012

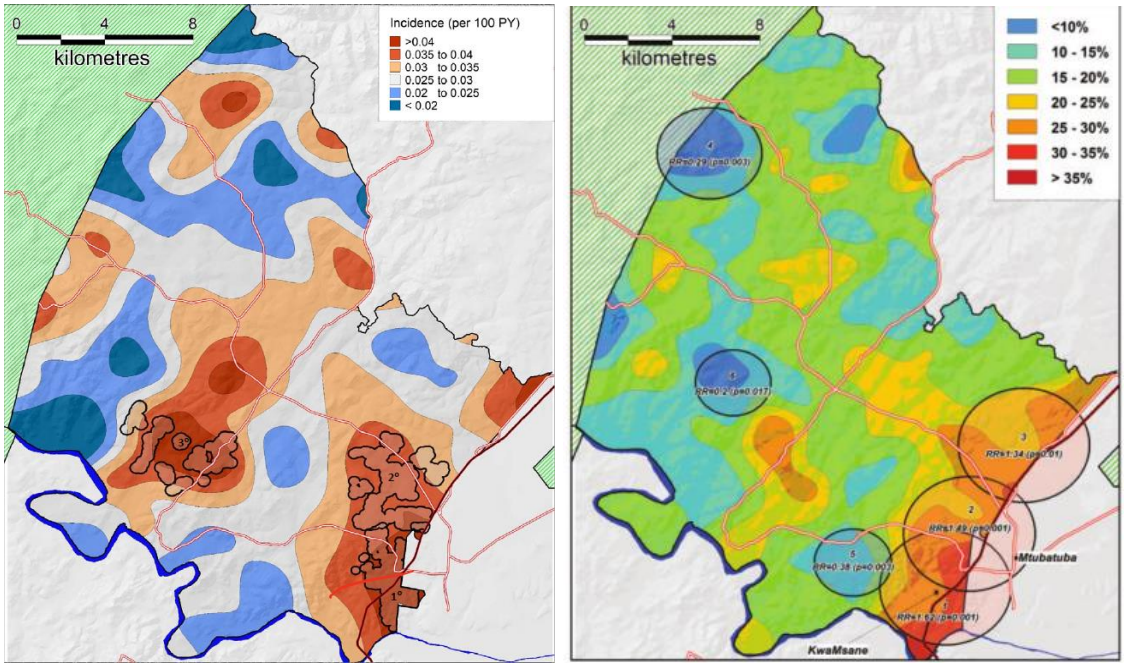


The HIV incidence was calculated using the same Gaussian kernel approach outlined above, which estimates the density of HIV infections per km^2. This allows the derivation of the number of HIV seroconversions per km^2 across the whole study area. Tango's flexibly shaped spatial scan statistic was then applied to detect a local excess of events and test whether such identified excess could have occurred by chance, and also identifies irregularly shaped clusters of high HIV incidence[131].

Figures 3.5 shows maps of AC DSA with HIV incidence (5a) and prevalence (5b) showing geographical clusters of infection. This confirms the high HIV prevalence (6-36%) and incidence (0.86-5.21%) in this area.

Figure 3.5: HIV incidence and prevalence in AC DSA.

A) Incidence map **B) Prevalence map**



Map A shows HIV incidence in the AC DSA (2017) – indicating 3 incident geographical clusters/hotspots (denoted by the three circles with 1, 2 3 in the red/orange zones)[131]. Map B shows HIV prevalence across the DSA categorized in 5% bands (2009)[130].

National electronic clinical dataset

The South African Department of Health together with the AC developed the Hlabisa HIV Treatment and Care Programme in 2004. This was in response to national policy aimed at achieving a three-fold increase in the number of individuals accessing anti-retroviral treatment (ART) between 2007-2012. ART is distributed through 16 primary health care clinics and the local district hospital (Hlabisa). A database collecting information on patients seen at these clinics and the hospital, as well as laboratory results, was developed to support this programme[132]. This database was called ARTemis and includes information on age, sex, contact information, clinic visits, laboratory data (including CD4 and VL) and treatment records. The data from ACDIS can be matched to ARTemis using the individual's unique South African identification number[143], or the AC individual identifier in some cases (IIntId)[143].

B) HIV SEQUENCE DATA

A total of 2,179 HIV-1 subtype C partial *pol* gene sequences sampled between 2000 and 2014 from South Africa were analysed. 1,376 of these sequences were obtained from unique individuals from the Africa Centre database and the other 803 sequences are a convenience sample of all South African HIV-1 subtype C sequences sampled between 2000-2012 available via the publicy available Los Alamos database[144]. The wide sampling frame was required to include an acceptable number of control sequences (e.g. 50% of the AC study sample size). Duplicates were excluded. The South African sequences were included as controls and to contextualise the Africa Centre samples.

The Africa Centre has conducted longitudinal population-based HIV surveillance, including DBS testing, since 2003 and this thesis used all samples collected routinely in the 2010-2012 and 2014 surveillance rounds. Sequence data from 2013 were not available. As described above, the routine surveillance includes a dried blood spot (DBS) sample for all consenting adults over the age of 15 years. A number of bio-measures are performed on these DBS samples, including HIV sero-status testing (SD Bioline HIV 1/2 3.0 ELISA, Yongin, South Korea)[145], HIV viral load (Biocentric HIV-1 Charge Virale assay, Biocentric, France), and HIV sequencing of selected samples[145,146]. Samples were selected for sequencing based on HIV positive sero-status and a viral load >10,000 copies/ml. Viral RNA was extracted, from all DBSs found to be positive by serology, using an automated platform (NucliSens®-easyMAG™, bioMerieux, Marcy-l'Etoile, France). The SATuRN/Life Technologies genotyping system was used for genotyping[145]. This protocol amplifies a 1.3kb region of the *pol* gene. For patients with multiple sequences, the first sequence available was selected for analysis. All sequences were associated with a date of sampling. The AC sequences were linked anonymously to the other demographic, clinical and behavioural metadata, but the South African sequences were usually associated with year of sampling only and no other metadata were available. When the exact date of sampling was unknown, the mid-point of the year (1st July) was taken.

Sequences are 1,907 nucleotides long and span the entire protease (297 nucleotides) and first 1,248 nucleotides of the reverse transcriptase gene of the virus. The sequences were assembled using Geneious 8.0.3 software[147]. Sequence quality was assessed using the HIV-1 Quality Analysis Tool[148] and the Calibrated Population Resistance (CPR) tool[149]. The Quality Analysis Tool tested for contamination and sequence translational problems (such

as frame shift and stop codon mutations) by BLASTing (Basic Local Alignment Search Tool) each sequence against a reference dataset. The CPR tool also looked for quality and performs standardized genotypic estimation of transmitted HIV-1 drug resistance. HIV-1 subtyping was performed using the REGA HIV-1 Subtyping Tool v 3.0[150]. All sequences used in this thesis were already aligned on receipt. However, I have manually checked these alignments and made adjustments where necessary using AliView.

I have spent time in South Africa at AC to understand how the data and samples are collected, the steps involved in data management, sample processing, steps from DBS to sequence generation, sequence processing, quality control and alignment. My work focusses on the analysis of sequence data, but I did not personally undertake the sequencing as part of my research[‡].

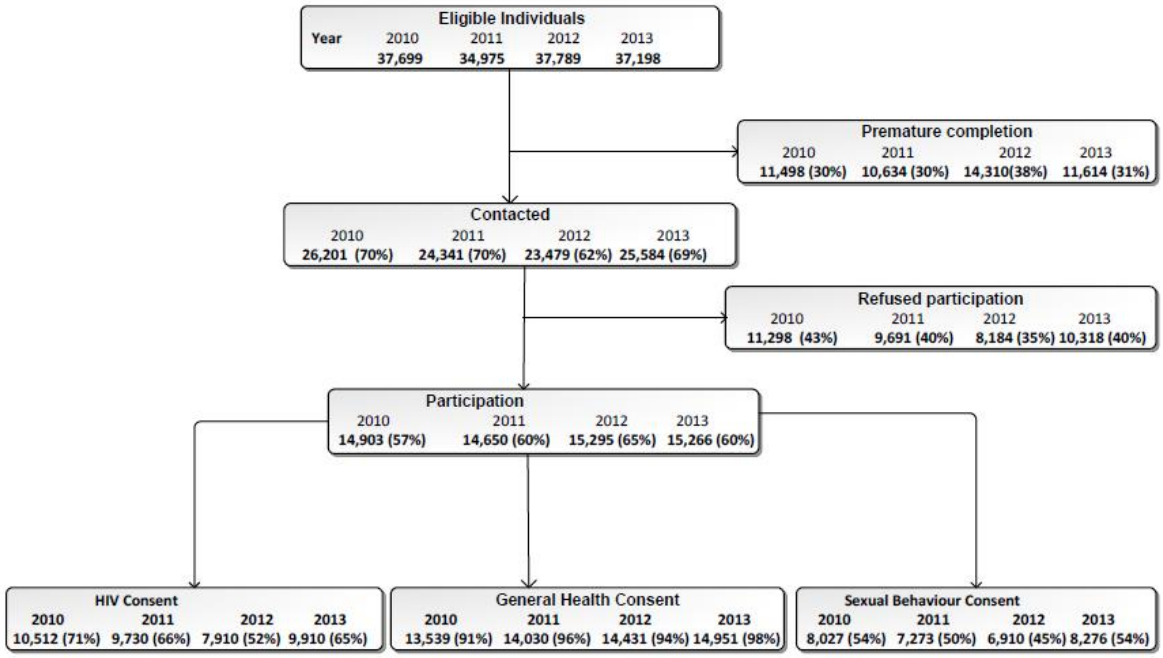
c) Participation in surveillance

Participation rates reflect the demographic dynamics within the population and are affected by many factors such as mortality, ageing, migration and employment, which in turn alter population dynamics from year to year. Refusal to participate leads to an important source of bias in any study. Participation in AC surveillance during this sequencing study (2010-2014) was between 57-65%[151], with a lower consent rate for DBS testing and sexual behavioural data collection. Figure 3.6 outlines the participation and consent rates for the Africa Centre surveillance between 2010 and 2013. It is important to note that even when a participant consents to participate, they may not complete all parts of the questionnaire. Therefore, in my analysis the amount of missing data is often much higher than the overall consent rate.

One of the main criticisms of population-based HIV surveillance programmes is the low annual participation rates. Previous studies have shown that refusal is associated with sex and age: men differentially do not participate, while women consent[152], and there is a differential uptake of health-seeking behaviour (including participation) and anti-retroviral adherence in men and women of older ages and in women more generally (likely as a result of antenatal screening and diagnosis). Although gender specific participation rates

[‡] The data collection and sequencing were undertaken by AC staff as part of the core function of AC. I undertook all analysis from raw sequence data onwards (collaborators contributions detailed where relevant).

Figure 3.6: Participation and consent rates within AC DSA (2010-2013)[151]



Eligible Individuals - resident and age >15.
Premature completion-individuals who could not be found because they had died, out-migrated or were temporarily away at the time of field visit.
Contacted-Individuals the field workers managed to speak to.
Participation- Individuals who accepted to participate in one or more of the following: HIV, sexual behaviour and general health (Figure 3.2). Each variable is not complete by all those who agree to participate, therefore, the amount of missing data is higher.
Consented- Individuals who consented to HIV, sexual behaviour and general health. Each variable is not completed by all those who agree to participate, therefore, the amount of missing data is higher.

are not available from AC, the gender consent discrepancy aligns with anecdotal reports from AC suggesting that men consent less frequently than women. This discrepancy is exaggerated as there are more females than males in the DSA population. Furthermore, evidence suggests that HIV prevalence is substantially higher among refusers[152]. This systematic difference between those who participate and those who do not results in bias. There are many reasons for lower participation rates in men, including they are less likely to consent and are often not present during surveillance due to migration for employment purposes. In addition, they have differential health seeking behaviours, likely due to the stigma associated with HIV diagnosis and clinic attendance, and are not involved in the kind of routine services that are available to women, such as maternity testing. Furthermore, it is reported that many men believe that additional tests are futile if they are already known to be positive, and that they may be put off testing given the knowledge that AC collaborates with the Department of Health. Therefore, HIV prevalence is likely to be underestimated in these settings. It is important to consider these factors when analysing the data to ensure results are interpreted correctly.

Participation in HIV positive population

Table 3.1 outlines the care cascade and laboratory results for HIV positive individuals by residential status in the DSA. Among the 5,746 HIV positive individuals who were resident in 2016 and had linked to care (i.e. have a clinical record on the national clinical database), 4,079 (71%) had participated in the DSS HIV serosurvey at least once and tested positive. 287 (5%) had participated in the DSS serosurvey but tested negative at the last participation. 1,380 (24%) had never participated in a DSS serosurveys. Therefore, 1,667 (18%) of the 9,407 HIV positive DSS residents are identified exclusively through linkage with the electronic national HIV database (not through AC surveillance). In contrast, among the 1,836 HIV positive individuals who were non-resident in 2016 and who had linked to care, 695 (38%) had participated in the DSS HIV serosurvey at least once and tested positive. 260 (14%) had participated in the DSS serosurvey but tested negative at the last participation. 881 (48%) had never participated in a DSS serosurveys. Therefore, 1,141 (35%) of the 3,287 HIV positive individuals who were non-resident in 2016 are only known about through linkage with the electronic national HIV database. These individuals are not included in the analyses in this thesis, which is based exclusively on AC DSS data.

Table 3.1: HIV Care cascade and linkage in AC DSA[153]

	Residents in DSA	Non-residents	Total population
Total (n=)	60,751	31,888	92,629
HIV positive	9,405 (15.5%)	3287 (10.3%)	12,692 (13.7%)
Linked to care (% of positive)			
Ever linked	5,746 (61.1%)	1,836 (55.9%)	
Never linked	3,659 (38.9%)	1,451 (44.1%)	
Currently in care (% of ever linked)	3,304 (57.4%)	828 (45.1%)	
Loss to follow-up (% of ever linked)	2,143 (37.3%)	689 (37.5%)	
Currently virally suppressed (% in care with results)	1,605 (84.7%)	357 (82.8%)	
Last VL >10,000	253 (5.0%)	79 (5.5%)	

Data up to 2016

HIV positive defined as positive in DSS serosurvey or via linked national clinical records. Linkage to care defined as clinical record on national database. Virally suppressed is VL<400 or CD4 >=500 (if no VL measurement available).

D) DATA MANAGEMENT

The Africa Centre has a Research Operations and Data Management team comprised of the deputy director of AC (Dr Kobus Herbst), two database scientists, a data repository manager, and two data cleaning personnel. This team is responsible for the overall design

of data platforms, and maintenance and management of the data repository. A separate team is tasked with the input of data, coding of all missing values, labelling of all data values, quality control and quality assurance of data collection, and input. Furthermore, AC staff are responsible for de-linking all records from identifiable information such as individual names to ensure confidentiality.

The data management group is experienced at managing large and complex multidisciplinary datasets to international regulatory standards. All datasets are stored on a secure central local intranet, which is backed up daily. Only authorized computer users with VPN access and login details are allowed to access the data management system. My unique database was stored on the UCL system and encrypted with password protection for security. Only two people had access to this dataset – me and Miss Anna Tostevin (data manager) who supported my linkage work.

E) DATA SHARING

The bespoke datasets I have created for this study will not be made publicy available and comply fully with the UK Caldicott principles[§] and governance on patient-identifiable data. The datasets will be made available to the AC data management team and future requests for access to the data will be governed by their systems.

For analyses in which the risks associated with identification are high, such as in analyses of transmission clusters within small communities, a second anonymisation was performed on datasets/results released outside of the supervisory team.

F) ETHICS

This project only involved data from cohorts and studies in which ethical approval for linkage of pathogen to clinical and epidemiological data has been obtained. The ethics approval from the Africa Centre data has been approved by UKZN ethics committee. UCL

[§] The 'Caldicott' principles were developed by the NHS to protect patient-identifiable information: (i) justify the purpose(s) of every proposed use/transfer; (ii) don't use it unless it is absolutely necessary; (iii) use the minimum necessary; (iv) access to it should be on a strict need-to-know basis; (v) everyone with access to it should be aware of their responsibilities; and (vi) understand and comply with the law.

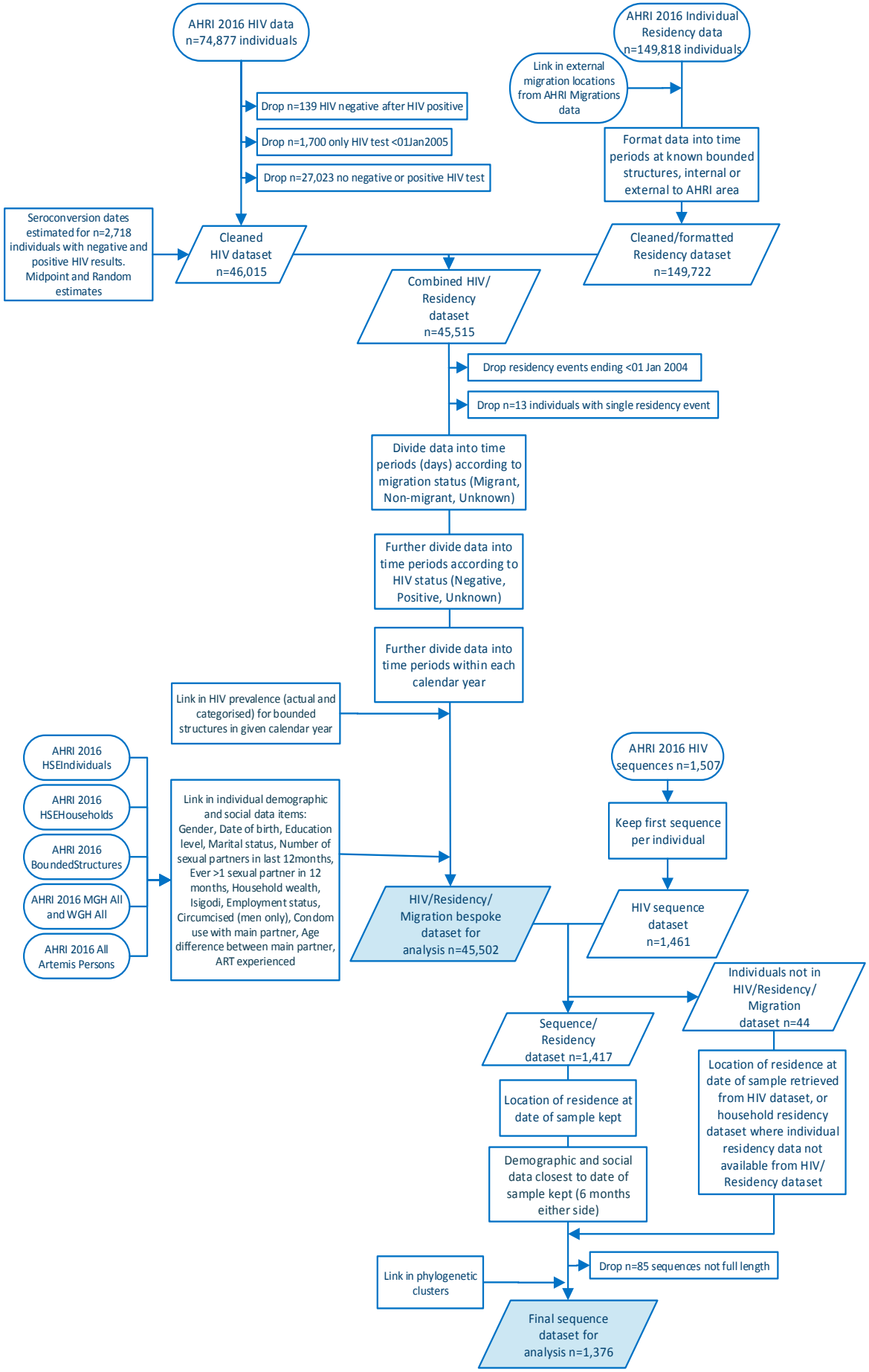
ethics committee has been informed of the project, but as it is only using data with pre-existing ethics approval, no further approval was necessary. I recognize the importance of an appropriate ethical and governance framework for using pseudo-anonymised genetic data and I ensured that this key aspect was incorporated appropriately into my thesis.

3.1.3 Data Linkage

The raw datasets were requested from AC data repository. Linkage across the datasets to create a bespoke longitudinal dataset was undertaken using individual and bounded structure identifiers (IIntId and BSIntId respectively). The raw datasets used and the steps undertaken to produce the dataset for my analysis are summarised in Figure 3.7. Each step of the linkage was reviewed manually, and the linkage was supported by a database manager, Miss Anna Tostevin.

Only direct linkages were included based on the individual identifier (IIntId). My linkage work created a unique pseudo-anonymised dataset of 45,502 individuals with 666,274 observations over 447,338 years of person-time at risk within the AC DSA. All 1,376 AC sequences linked to some metadata (100% linkage), although the fields completed varied depending on response to surveillance surveys.

Figure 3.7: Steps undertaken to obtain HIV dataset for the analyses in this thesis



3.2 Ebola: Sierra Leone

3.2.1 Study site

Sierra Leone is a West African country with a population of 7,075,641 (2015 national census)[154]. It is bordered by Guinea to the north and east, and Liberia to the south. The country is divided into five administrative regions, which are further divided into 14 districts. Freetown is the capital city located in the Western Area. A map of Sierra Leone is shown in Figure 4.2 (Chapter 4).

Sierra Leone became independent from the UK in 1961. Between 1991 and 2002 it endured a civil war which destroyed much of the country's infrastructure, led to 10,000s of deaths, and displaced over one third of the population. Trust in government and authority was significantly eroded during the civil war and has been slow to recover.

The life expectancy in Sierra Leone is 57.8 years[155]. Pre-Ebola there were 136 physicians in the country (0.022 per 1,000 population density). Sierra Leone is the country with the highest number of Ebola cases ever reported.

3.2.2 Data

The focus of this part of the PhD research was on a dataset of 554 *Ebolavirus* sequences. I first describe the collection and processing of these samples. I then describe the available epidemiological and clinical datasets available to allow linkage. The sequence data and epidemiological data generally used different patient identifiers, so the linkage often required several steps and is outlined below.

a) Sequence data

The sample collection and sequence generation were undertaken by my collaborators, Prof. Ian Goodfellow and colleagues from the University of Cambridge and the Sanger Institute. All sequencing was undertaken in Sierra Leone. I was provided with 554 raw sequences associated with a genome identifier and with a sample date. Other identifiers available, but not uniformly present for all sequences, included: a national Ministry of Health Identifier (MOH-ID), gender, laboratory sample ID, date of symptom onset, geographical information, age and gender. Where MOH-ID was available, this enabled

direct linkage to the metadata (if available), but where it was not, multiple steps of linkage were required.

I undertook the phylogenetic analysis of the raw sequence data (supervised by Dr Stéphane Hué). The more sophisticated fuzzy data matching was supported by Miss Anna Tostevin.

The methods of sample collection and sequence generation by Prof. Goodfellow's team are outlined below:

Sample collection

Samples were collected from patients in Ebola isolation and treatment centres in five districts of Sierra Leone: Makeni (Bombali), Port Loko, Kambia, Kerrytown (Western Urban) and as well as from Koinadugu. The samples were residuals from blood samples or buccal swabs taken for Ebola diagnostic confirmation.

Sample preparation and sequencing[156]

Total nucleic acid extracts were prepared from plasma obtained from collected blood samples or buccal swabs. Samples were tested for the presence of EBOV RNA[157] and were considered positive if cycle threshold (Ct) values were <40 (an inversely proportional proxy for viral load). Nucleic acid extracts from EBOV PCR-positive samples were then subjected to reverse transcription/PCR amplification. Amplicon purification and size selection were performed, followed by library quantification by qPCR. Libraries were subsequently sequenced and processed to produce a consensus sequence.

Sequence alignment

The 554 newly generated genomes were aligned to a EBOV Makona genome (the strain responsible for the West African outbreak, 2013-2016) downloaded from the NCBI Ebolavirus Resource.

Ethics

Ethical approval for this project was obtained from the Sierra Leone Ethics and Scientific Review Committee and also the London School of Hygiene and Tropical Medicine ethics board. This included the field data collection as part of the National Data Archive project,

which included patient identifiable information. The UCL ethics committee has been informed of the project, but as the work only uses data with pre-existing ethics approval, no further ethical approval was necessary.

In addition to the ethics approval above, this research is also covered by the ethics approval of two collaborators, each their own ethics approval from Sierra Leone. The ethical declarations for each collaborator state:

- Sequence data (from Cambridge University/ Wellcome Trust Sanger Institute collaborators– Prof. Ian Goodfellow): "The study was conducted in compliance with principles expressed in the Declaration of Helsinki, and ethical approvals for the use of residual diagnostic samples for sequencing were obtained from the Sierra Leone Ethics and Scientific Review Committee and the Ministry of Health of Sierra Leone. The Sierra Leone Ethics and Scientific Review Committee approved the use of diagnostic leftover samples collected by EMLab and corresponding patient data for this study".
- For clinical/outcome data and linkage analysis (from IMC collaborators – Dr Adam Levine): "Ethical approval for this study and exemption from informed consent was provided by the Sierra Leone Ethics and Scientific Review Committee, the University of Liberia – Pacific Institute for Research & Evaluation Institutional Review Board, and the Lifespan (Rhode Island Hospital) Institutional Review Board".

I recognize the importance of an appropriate ethical and governance framework for using pseudo-anonymised genetic data, and it is an important aspect of this study.

b) Epidemiological, clinical data

The following datasets were available and were used to link the sequence data to metadata. I have described each data source in turn, how the data were obtained, a summary of the relevant information obtained from the data, and key identifiers to allow linkage.

i) District level data

These data were collected as part of the National Data Archive project commissioned by the government of Sierra Leone and funded by the Department for International Development (DFID), UK. This project was undertaken as a collaboration between the Ministry of Health and Sanitation in Sierra Leone (MOHS-SL), DFID, WHO, CDC and LSHTM. I led the in-country work and data collection for this project, travelling to Sierra Leone and

around the country to visit all affected districts. The aim of this project was to build a repository of all data and information about the Ebola response in Sierra Leone. The objective for doing this was two-fold: firstly, for archiving purposes to have a historical information resource about the outbreak and the response; and secondly, to allow national level analyses to be undertaken and provide a source of knowledge about the Ebola response for future outbreaks.

Two teams (including me) travelled around the country to District Ebola Response Centres (DERCs - the coordinating centre of the Ebola response), each of which had six core 'pillars' (workstreams) including alerts, contact tracing, case management, quarantining, safe burial, and social mobilization. All available data related to the Ebola response were collected, although the data varied enormously between each district and pillar. The data included information about possible, suspected, and confirmed cases, as well as information concerning implementation of interventions.

The core dataset collected and used in this work is the **'Viral Haemorrhagic Fever' (VHF)** clinical/epidemiological dataset collected at a district level (e.g. Kambia VHF and Port Loko VHF), set up by the CDC at district level, and includes clinical, contact tracing and outcome data on all possible, suspected, and confirmed cases of Ebola. The main patient identifier used in this dataset is MOH-ID. Where this dataset was not available, other datasources were used to optimize the sample matching coverage rate.

ii) IMC data

The International Medical Corps (IMC) set up and ran several Ebola Treatment Centres (ETC) during the Ebola epidemic. As the epidemic ended, IMC collected and collated data from across all ETCs with the aim of forming a unified platform containing all clinical information. I am collaborating with IMC on this work and have access to the database from Sierra Leone.

Data collection

Patient demographic, clinical, and psychosocial (PSS) data were recorded from admission to discharge on standardized paper forms by trained nurses, physician assistants, or physicians. All data were collected as part of routine clinical care and for epidemiologic purposes. Images of patient files were scanned into PDF, JPG, or TIFF formats within the

low risk zone of each ETC. Data from all forms were transferred to separate electronic databases at each ETC and later combined to form a unified database. This database includes ten separate tables encompassing patient demographic, triage, rounding, treatment, laboratory, psychosocial, outcome, and follow up data.

At the conclusion of IMC's ETC programming, patient paper records from the low-risk zones of the ETCs were electronically scanned and stored in IMC's secure network drive and the hard copies were transferred to the MOH. Laboratory data, including EVD RT-PCR test result and Ct values, as well as malaria test results, were obtained directly from laboratories to which patient blood tests were sent, and were linked to patient data in IMC's unified database.

Quality control

Lot Quality Assurance Sampling (LQAS), a random sampling methodology, was used by IMC to assess the quality of data entered into the databases from original patient charts. As per LQAS methodology, a random sample of 19 patient ID numbers from two sub-strata (EVD+ and EVD-) from each ETC were selected for this data quality audit. Due to a high number of discrepancies between data on scans of triage, rounding, and treatment patient charts, as well as data in the unified database, re-entry for these data were performed using the scanned copies of the original charts. Once re-entry was complete, another data quality audit utilizing LQAS was completed. All discrepancies were recorded as an error. The number of errors per patient chart was divided by the total number of data points for the specific patient; this number differed per patient depending on length of stay. The total percentage of errors was then calculated. Results from the audit conclude that approximately 99% of the data in the IMC unified database are consistent with information from scans of patient charts.

iii) WHO data

Due to unforeseen circumstances, the collection of district level data were not completed for four districts. Furthermore, data were misplaced in transit back to the UK by a collaborator on the National Ebola data archive project. This left me with an incomplete dataset. Therefore, I contacted collaborators to see if it was possible to access additional data in order to minimise this gap.

Firstly, the mathematical modelling of infectious diseases department at LSHTM is a collaborating centre with the WHO and provided real-time support during the Ebola outbreak. It was this team with whom I worked in-country on the National Ebola Archive Project. Due to their status as a collaborating centre, they had access to WHO individual level data. I was given access to this data under a collaborating agreement. However, the patient identifiers were coded and I was unable to access the codebook to enable linkage to my dataset. However, in <5% of cases, an additional field including the MOH-ID was completed and this enabled linkage of a few cases.

Furthermore, the Chief Medical Officer in Sierra Leone gave permission for WHO to collaborate with me directly (as WHO remain custodians of the national data) to help link the unlinked sequence cases. Unfortunately, due to linkage problems within WHO, I only gained access to the WHO laboratory dataset. I provided WHO with a list of individual identifiers associated with each unmatched sequence (unmatched after using all data sources above) and they sent me a dataset with information including laboratory ID, genome ID and MOH-ID, outcome at time of sample, CT value, gender, and some geographical information for those they were able to match.

iv) Laboratory datasets

Finally, Prof. Goodfellow had additional laboratory information on some of the sequenced samples (from 'Hastings', 'Nig-EM', 'POW & Kam' laboratories). This included information obtained by the laboratory on the form accompanying the sample. In a few instances, this allowed linkage of the laboratory number to a MOH-ID, which enabled linkage to the epidemiological data.

v) Village X data

Chapter 5 in this thesis focusses on a Village outbreak of Ebola, which included cases in the 554 sequences. Whilst undertaking field work in Sierra Leone, I was asked to collaborate on this work and undertake the analyses, which combined epidemiological, clinical and sequence data from this Village. I did not collect these data, but was sent the raw data to analyse. These data were collected as part of the routine response to the outbreak by local district response teams, supported by WHO and CDC. The work was part of a collaborative enterprise including CDC (epidemiological data), IMC (clinical data and experience) and University of Cambridge (sequence data), and the project was conceived, initiated and approved by the University of Cambridge, University of Rhode Island, and the Sierra Leone

ethics committees prior to my involvement in the project. My contribution to this work was undertaking all analyses, writing the manuscript, and coordinating input from all agencies involved.

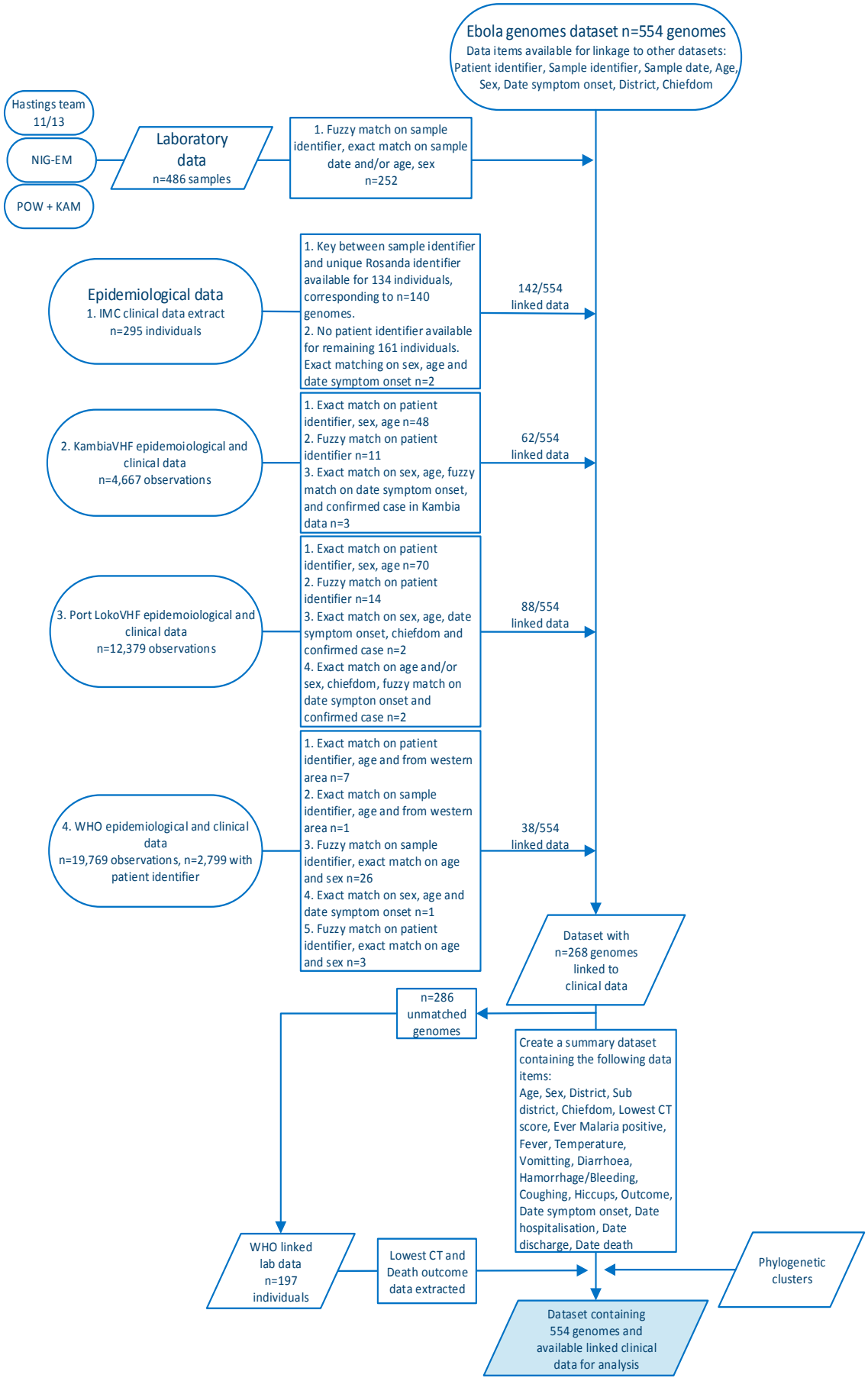
3.2.3 Data Linkage

Metadata were matched to genomes using an algorithm of unique patient identifiers. A summary of the steps taken to maximise the linkage between sequence data and associated metadata are summarised in Figure 3.8.

During the linkage, some genomes linked to multiple data sources, e.g. both IMC clinical data and VHF data. In these cases, the IMC data were used as first line data and took priority over other data sources if variable information did not match. This was on the basis that the IMC dataset was predominantly clinical data, collected and cleaned by teams who cared for the patients, rather than VHF data which involved data entry from paper forms with the potential to induce errors. The IMC data had also passed through quality assurance checks. Therefore, on balance, the IMC data were more likely to be reliable. Any incomplete fields were then supplemented by data from other sources in the following order: VHF regional datasets (Kambia or Port Loko), laboratory datasets, and finally WHO datasets. This order was determined by the order in which the datasets were acquired, and newly acquired datasets were used to supplement and improve current linkage. Where no direct link was available, probabilistic fuzzy matching was undertaken using the *reclink* command in Stata v14 (StataCorp LP, College Station, TX, USA). This involved matching records from two different data files using two or more variables that are not unique to an individual, but in combination might provide a reliable linkage, e.g. age, gender, district and date of onset of symptoms. All linkage steps were reviewed manually.

Unfortunately, due to the data loss described above and the challenges of accessing relevant data, the sequence to metadata linkage coverage was only 52%, and within those that linked many fields were sparsely populated. The missingness of this data meant that this was too low to yield meaningful results in the analysis of the association of viral lineage to outcome or clinical presentation. Nonetheless, it did enable the Village X work to be contextualised more generally in terms of the Ebola epidemic in Sierra Leone.

Figure 3.8: Summary of steps take to link Ebola sequence data to metadata



3.3 Conclusion

There are multiple data sources used in this thesis, including clinical, epidemiological, demographic, behavioural, and sequence data. There are many challenges associated with linking data across multiple data sources. This is particularly evident in linking routinely collected data where the data linkage has not been previously considered and, therefore, not managed in a way to facilitate linkage across different datasets. The primary limitations were poorly recorded patient identifiers and lack of consistency in use of patient identifiers across datasets. This limited the potential of the data to answer many key questions. Going forwards, guidance and systems to promote good data management and archiving of data are needed, especially in outbreak settings. These improved systems would maximise the data able to be reused for scientific advancement. Recently, there have been numerous calls for data to be made more easily accessible, and my work highlights the challenges of accessing data as well as data security issues. These challenges are described further in Chapters 9 and 10.

Chapter 4

THE EBOLA OUTBREAK, 2013-2016

OLD LESSONS FOR NEW EPIDEMICS

This chapter gives a detailed descriptive epidemiological review of the most recent Ebola outbreak in West Africa (2013-2016). A brief review of the global HIV epidemic is provided in Chapter 2, but as the West African Ebola epidemic occurred during my PhD, this work is novel and has been published as a comprehensive overview.

I aim to understand the key factors and disciplines needed in traditional outbreak responses. I summarise the published literature and supplement this with some relevant grey literature to describe the outbreak chronology in West Africa. I discuss the factors that contributed to the rapid and extensive propagation of this outbreak, highlighting the key failures and successes. This outbreak was the first time that molecular epidemiology had been used in the field to understand an acute emerging epidemic. Therefore, I explore its role and potential benefits for future outbreaks. Finally, I highlight the key lessons learnt form the world's largest Ebola outbreak.

This work has been published: Coltart CEM, Lindsey B, Ghinai I, Johnson AM, Heymann DL. The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Phil. Trans. R. Soc. B,* 2017: 372; 20160297; DOI: 10.1098/rstb.2016.0297[160].

4.1 Introduction

Ebola viruses have significant epidemic potential, as shown by the 2013-16 West African outbreak. Caused by the *Zaire* strain, this outbreak was unprecedented in scale, being larger than all other outbreaks combined, with 28,646 reported cases and 11,323 reported deaths[158]. This Ebola outbreak was the first to lead to a major global public health threat and the first in which the virus spread across multiple international boundaries.

Although Ebola is known to cause outbreaks in central and eastern Africa, no sporadic human cases or outbreaks had previously been reported in West Africa. Notwithstanding this, in 1994 there was a major outbreak among non-human primates in Cote d'Ivoire with one veterinarian subsequently infected. Because of this, in the early phase of the West African outbreak, Ebola was not considered as a differential diagnosis, with Lassa fever - another viral haemorrhagic fever that is known to occur in humans in West Africa - being considered a more likely cause.

The widespread nature of the West African outbreak is thought to be related to the highly mobile communities and densely populated regions affected in the early stages of the outbreak. Previous outbreaks had been limited to remote, rural areas allowing initial containment efforts to be more effective. The majority of cases in the West African outbreak were localised to three countries with intense transmission: Sierra Leone, Liberia and Guinea. Seven other countries had minor outbreaks with non-sustained transmission or isolated cases, all with origins attributable to West Africa: Nigeria, Mali, Senegal, Spain, the UK, the USA, and Italy. Other countries also accepted evacuated cases from West Africa for hospitalisation, including Germany, France, Switzerland, the Netherlands and Norway.

This chapter aims to provide a detailed description of the evolution of the outbreak. I outline the events by country, in chronological order, including epidemiological parameters and implementation of outbreak containment strategies. Molecular epidemiology was used for the first time in an acute emerging epidemic during this outbreak. I summarise the contribution of molecular epidemiological studies to understanding of the outbreak and briefly compare the relative roles played by traditional versus molecular epidemiology in understanding transmission dynamics.

Finally, I summarise the factors that led to rapid and extensive propagation, as well as discussing the key lessons learned from this outbreak and from the response. It is vital that the lessons learned from the world's largest Ebola outbreak are not lost.

To contextualise this chapter, Chapter 2 provides the background to Ebola Virus Disease (EVD) including details of previous outbreaks, as well as the epidemiology, clinical characteristics, investigations, treatment and prevention strategies for EVD.

4.2 Outbreak evolution

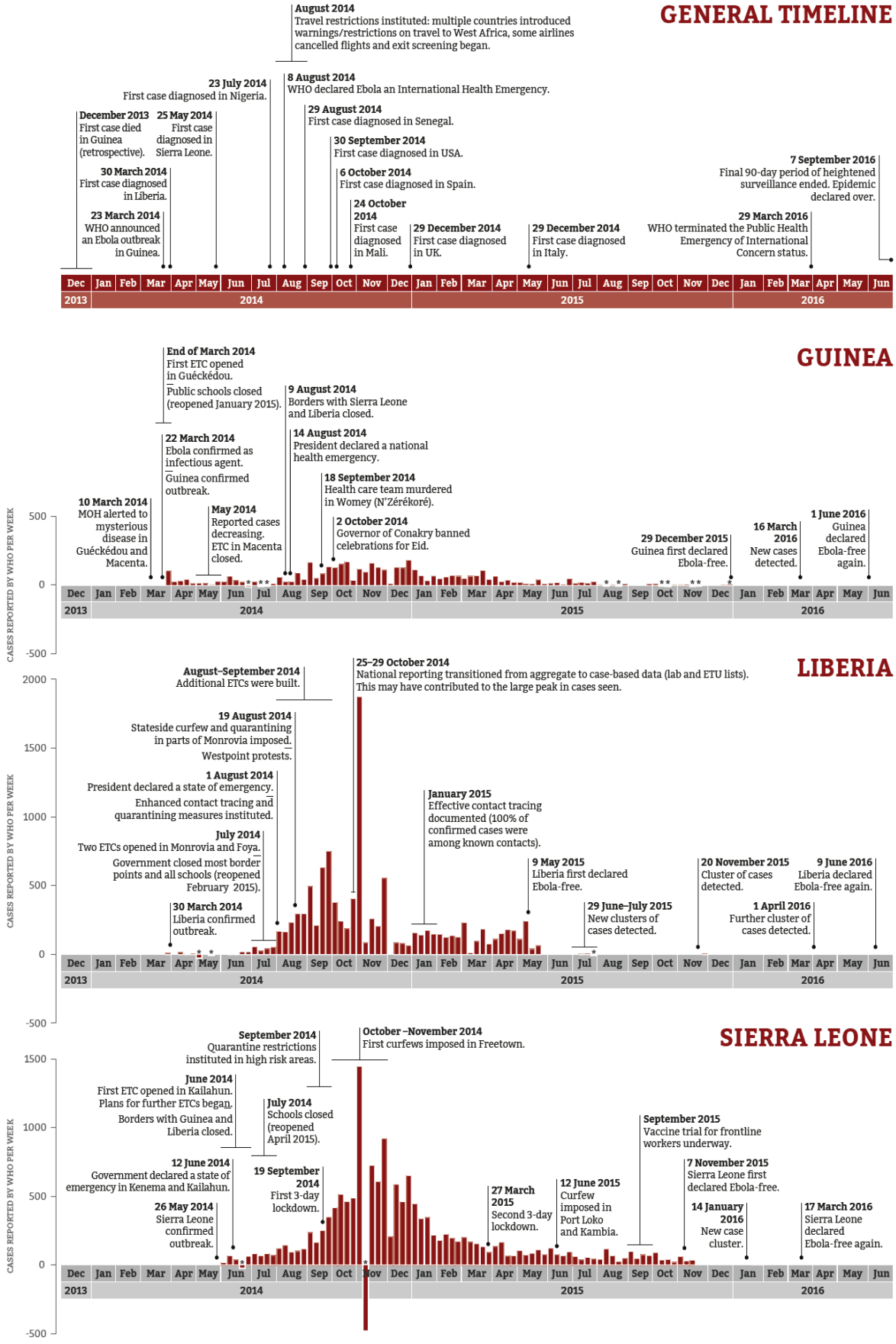
The outbreak began in Guinea, with the first case retrospectively identified as having occurred in late 2013, and spread to several other countries, with Sierra Leone and Liberia most severely affected. The evolution of the 2013-16 West African outbreak is described below in chronological order, by country. Figure 4.1 documents a timeline of the major outbreak events in each of the three main affected countries, including epidemic curves. Figure 4.2 shows a map of the three countries in West Africa with intense transmission, to aid visualisation of the geographical spread. Table 4.1 shows a summary of basic country statistics pre-Ebola. Box 4.1 highlights key outbreak related definitions used throughout this work.

Box 4.1: Outbreak related WHO definitions used during an Ebola outbreak[159]

Suspected case	Any person, alive or dead, suffering or having suffered from a sudden onset of high fever and having had contact with: • A suspected, probable or confirmed Ebola case • A dead or sick animal Any person with sudden onset of high fever and at least three of the following: • Headaches • Anorexia • Stomach pain • Vomiting • Diarrhoea • Lethargy • Myalgia or arthralgia • Difficulty swallowing • Difficulty breathing • Hiccups Any person with inexplicable bleeding or sudden, inexplicable death
Probable case	Any suspected case evaluated by a clinician Any deceased suspected case having an epidemiological link with a confirmed case (where it is not possible to collect samples for laboratory confirmation)
Confirmed case	Laboratory confirmed cases: Any suspected or probable cases with a positive result for virus antigen by detection of virus RNA by reverse transcriptase-polymerase chain reaction (RT-PCR), or IgM antibodies against Ebola

NB The use of 'case'/'case number' refers to the total reported suspected, probable and confirmed cases provided in WHO situation reports

Figure 4.1: Timeline of key events with country-specific epidemic curves[160]



Case numbers are total reported suspected, probable, and confirmed cases provided in WHO situation reports throughout the epidemic. We have calculated weekly case number (Monday-Sunday).
*Negative case numbers are reported (as per WHO data) when suspected/probable cases subsequently test negative for Ebola PCR or as a result of data errors.
Abbreviations used: MOH- Ministry of Health, MSF- Médecins Sans Frontières, ETCs – Ebola Treatment Centres. Where events happened multiple times, only the first occurrence has been shown.

Figure 4.2: Geographical map of Guinea, Sierra Leone and Liberia showing districts and total number of confirmed cases by district (adapted from WHO[161])

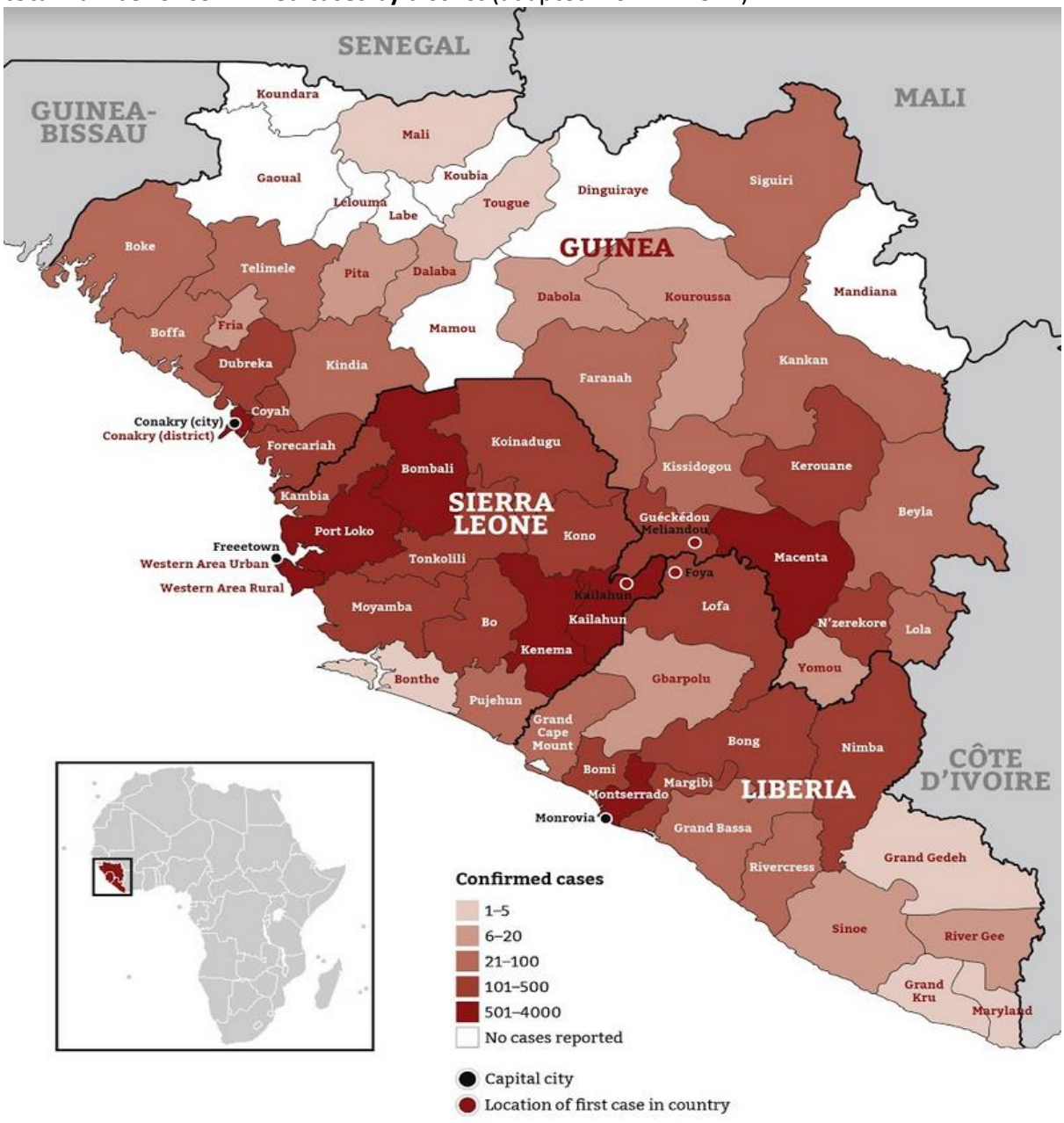


Table 4.1: Basic country statistics from the three main affected countries

Country Statistic	Guinea	Liberia	Sierra Leone
Population	12.3 million	4.4 million	6.3 million
Rural population (% of total)	63.3	50.7	60.4
Gross domestic product per capita (US$)	539.6	457.9	792.6
Capital city	Conakry	Monrovia	Freetown
Physicians per 1000 people (as of 2010)	0.1	0.014	0.022
Total number of reported Ebola cases (WHO-2013-16)	3811	10678	3956
Total number of Ebola deaths (WHO 2013-2016)	2543	4810	14124

Table constructed from World Bank data. As of 2014 unless otherwise stated[162,163].

4.2.1 GUINEA

On 10th March 2014, the Ministry of Health in Guinea was alerted to an outbreak of a mysterious disease characterized by fever, severe diarrhoea, vomiting, and a high fatality rate in the prefectures (regions) of Guéckédou and Macenta in south-east Guinea**. Two days later, Medicins sans Frontières (MSF), which had worked in the region primarily on malaria projects since 2010 was also notified. A team sent by the Ministry of Health reached the outbreak area on 14th March, with a European MSF team arriving on 18th March. Epidemiological investigation and blood samples (sent to the biosafety level 4 laboratories in Lyon, France and Hamburg, Germany) confirmed EVD. The WHO subsequently announced an Ebola outbreak on 23rd March 2014[164-166].

Retrospective traditional epidemiological analyses traced the source of the outbreak to Meliandou village within the prefecture of Guéckédou, in the forested region of south-eastern Guinea (see Figure 4.2). The suspected index case was a 2-year old child who fell ill on 2nd December 2013 and died four days later[167]. A second epidemiological investigation confirmed the source village and index case, but the date of death of the index case was documented as the end of December 2013. Other family members rapidly became unwell and died between 13th December 2013 and 1st January 2014 (mother, sister, grandmother). A village midwife who cared for the index case during his illness also fell ill – during her hospitalisation in the nearest town, Guéckédou, she likely infected another healthcare worker (HCW) who was hospitalised in Macenta hospital and is thought to have triggered the spread of the infection to larger town. The midwife also had epidemiological links to cases in villages around Guéckédou prefecture (Dandou Pombo, Dawa, and Gbandou Villages) between January and March 2014 with 17 reported deaths. The initial case fatality rate was 86% (12 of the 14 original patients with a known outcome died). Baize *et al*. reported this initial transmission chain in October 2014 and an adaptation of the initial transmission tree is shown in Chapter 5 (Figure 5.2)[168].

Between December 2013 and the end of March 2014, when the *Ebolavirus* was identified as the infectious agent, a total of 49 cases had been confirmed, with 111 clinically

**To avoid confusion, it is important to note that all eight prefectures in Guinea have a district and city (the regional capital) with the same name, therefore we have tried to be explicit in stating the prefecture, district, and town level within this paper.

suspected cases and 79 deaths recorded on the basis of clinical symptoms[165,166]. The cases were recorded in the prefectures of Guéckédou, Macenta, and Kissidougou.

The first cases were detected in the capital city, Conakry, in March 2014. However, it was not until May 2014 that sustained transmission in Conakry was documented. Conakry is a city of approximately two million inhabitants located in the West of the country on the coast[165]. Interestingly, while most districts had a peak, or peaks of infection, both Guéckédou and Conakry had fairly consistent levels of transmission spanning the majority of the outbreak period in Guinea.

By May 2014, the number of cases appeared to be declining in the initial epicentre and the MSF treatment facility in Macenta was closed. New cases continued to be reported in other parts of the country, but these were attributed to introductions from Sierra Leone and Liberia. In time, it became apparent that case numbers had not fallen, instead cases had been hidden due to a number of conspiracy theories which arose when patients were taken to Ebola Treatment Centres (ETCs) and died there. In September 2014 one newspaper reported that "many Guineans say local and foreign HCWs are part of a conspiracy which either deliberately introduced the outbreak, or invented it as a means of luring Africans to clinics to harvest their blood and organs"[169]. Furthermore, traditional burial practices were forbidden (initially at a local level, but subsequently at a national level) to reduce exposure to infected bodies (see Box 4.2). However, this violation of cultural beliefs bred fear that the deceased and their relatives would be cursed for failing to perform a proper ceremony. The unfamiliar sight of workers wearing full protective clothing heightened tensions; fear reached a climax when riots broke out in N'Zérékoré after rumours spread that healthcare workers disinfecting a market were in fact contaminating people[170]. In September 2014, eight members of a health care team were tragically murdered by residents of Womey near N'Zérékoré[170-172].

Data during this early phase of the outbreak was irregularly collected and resources were focussed on providing clinical care. Despite multiple requests from MSF and other agencies working in the country, it was not until 8th August 2014 that the WHO declared the outbreak a public health emergency of international concern. The President of Guinea, Alpha Conde, followed suit, declaring a national health emergency. Containment efforts included automatic admission to hospital for suspected cases, compulsory quarantining of

Ebola contacts, travel restrictions including enhanced border controls, and preventing dead bodies being transported between towns (common for cultural repatriation). Contravening these restrictions would be subject to law enforcements[173].

Despite the increased international support (September 2014 onwards), weekly confirmed case numbers remained stubbornly between 75 and 148 from September to December 2014[158]. In early October, MSF reported a spike in cases in the capital, Conakry, with one treatment centre receiving 22 patients in a single day. As a result, the governor of Conakry banned all cultural celebrations for Eid[174,175].

During October, districts which were previously disease-free started reporting cases including Lola, Kankan and Faranah districts[158]. Transmission in Kankan and Faranah districts was of particular concern due to their proximity to national borders; Kankan is adjacent to Côte d'Ivoire and forms part of the major trade route to Mali, while Faranah district neighbours Koinadugu in Sierra Leone, which was beginning to report cases. The potential cross-border transmission highlighted the need for national border surveillance.

On 23rd October 2014, the government of Guinea announced that they had started compensating the families of HCW's who died after becoming infected with a lump sum payment of $10,000 to each family. Until that date 42 cases of HCW deaths had been reported[176].

Intense transmission persisted through November and December. Concerns were raised on 20th November 2014 when the Red Cross sent blood samples to a testing centre via courier taxi. The taxi was robbed near the town of Kissidougou with the robbers unwittingly stealing the cooler bag with the infected blood. Despite public appeals, the samples were never recovered[177], although there was no documented evidence of any resulting transmission from this event.

By mid-December 2014, cases were reported in the northern district of Siguiri bordering Mali[158] and one month later the virus had spread to the western district of Fria for the first time; all 19 districts were reporting transmission events on a weekly basis and transmission in the capital remained high throughout[158].

All countries struggled to provide the necessary bed capacity to isolate and treat all suspected, probable and confirmed cases of Ebola, particularly in the early phases. In total, Guinea had nine ETCs during the outbreak. The early ETCs were run by MSF and French Red Cross, with many other players forming crucial collaborations and providing resources and staff. The largest ETC was the first one in Guéckédou, which had a maximal capacity of 170 beds. Other centres included Macenta, Conakry (Donka Hospital), Coyah, Beyla, N'Zérékoré, Kerouane, Wonkifong, and Kankan[178]. Setting up an ETC was a huge undertaking and took time, meaning that the timings of capacity and demand rarely coincided.

Incidence began to fall in early January 2015 to approximately 50 cases per week, but this fluctuated[158]. Throughout February and March, transmission was concentrated in the western districts including Conakry, Coyah, and Forecariah, which borders Kambia in Sierra Leone – by now in the midst of its own outbreak (see below). These two areas formed 46 % of the weekly new case counts across all three West African countries towards the end of March 2015[158].

During April, May and June 2015, the national weekly case count declined to around 20 cases reported per week, but fluctuated until a decrease at the end of July. Transmission remained concentrated around the Western Districts in Conakry and Forecariah, but with cases re-emerging elsewhere (Dubreka and Boke). Through August Guinea reported just a few cases per week focused around Conakry and Coyah and 13th September 2015 marked the first week with no new cases documented. However, a small number of further cases were intermittently reported in the Forcariah district between mid-September and 29th October[158].

Guinea was first declared Ebola-free on 29th December 2015 after a 42-day period without new cases. However, on 16th March 2016, within the 90-day high level surveillance period, Guinean health officials reported three deaths with symptoms compatible with Ebola in the village of Koropara[179]. Investigators from the Ministry of Health, WHO, the US Centres for Disease Control (CDC) and UNICEF arrived the following day. Samples were taken from four contacts of the deceased, and two tested positive for Ebola (the mother and sister of one of the deceased). Cases were rapidly admitted to a treatment facility, together with rapid mobilisation of epidemiologists, surveillance experts, vaccinators,

social mobilizers, contact tracers and an anthropologist as part of the inter-agency response team. Guinea was again declared Ebola free on 1st June 2016[158]. Box 4.2 outlines the WHO criteria for declaring an outbreak over.

Box 4.2: WHO Criteria for declaring the end of the Ebola outbreak[180]

End of Ebola outbreak: 42-day count	The outbreak is considered ended in a country after 42 days have passed since the last confirmed case has met one of three criteria: 1. Been isolated, recovered and subsequently tested negative for the virus on two blood samples (collected at an interval of at least 48 hours) 2. Been isolated and subsequently died in the ETC with a safe burial organized by the ETC. The 42-day count begins on the day following burial 3. The case was a contact of a confirmed Ebola case. He/she died and was buried in the community and was either a confirmed case or a probable case. The 42-day count begins on the day following burial. The outbreak in the West Africa sub region was declared when the 42-day period elapsed in the last affected country. The rationale for 42-days is based on twice the maximum incubation period for Ebola, as this can be expected to confirm the interruption of human-to-human transmission. During this time, each country should: a) Ensure active case finding around confirmed cases and transmission chains b) Implement both active and passive surveillance for EVD (e.g. through regular health facility visits and by maintaining a nationwide system of alerts) c) Conduct post-mortem testing for EVD following deaths in the community d) Offer testing of semen samples among survivors and, for those who test positive, monthly testing thereafter until 2 negative results are obtained e) Ensure screening of blood donors and products f) Sentinel surveillance among patients with febrile illness
90-day heightened surveillance period	After the 42-day period has elapsed, each country should maintain a system of heightened surveillance for a further 90 days, ensuring ongoing EVD surveillance and notification. The rationale for 90-days of heightened surveillance is due to: 1. The continued risk of new importations of EVD within the West Africa sub-region, given the outbreak occurred in multiple countries 2. The possibility of sexual transmission (current evidence suggest that viable *Ebolavirus* can persist in semen for at least 82 days after symptom onset and possibly longer than 6 months) 3. The possibility of a missed transmission chain 4. The possibility of a new emergency from an animal reservoir During this time a combination of active and passive surveillance should be maintained, integrated with surveillance for other important epidemic-prone diseases. Post mortem testing and testing of survivor semen samples should also be continued for 90 days. Passive surveillance should then be continued indefinitely and EVD preparedness plans should be in place and monitored in all countries previously affected by EVD.

Despite ultimately having the lowest confirmed number of cases and deaths of the three West African countries with major outbreaks, Guinea witnessed 2,543 Ebola deaths among 3,811 confirmed cases[158]. Of the 34 districts in Guinea, 24 districts were affected by Ebola, in contrast to both Liberia and Sierra Leone where every district had reported cases[161].

4.2.2 LIBERIA

The first cases of EVD were confirmed in Liberia on the 30th March 2014, eight days after the outbreak was reported in Guinea. Two cases were confirmed in Foya district, Lofa County, close to the border with Guinea and Sierra Leone (Figure 4.2)[181] and RNA sequencing confirmed the virus was imported from neighbouring Guinea[106].

The Liberian government responded to the outbreak by forming a high-level National Task Force composed of the WHO, UNICEF and multiple international NGOs[††]. Shortly after the formation of this Task Force the CDC sent teams to assist the response[165]. The initial response included enhanced surveillance, contact tracing, training of medical staff, community awareness campaigns and supplying personal protective equipment to health facilities[181].

Within two weeks, five other counties had reported suspected cases; Margibi County, Bong, Nimba Grand Cape Mount, and Monteserrado[165]. Within the first month of the outbreak in Liberia 13 cases were confirmed, of which 11 died[182]. The initial wave of infection was effectively contained within Liberia with no new cases reported for nine weeks between the 6th April and the 7th June[165]. Phylogenetic analysis supports this, with no future samples of *Ebolavirus* identified from the lineage responsible for the first wave of the outbreak[183].

Believing the outbreak to be "relatively small"[184], the WHO and CDC began to withdraw from Liberia[8], despite ongoing transmission in neighbouring Guinea and calls from MSF that more support was required due to the unprecedented geographical spread of the outbreak[185]. This withdrawal later formed part of the wider criticism of the international response to the outbreak[8].

A new laboratory confirmed case of EVD was reported on the 7th June 2014 in Foya district, sparking the beginning of the second wave of transmission in Liberia[165]. Evidence from both contact tracing and phylogenetic analysis suggests that the virus was reintroduced from Sierra Leone and was a distinct viral lineage to previous cases in Liberia[183]. This was the earliest contribution of phylogenetic analysis to the outbreak investigation. The

[††] The International Red Cross, Samaritan's Purse, Pentecostal Mission Unlimited-Liberia, CHF-WASH Liberia, PLAN-Liberia, UN FPA

outbreak soon spread to the capital city, Monrovia – where a quarter of the country's population live – and claimed the lives of a nurse and the head surgeon from Redemption Hospital, a reminder of the significant risk posed to HCWs during Ebola outbreaks[186,187]. By the end of June 2014, 107 cases of EVD were reported in Liberia, including 52 laboratory confirmed and 65 deaths, occurring in Lofa, Montserrado and Margibi. Genetic sequencing suggests that this second distinct viral lineage (which rapidly spread to Monrovia and Margibi County) was responsible for future intra-country spread, rather than the initial viral lineage (from Foya in March 2014)[183].

In July 2014, two ETCs opened in Monrovia and Foya, run by the charity Samaritan's Purse. Capacity was minimal and centres filled quickly, the centre in Monrovia having only 40 beds[185]. Within a month, two American volunteers were infected with Ebola – Samaritan's Purse subsequently suspended activities in the country and evacuated its staff, with MSF stepping in to manage the ETCs[185]. July also saw the Liberian Government close most border points and all schools in order to attempt to minimise transmission[188]. Despite this, by the end of July, case numbers had tripled and EVD had spread to 7 of Liberia's 15 counties (Lofa, Montserrado, Margibi, Bomi, Bong, Nimba and Grand Gedeh); there were 329 suspected cases and 156 deaths[165].

Transmission increased rapidly throughout August and September 2014 (total cases 1082 and 3458 respectively) and all but two counties had reported cases. Additional ETCs – run by MSF, Save the Children, International Medical Corps and the Liberian Ministry of Health – were built but struggled to cope with the high volume of cases[189]. The 120-bed 'Island Clinic' in Monrovia, especially built by the Liberian Ministry of Health, reached capacity within 24 hours[190]. The total bed capacity for the whole of Liberia was 314 by the beginning of September with an estimated deficit of 760 in Monrovia alone[158].

October 2014 saw the largest number of new cases in a month, with 3077 cases suspected or confirmed, nearly doubling previous case counts[182]. All 15 counties had now reported at least one confirmed case, with Montserrado, Magibi, Bong and Nimba worst affected, and transmission decreasing in Foya[158].

In November 2014, the United States government proposed to send $2.89 billion and deploy 3000 troops to West Africa, with their response focused on Liberia[191,192]. This

resulted in the building of new ETCs, an increase in laboratory capacity, air transport of supplies and awareness programs. EVD transmission decreased significantly before many of the ETCs had become operational, with some ETCs treating no patients[192].

Mistrust between communities and authorities was a common theme among countries affected by this West African outbreak. This was epitomised by protests in mid-August 2014. Residents of the West Point District tried to dismantle an Ebola Screening Unit, which they viewed as a risk to their safety. This lead to violent clashes between soldiers and the protesters and the eventual quarantine of the whole West Point District[193]. Mistrust of the government, as well as fear from stigmatisation, led some to avoid seeking medical help for suspected EVD, and reluctance to engage in surveillance and contact tracing[189]. In a heavy-handed response, the Liberian Government made it illegal to conceal an Ebola infected patient, with a prison sentence of two years[194].

Community interventions played a significant role in curtailing the outbreak, but were not initially emphasised by the international teams. Examples of successful interventions included communities buying megaphones to counter myths related to infection and to encourage people to seek treatment instead of hiding from authorities[192]. Liberia is a country with strong community connectedness, with trust for community leaders far greater than that of government[195]. Working alongside this community network proved essential for effective outbreak response.

The peak in reported cases occurred in late September 2014, but by late 2014, transmission began to decrease. By early January 2015, nine months from the first reported cases in Liberia, approximately 150 new cases were being reported per week, and transmission was limited to two counties, Montserrado and Grand Cape Mount[165]. Effective contact tracing and monitoring were now well established, and all registered contacts were being monitored daily, and 100% of confirmed cases were occurring among known contacts[158].

The first week of March 2015 was the first time no new cases of EVD were reported in Liberia. One additional case was confirmed in Monrovia later in March, with no additional infections. Forty two days later on the 9th of May 2015, Liberia was declared Ebola free[158].

Liberia was declared 'Ebola free' three more times, after small clusters of infection in June 2015 and November 2015, and April 2016. The first of these clusters was of six confirmed cases near Monrovia, and phylogenetic analysis suggested re-emergence from an EVD survivor in Liberia rather than cross-border spread [158]. The second cluster occurred in Monrovia among three members of the same family. This was also attributed to long-term viral carriage in a survivor. The final cluster occurred in April 2016 and was thought to have been imported from Guinea when a woman travelled to Monrovia from Macenta, Guinea to visit relatives after the death of her husband from EVD. The virus spread to her two sons but with rapid diagnosis, early treatment and isolation in an ETC and contact tracing the virus did not spread further[196]. Liberia was again declared Ebola free on the 9th June 2016. The total number of suspect and confirmed cases in Liberia was 10,678, with 4,810 deaths[182].

4.2.3 Sierra Leone

EVD was first confirmed in Sierra Leone on May 25th 2014[197]. The first case was a young woman in Kenema, Sierra Leone's third largest city, 50 km from the Liberian border and 100km from the border with Guinea. Given the situation in neighbouring countries, Sierra Leone had already begun an enhanced surveillance programme based in the Lassa fever isolation ward in Kenema General Hospital[105]. Within a month of the outbreak being confirmed in Sierra Leone, over 150 people were reported infected[158], #0} and case numbers appeared to be increasing explosively. The government declared a state of emergency in the 'eastern hub' of Kenema and neighbouring Kailahun on 12th June 2014 and WHO reinforced its representation in the country from mid-June 2014[198]. Within six months of the first reported case, the outbreak in Sierra Leone peaked (November 2014) with up to 150 people a week being infected[182].

The steep gradient of the early part of the epidemic curve hints at a "slow and silent", undetected early phase to the outbreak in Sierra Leone[199]. Retrospective analyses suggests that the EVD was introduced to Sierra Leone from Guinea more than five months before the first officially reported case. This analysis identified a female who travelled from Meliandou, Guinea to Sierra Leone and subsequently died from EVD in January 2014[199]. Further genomic analysis suggests two distinct linages of *Ebolavirus* were introduced into Sierra Leone from Guinea in early 2014[105] and retrospective testing of archived clinical

samples during the 2013-16 outbreak revealed EVD was causing clinical syndromes in Sierra Leone as early as 2006[200].

MSF opened the first ETC in Sierra Leone in Kailahun in mid-June with diagnostic support from Public Health Canada. Like many ETCs at this stage of the outbreak, this facility was rapidly overwhelmed. Further beds were provided in the Lassa fever isolation ward at Kenema District Hospital, which was uniquely placed to deal with the emerging threat of EBV. However, this too was overwhelmed by the sheer volume of cases and it was forced to move patients into general medical wards where isolation and infection control were inadequate. More than 40 HCWs from this hospital were infected in 2014[201], with many of them dying, including Sierra Leone's only national expert on haemorrhagic fevers[202].

By mid-2014, the UK Government took an active role in supporting the National Ebola Response in Sierra Leone. Together with NGOs, such as Save the Children, they provided four ETCs with 700 beds in major urban centres, including one specifically for HCWs with EVD, which was led by the British Military[203].

Once the virus was established within Sierra Leone, molecular epidemiological evidence suggests that sustained human-to-human transmission occurred within the country, rather than through repeated cross-border reinfections or recurrent zoonotic events[204]. EVD appears to have spread long distances following major road networks, while many smaller chains of transmission went unnoticed and uncontrolled in remote, isolated villages[205]. Just as the outbreak in one area was thought to be coming under control, these undetected links surfaced in new geographical areas, causing wave-like spread across the country from east to west.

By September 2014, sustained transmission was reported in the densely populated capital, Freetown[206]. The spread from the eastern hub to Freetown, which resulted in intense transmission, marked a serious escalation in the outbreak[207]. In response, a three day national lockdown was imposed (19th -21st September 2014), designed to give HCWs time to identify new cases, to decrease the movement of people, and to increase awareness of EVD through door-to-door campaigns[208]. Subsequently, quarantine restrictions were put in place in high risk areas with a curfew imposed in Freetown. During the curfews, which

lasted anywhere from 21-days to several months, movements were restricted between 6pm and 6am daily.

The government of Sierra Leone, out of "a desperate need to step up [their] response"[145], began a range of other measures aimed at containing the outbreak: a state of emergency was declared at various times across different regions; schools and other public places (including restaurants) were closed; screening at land borders was strengthened[199]; and all large gatherings including sporting events were cancelled[209].

Mass quarantine proved controversial – at one point one third of the population of Sierra Leone was under quarantine[210]. The president acknowledged the situation as "difficult"[211], though this significantly underplays the food shortages that affected some of the quarantined areas and forced people to break quarantine[212]. Aside from the ethical concerns, many felt that mass quarantine measures were ineffective for Ebola as patients are not infectious until they become symptomatic[213], and they may have been counterproductive by preventing the free movement of necessary medical supplies and personnel[214].

By October 2014, cases were reported from the northern district of Koinadugu, the last remaining Ebola free district of Sierra Leone[215]. However, by November 2014 the outbreak in the eastern hub began to see reduced transmission[207].

More HCWs were infected and died in Sierra Leone than in any other country, both in absolute numbers and relative (proportion of cases) terms[201]. Sierra Leone was the only country in which there were strikes by front-line HCWs because of working conditions and pay. In late 2014, burial workers went on strike over unsafe conditions and a lack of hazard pay[193], in contrast to the relatively generous compensation paid to HCWs and their families in Guinea[176]. Then doctors and nurses withdrew their labour[199] seeking assurance that a new UK-built treatment centre for HCWs (in Kerrytown) would accept local, as well as international staff if they became infected[216].

Despite an apparently well-functioning contact tracing system[217], cases continued in the north of the country across the first half of 2015, in part fuelled by secret and unsafe burials[218]. In June 2015, Operation Northern Push was launched by the government of

Sierra Leone, in collaboration with international partners, to eliminate Ebola from Port Loko and Kambia districts, then hotspots of transmission[219]. This included the imposition of curfews in both districts (12th June 2015), enhanced surveillance, active contact tracing, intense community engagement, and mass quarantine. With these intense response strategies, the outbreak in Sierra Leone appeared to be coming under control, but despite the strictness of Operation Northern Push[220], new chains of transmission proved stubbornly resistant to detection. News of a case in Kambia in September 2015 – three weeks after the last reported case and with no link to any known chain of transmission – led to a flurry of activity, including the first trial of ring vaccination in Sierra Leone for EVD[221]. This was in addition to further vaccine trials for frontline workers undertaken in Sierra Leone[222].

After this concerted effort, one year on from the epidemic peak, in November 2015, Sierra Leone was declared free of EVD when two incubation periods had passed since the last positive patient had tested negative for the second time[223]. However, two months later in January 2016 a 22-year-old woman died of EVD and her carer was subsequently found to be infected. Applying many of the lessons from the previous 18 months, the public health system responded rapidly and effectively, quickly containing the flare-up and preventing spread. Four months later, and almost two years after the first confirmed infection, Sierra Leone was again declared free of EVD on March 17th 2016[224].

The outbreak in Sierra Leone claimed the lives of 3,956 persons and is believed to have infected 14,124 (8,706 laboratory confirmed)[165]. Sierra Leone is now the country with the largest number of Ebola cases in history.

4.2.4 Other countries affected

Several other neighbouring countries had confirmed Ebola infection during this period, and they had epidemiological and phylogenetically proven links to the outbreaks in Guinea, Liberia and Sierra Leone. Only Nigeria and Mali had foci of local transmission. There was also importation of infection in to several European countries and the USA, linked to the West African outbreaks, and in the USA and Spain isolated local transmission also occurred to HCWs. Table 4.2 highlights the Ebola infections diagnosed outside Guinea, Liberia and Sierra Leone by country. There was also the coincidental, but unrelated, outbreak in the Democratic Republic of Congo (DRC), which occurred at the same time.

Table 4.2: Cases diagnosed outside of West Africa related to this outbreak

	No. of cases	No. of deaths	Dates of outbreak	Details
Nigeria[225]	20	8	23/07/2014 to 19/10/2014	Index case travelled by air from Liberia. Local transmission to 19 people (12 first generation cases, 3 waves of transmission).
Mali[226-229]	8	6	24/10/2014 to 18/1/2015	Two importations: 1. October 2014: 2 year old girl from Guinea whose father was a Red Cross worker who died – no local transmission. 2. November 2014: An Iman from Guinea, thought to have partaken in traditional burial ritual ceremonies across the border in Sierra Leone. Local transmission occurred and six others infected. Phylogenetic analysis suggests these were two separate introductions from Guinea: one in October and one in November 2014[229].
USA[230]	4	1	30/9/2014 to 21/12/2014	Two episodes: 1. Liberian national visiting family in Dallas. Local transmission to two HCWs. 2. HCW returned from Guinea. No local transmission.
Spain[231]	1	0	6/10/2014 To 2/12/2014	Nurse caring for a repatriated HCW NB HCW diagnosed in Sierra Leone – local transmission occurred to nurse in Madrid. A second HCW was repatriated from Liberia, but no local transmission occurred. Therefore, total cases cared for in Spain = 3.
UK[232]	1	0	29/12/2014 to 10/3/2015	HCW returned from Sierra Leone (multiple hospital admissions for Ebola) – no local transmission. NB Two other cases were repatriated from Sierra Leone after diagnosis and cared for in UK – no local transmission. Total country case cared for in UK: 3
Senegal[233]	1	0	29/8/2014 to 17/10/2014	Traveller from Guinea – no local transmission.
Italy[234]	1	0	12/5/2015 to 20/7/2015	HCW returned from Sierra Leone – no local transmission.

NB Other countries also accepted evacuated cases from West Africa for hospitalisation, including Germany, France, Switzerland, Netherlands and Norway.

4.3 Outbreak: propagation and failure to control

In the 2013-16 outbreak, social, biological and structural drivers of transmission combined to allow a perfect storm resulting in an unprecedented outbreak with devastating consequences[235]. This was exacerbated by a failure in the response at both national and international levels. Table 4.3 outlines the key factors leading to the failure in controlling this outbreak. This section discusses these factors in more detail, with specific focus on the role of interventions on limiting outbreak size.

Table 4.3: Factors leading to failure to control the outbreak

Factors	
Population structure/ Geography	Mobile populations Rural to urban migration affecting densely populated areas Zoonotic emergence event at the intersection of three countries and near a road network Porous national borders (see Figure 4.4) Multi-country spread
Economic factors / Lack of infrastructure	Fragile states following recent civil wars Lack of governmental trust following historical corruption Weak health systems Road networks along which infection spreads Poor transportation networks Poor telecommunication networks International air links Lack of vehicles to access remote sites
Cultural and Behavioural factors	Traditional burial rituals Dependence on traditional healers Secret societies Community resistance, fuelled by lack of trust and disregard of cultural sensitivities at times Conspiracy theories e.g. hiding cases Civil disobedience
Interventions / Failure in response	Delayed identification Delayed and poorly coordinated international response Weak governance and lack of local accountability within national/local response Lack of evidence on effectiveness of interventions Lack of experience in managing an outbreak on this scale Lack of communication Shortages of HCWs Healthcare associated spread augmenting outbreak Initial lack of community engagement and public information

Of note, multiple estimates of the basic reproductive number, R_0, were carried out. The basic reproductive number is defined as the number of secondary cases that arise from a primary case in a completely susceptible population. The outbreak will only begin to

recede once R_0<1. In Guinea, estimates ranged from 1.5 to 1.71. In Liberia, estimates ranged from 1.36 and 1.83. In Sierra Leonne estimates ranged more widely from 1.4 to 2.53[236].

4.3.1 Population structure/Geography

Emerging diseases such as Ebola often arise from close animal contact at the zoonotic interface. It is therefore common for outbreaks to occur in isolated rural areas, and most previous Ebola outbreaks have remained contained in these settings. In the initial phases of this outbreak, disease transmission went undetected and likely led to chains of transmission within the Kissi tribal area which spans the borders of Guinea, Liberia, and Sierra Leone (Figure 4.3). The population within the Kissi tribal area is mobile across these three countries, the borders of which are porous, and this population accounts for the vast majority of early cases within each of the three countries.

Figure 4.3: Where the outbreak began - map to show Kissi tribal area spanning Guinea, Sierra Leone, and Liberia

Failure to control transmission in the early phases of the outbreak allowed mobile populations and migration to spread transmission chains from rural to urban areas. Recent studies estimate that population mobility in the major affected countries is seven times higher than elsewhere in the world, thought to be due to poverty driving mobility as people look for work or food[237]. The increasing connectivity of distant rural communities[219] means that outbreaks of emerging diseases are more likely than ever to

reach densely-populated centres [238,239], such as Freetown, Monrovia and Conakry. These major cities provide hubs for international spread[240-242]; the world is increasingly globalised and infections do not respect national borders[242].

4.3.2 Lack of Infrastructure

Guinea, Liberia, and Sierra Leone are among the poorest countries in the world, and have only recently emerged from civil wars. Their damaged health infrastructure was ill-equipped to deal with the scale of this outbreak. Pre-outbreak, HCW capacity was already critically low at approximately one or two HCWs per 100,000 population (Table 4.1), and this was further diminished by the epidemic[163,237] .

The nonspecific nature of initial clinical presentation of EVD means fast and accurate laboratory diagnosis is essential, yet in rural West Africa both laboratory and human resource capacity is limited. Therefore, timely response mechanisms were initially hindered by lack of diagnostic facilities. Furthermore, road systems, transportation and telecommunications networks were weak in all three countries, especially in rural settings. This delayed the transportation of patients, diagnostic samples and public information campaigns[237]. For example, in some settings, clinical samples had to be transported across large geographical areas, with poor transport infrastructure, and diagnostic confirmation often took several days.

4.3.3 Cultural factors

Burial practices

High-risk behaviours and lack of infection control measures around death and traditional burial practices have long been known to propagate transmission events[243](Box 4.2). During August 2014, 60% of new infections in Guinea were linked to funeral practices, with 80% linked to traditional burials in Sierra Leone in November 2014[237]. One funeral alone is thought to have begun a huge chain of transmission with several hundred infections[244].

Safe burials were, therefore, integral to the Ebola response – modelling based on the outbreak in Liberia, suggested that interrupting funeral transmission could have had the greatest impact of all interventions on outbreak prevention[245]. Despite this, repeated assessments revealed widespread risks in funerals, including a lack of trained

burial teams, a shortage of burial space, no clear guidelines on collecting diagnostic specimens from the deceased, and a lack of community engagement[246].

Funerals in Eastern Sierra Leone, for example, are steeped in cultural significance, but provide fertile ground for Ebola transmission; the strong sense of family and local allegiances, often tied together by marriages and dowries, are thought to have allowed these remote communities to survive through civil war when government safety nets were non-existent. The practice of taking wives from distant villages saw sisters and close female relatives travelling long distances in order to wash the bodies of women who died from Ebola according to the Muslim tradition (Box 4.3), providing an important social pathway that facilitated the spread of Ebola to new geographical areas[243]. This was compounded by widowers travelling back to their wives' home villages to complete any outstanding dowry payments though labour, providing further opportunity for onward disease transmission[243].

New WHO standard operating procedures (Box 4.3) were insufficient alone to successfully reduce risk behaviours during funerals; it was only when these social pathways were recognised, acknowledged and addressed that the number of safe and dignified burials met international guidelines and the epidemic curve began to fall[247].

4.3.4 Interventions

Prior to this outbreak, the mainstay of interventions to combat Ebola outbreaks were contact tracing and follow up for exposed contacts, prompt treatment and isolation of suspected and confirmed cases, strict infection control and safe burial, underpinned by a strong commitment to community engagement[248]. These measures continued to be effective during the 2013-16 outbreak, and the epidemic confirmed knowledge and protocols established from previous outbreaks.

However, progress in understanding the exact benefit of each intervention is limited due to lack of evidence; the decline in cases across all countries coincided with simultaneous implementation of multiple interventions and disentangling the role of each requires further study. For example, many interventions were part of an 'improved package' in Ebola treatment, where treatment beds and improved community-based infection control were implemented in tandem[249].

Box 4.3: Key facts about traditional burial practices – why do we need safe burials?

The purpose of the burial rituals are three-fold:

1. To honour the dead relatives in the traditional way
2. To say good-bye
3. To accommodate the deeply held beliefs about the obligations of the living to the dying and dead, respecting the cultural view of life after death.

Many tribal ceremony rituals are closely held secrets, therefore not well documented.
One of the main rituals common to all groups is the washing of (with bare hands) and spending time with the dead body, which is highly infectious in the case of Ebola. This is one of the main aspects leading to the enhanced transmission of infection and, thus, the need for safe burial practices.

Burial ceremony traditions depend on tribe and religion:

- Christians close the eyelids of the dead, wash and dress them.
- Muslims wash the dead as well, but wrap them in a white cloth.

NB The populations affected by the Ebola outbreak consisted mainly of both Christians and Muslims.

Special circumstances:

- Tribal leaders: additional rituals are undertaken to transfer powers to a successor
- Pregnancy: many cultural groups feel that the fetus needs to be removed from the mother's body before burial as it disturbs the world's natural cycles
- Some tribal rituals involve animal sacrifice, or inspecting the dead body to determine if the deceased had been a witch – if so, the spirits must be rendered innocuous before burial.

If not undertaken properly, there are consequences for both the deceased and the living relatives, for example the dead are thought to wander the earth eternally and plague the community if they don't reach the village of the dead, which is facilitated by the burial rituals.

WHO protocol for the management of a safe and dignified burial includes:

1. Always take into account cultural and religious concerns and obtain family consent in burial plans
2. Only trained personnel should handle remains during the outbreak
3. Use Personal Protective Equipment (PPE)
4. Place the body in the body bag
5. Place the body bag in a coffin where culturally appropriate
6. Sanitize family's environment
7. Remove PPE, manage waste and perform hand hygiene
8. Transport the coffin or the body bag to the cemetery
9. Burial at the cemetery : place coffin or body bag into the grave
10. Engaging community for prayers as this dissipates tensions

One of the main challenges in implementing safe burial practices are finding culturally acceptable methods in accordance with safety procedures. In order to do this collaboration was needed between political, health, tribal, and religious leaders. Examples of adaptations to burial practices used in the Ebola outbreak include:

- Guidance given by religious leaders not to wash the corpse and praying for the deceased in absentia was sufficient – this was believed to have occurred in historical outbreaks documented in the scriptures
- Burial workers would try to honour reasonable requests from families of the dead e.g.
    1. As they were dressed in protective suits they could dress the dead in outfits chosen by family before the corpses were placed in body bags
    2. Money and jewellery or other sentimental items were allowed to be placed in body bags with the corpse as a 'toll', for the deceased must pay to cross over to the village of the dead
    3. Brief prayers were allowed for loved ones, either while standing two meters (6.5 feet) from the white body bag before removal for burial, or at the grave site after burial.

Adapted from WHO and National Geographic[2,3]

a) Community engagement

Central to all interventions was the need to work with affected communities to serve effectively their needs[250]. As described throughout this paper, a failure to engage communities effectively had a detrimental effect on the outbreak response. Mistrust between the organisations and individuals directly impacted on the effectiveness of surveillance, contact tracing, healthcare seeking behaviour and safe burial initiatives, all individually propagating the spread of the virus[251,252] . Conversely, where achieved, community engagement was integral to controlling Ebola; modelling of the outbreak in Liberia concluded that the increase in bed capacity, credited with causing a reduction in disease incidence, was insufficient to bring case numbers down without significant public engagement[253].

One report identified methods of achieving community engagement for Ebola by drawing on experiences from a wide range of stake holders, including individuals and affected communities, local experts, and international actors. The key steps included identifying both male and female community leaders to champion key messages; organising regular community meetings (to both understand concerns and educate); utilising varied communication methods; tailoring global policies to local settings; and involving family members in care actions which did not expose them to increased risk[254]. For instance, WHO published new protocols for safe and dignified burials in October 2014, designed in conjunction with affected communities, which recognised the need for the family of deceased Muslims to wash and shroud the body and suggested alterations to the standard ritual washing to minimise risk, such as dry ablutions performed in personal protective equipment and shrouding the body in white body bags rather than the traditional white cotton[3].

Effective community engagement benefited policy making: strategies designed to incorporate cultural values, customs and concerns of affected communities were more effective[254]. Implementation also benefited from community engagement; at various stages of this outbreak, transmission was fuelled by a reluctance of populations to seek care in designated facilities, to engage in adequate contact tracing, to respect quarantine regulations, or to reveal deaths in order to allow safe burial[237] – all of which improved after programmes of community engagement[255].

All other policies outlined below must therefore be viewed not as isolated technical interventions, but as part of a wider programme of disease control activities, with community engagement chief amongst these.

b) Contact tracing

The ability to identify, and subsequently interrupt, chains of transmission is crucial to the success of containment efforts. The success of contact tracing is, like all Ebola interventions, determined by the extent to which communities trust those attempting to curb the outbreak. For example, in Nigeria, early recognition and implementation of public health measures averted spread of infection. However, this was at the cost of the life of the clinician who identified EVD[225,256].

One proxy indicator of the success of current contact tracing is, therefore, the proportion of new confirmed cases who were already being monitored as contacts of known, existing Ebola cases. These contacts can then be followed up with temperature checks for 21 days (the usual incubation period) and receive prompt isolation if diagnosed with Ebola, thereby improving treatment outcomes and reducing exposure to further potential contacts. In fact, mathematical modelling from early in the outbreak in Sierra Leone and Liberia identified contact tracing (with prompt isolation and infection control) would have a more substantive effect on the epidemic than even potentially curative medical therapies[257]. Despite this, evidence from Guinea and Sierra Leone suggests that contact tracing was far from adequate – as a result very few new cases were from identified contacts and chains of transmission proved stubbornly difficult to interrupt[252,258,259].

In contrast, the rapid public health response to the various “tail end” epidemics across the region involved effective contact tracing. A high proportion of secondary cases were identified and followed up as contacts prior to their diagnosis with EVD. Had this been achieved earlier in the epidemic, a substantial number of cases might have been averted, as was seen in Nigeria.

c) Infection prevention and control measures

Effective infection control in healthcare facilities is a crucial step in interrupting chains of transmission. Lack of knowledge and resources to provide effective infection control were major factors leading to an amplification of this outbreak. Transmission was propagated by

HCWs who became infected and inadvertently spread infection to their family members and communities, particularly in the early phase of an outbreak, in a similar pattern to historic outbreaks[100,168,260]. Early in the outbreak, the relative risk for acquiring EVD was around 100 times higher for HCWs compared to the general population[261], although this risk decreased as barrier precautions were more effectively implemented, and personal protective equipment became available[262]. This is a tragic reminder of the risks frontline HCWs face in weakened and understaffed healthcare systems.

In an effort to strengthen infection prevention and control (IPC) practices, a partnership between Ministries of Health, CDC, and WHO was established in August 2014 to improve IPC at non-ETC health-care facilities and to decrease the risk for Ebola transmission to HCWs. A National Infection Prevention and Control Plan was developed and published in each country by late 2014, and a national IPC task force established to coordinate infection control efforts. However, progress was slow. In October 2014, almost five months after the first reported case, and despite infection control teams being deployed in northern Sierra Leone, major gaps in IPC practices remained. None of the six Sierra Leonean districts visited by a CDC-led monitoring team had standard operating procedures or adequate equipment[263]. Progress was hindered by delays in importing personal protective equipment and a lack of engagement with community partners[254]. Thus, effective implementation on the ground was impossible despite extensive training[264].

d) Ebola Treatment Centres (ETCs) and bed capacity

The lack of bed capacity in ETCs resulted in a massive inpouring of support to increase treatment facilities. Offering patients care with stringent infection control, supported by accurate PCR diagnosis, increases willingness to be hospitalised and facilitates contract tracing. ETCs have been integral in the control of previous outbreaks, though never before have they been deployed on this scale[265]. However, the role of increasing bed capacity in controlling the outbreak remains controversial. Modelling from the CDC, in September 2014, indicated that 70% of all Ebola cases would need treatment in ETCs to begin to curb the outbreak[266], necessitating a huge scale up in the provision of ETCs from the international community[267]. MSF had been operating several of the early ETCs, and several other players (e.g. Save the Children) began to take an active role in providing more ETCs across all three countries.

The UN Mission for Ebola Emergency Response target of two ETC beds per suspected or confirmed case was eventually met in January 2015[268], but by this time the peak of the outbreak had passed and case numbers were already on the decline in most areas across the three countries. Despite this, several reports suggest that this intervention was perhaps the most crucial in controlling the outbreak[249,269], while others, in contrast, suggest that it occurred too late in the outbreak evolution to have significant impact[249,270]. Part of this discrepancy may be explained by the assumption in certain mathematical models that admission to ETCs led to near perfect isolation of infected individuals[271]. Furthermore, constructing an ETC and training the relevant staff is a time-consuming process, at a crucial time when a single week can make a big difference to the disease burden. Therefore, future outbreak response should begin with construction of required ETCs early to maximise impact.

e) Quarantine

Some interventions though, appeared to be ineffective and sometimes even counterproductive. Epidemiological modelling provided little evidence to support the use of mass quarantines[272] by incorporating the evidence that when patients are asymptomatic they are not infectious[214]. For example, modelling after the Liberian Westpoint quarantine suggested it had little effect in bringing down the reproductive number[245]. Additionally, the quarantine of HCWs from high-income countries, as established by New York and New Jersey, was done with an aim to reduce risk, but its major impact appears to have been to dissuade other volunteers from travelling to West Africa[214].

4.3.5 Intervention summary

Given the limited resources, efforts were rightly focused on outbreak response and clinical care, rather than scientific advancement and research. As a result, we continue to lack sufficient data to inform a robust, evidence-based approach to control Ebola outbreaks and to determine how, when, and in what order to deploy interventions. Prospectively collected data on outbreak response are sparse, preventing rigorous evaluation of the impact of the response on both incidence and relative merits of each intervention[249]. The assessment of specific interventions was further limited by the contemporaneous introduction of many of the major strategies. It is likely that no

single intervention was responsible for reversing the epidemic curve[270], but collectively the traditional control measures instituted appeared ultimately to help.

The limited public health infrastructures within the affected countries contributed to the failure to implement traditional public health measures early enough to deal with the scale of the outbreak. The shortcomings of these responses were compounded by the lack of cultural and contextual knowledge among the high-level international response organisations. It appears that one of the key factors in controlling the outbreak was the successful engagement of communities to help overcome cultural barriers and to pass on educational messages about the importance of infection reporting, contact tracing, and safe burial. Once communities understood that they were able to contribute to the response, successful implementation of all interventions increased.

Most of the progress in knowledge regarding successful interventions occurred via a process of elimination, i.e. by excluding interventions that did not play a crucial role in curbing case numbers based on timing of implementation (ETCs, vaccines and therapeutics). The proven interventions – exhaustive contact tracing, strict infection control measures (during burials and in healthcare settings) and community engagement- are not novel strategies. All were established during the response to the first Ebola outbreak of 1976 and they remain the mainstay of Ebola control today[250,253,254,273,274]. These traditional public health measures remain critical and need to be communicated and reinforced early on during future epidemics, especially in resource-poor settings. The response to this outbreak failed not in the development or deployment of novel technologies, but in the prompt recognition by local and global organisations of the scale of the outbreak. Box 4.4 suggests ten components of an effective Ebola Response based on our research for this article. They are in no particular order, and incorporate both practical solutions for which some evidence exists, and recommendations from expert bodies.

Box 4.4: Top 10 components of an effective Ebola Response

1. Early identification and recognition of outbreak
2. Effective collaboration and coordination between national and international players, with sound governance
3. Quick mobilization of professional and community resources
4. Improved communication and awareness
5. Improved community engagement
6. Training of HCWs in infection control
7. Organisation of contact tracing and isolation
8. Good surveillance and case detection (including molecular epidemiological methods)
9. Safe burial practices
10. Consideration of vaccination strategies based on the latest evidence

4.3.6 Failure of response

There has been much speculation and criticism regarding the timing and delay of both the national and international responses, with WHO in particular singled out for criticism[8]. This crisis exposed organisational failings in the functioning of WHO leading to a "significant and unjustifiable delay occurring in the declaration of a Public Health Emergency of International Concern"[9]. WHO has already accepted the need for transformation of organisational culture and delivery, but this will not be enough to prevent and mitigate future outbreaks – Member States must also transform to ensure full political support for implementing the core capacities for public health outlined in the International Health Regulations (see below).

The scale of the outbreak was underestimated by experts and minimized by authorities. Although some key international players were quick to respond, the early phase of the outbreak was marred by resources being quickly overwhelmed and the inability of local public health infrastructure to cope with the rapidly amplifying case load. Guinea, Liberia, and Sierra Leone had a shortage of public health surveillance capacities to detect, report, and respond rapidly to the outbreak. The response was hampered by a lack of trained and experienced personnel willing to be deployed to West Africa, inadequate financial resources, a limited understanding of effective response methods, ineffective community engagement, and poor coordination[9]. Furthermore, many suspect that the lack of good governance and accountability at a national/local level was a key factor in the genesis of the perfect storm that was the 2013-16 Ebola outbreak.

4.4 Further questions

In the writing of this paper many crucial questions were identified which remain the unexamined part of the story. These include: how might one account for the different epidemiological curves, geographic distributions, and mortality rates across the three countries?; why did the outbreak not spread more widely across porous borders to other neighbouring countries?; what is the role of protective immunity in controlling the outbreak?; is there any evidence for host genetic susceptibility to infection?; would the outbreak have died out without any intervention? This paper summarises what is known and hence we have not speculated on any of these matters. However, we would like to highlight how much more work there is to complete before the outbreak can fully be understood.

4.5 Key successes

Overall, progress in our knowledge of EVD transmission dynamics, control measures, vaccines and therapeutics has been mixed. The limited research that did occur during the outbreak was focused on novel solutions, e.g. vaccines and therapeutics, rather than fully understanding traditional public health measures or clinical management (e.g. fluid and electrolyte balance). Areas of significant advancement include: establishing a pipeline for clinical therapeutic trials in emergency settings; vaccine and therapeutics development; the incorporation of near real-time molecular techniques into transmission dynamic studies; and the use of modelling to assist in decision-making.

4.5.1 Areas of scientific progress include

i) **Development of a process, or "pipeline", for rapid approval and implementation of clinical trials in emergency settings**: This was one of the greatest advancements brought about by this outbreak– namely in relation to vaccines and therapeutics. The expedited ethical and regulatory approvals meant that patients could potentially benefit from the latest treatments without any meaningful delay. Although these vaccines/treatments lack the robust evidence base that would normally be required and hence are accompanied by potential risks, their use is justified in situations with significant associated mortality and limited opportunities to study the disease. This provides the possibility of being able to provide potentially life-saving treatments to some patients and to improve knowledge of disease management for future outbreaks.

During this outbreak, the clinical trials occurred too late to have any significant impact, although hold promise for the future.

ii) **Vaccine development:** To date, no vaccines have been approved for clinical use in humans. However, there are several promising vaccine candidates, from both non-human primate trials and as a result of the first human clinical trials of Ebola vaccine which occurred during this outbreak. Candidate vaccines include adenovirus, vesicular stomatitis, parainfluenza vectors[275-277]. One vaccine, in particular- a vesicular stomatitis virus vaccine, has been shown to be highly effective (70-100% efficacy) and safe[277] and will undoubtedly be used as a key preventative strategy in the future. Vaccination strategies for Ebola outbreaks are discussed further in Chapter 5.

iii) **The use of cutting-edge molecular techniques**: This was used in two main ways. Firstly, rapid polymerase chain reaction (PCR) was made available for the prompt and accurate diagnosis of suspected cases, facilitating their early isolation. Secondly, this outbreak was one of the first instances of near real-time sequencing being used in an acute setting to enhance outbreak understanding. For example:

   a. Real-time genetic sequencing helped to identify specific clusters that could then be linked to risk behaviours. The large outbreak related to the funeral in Kenema, for example, was mapped after sequencing the virus in infected patients, and allowed future responses to be targeted to high risk 'super-spreading' events e.g. funeral ceremonies[244]. Only when chains of transmission were identified, using a combination of conventional traditional field epidemiology and, later, cutting-edge, near real-time molecular techniques, were those infectious networks interrupted. Therefore, novel technologies can provide additional benefit when used in tandem with traditional techniques to supplement established strategies.
   b. Genomic surveillance also helped to describe the outbreak by facilitating an understanding of the origins of the outbreak and subsequent emergent clusters. Further, phylogenetic trees constructed during and after the outbreak demonstrated that disease transmission was almost exclusively

within country, human-to-human transmission, rather than repeated zoonotic events[105].

c. Genetic sequencing also established that the Ebola outbreak in the Democratic Republic of Congo, in 2014, was a result of a separate, unrelated zoonotic event[113]. This highlights that there is a role for incorporating viral genetic sequence data into the standard outbreak response toolkit, as it allows epidemiologists responding to the outbreak to focus their resources most effectively to interrupt transmission.

Table 4.4 highlights the key West African Ebola outbreak phylogenetic studies arising during the course of this research. It highlights the utility of phylogenetic techniques in outbreak investigation.

Table 4.4: Phylogenetic publications and developments during the West African Ebola outbreak

Aim	Key findings
Confirming the emergence of a new EBOV strain Baize *et al*.[167]	Confirmation of Zaire *Ebolavirus* as the causative agent The outbreak was due to a new clade, previously unknown, arising from strains of EBOV from DRC and Gabon. Suggesting parallel evolution All cases link back to the index case Therefore, confirming the emergence of a new EBOV strain
Historical origins and zoonotic introductions Gire *et al*.[105]	The West African variant diverged from a central African lineage around 2004 The virus crossed over from Guinea to Sierra Leone in May 2014 The epidemic is propagated by human-to-human transmissions only, and there were no additional zoonotic sources Demonstrated rapid accumulation of inter-host and intra-host genetic variation, including a number of genetic changes distinct to this lineage
Evolution of the outbreak Carroll *et al*.[106], Simon-Lariere *et al*.[229], Ladner *et al*.[183]	The outbreak included multiple distinct viral lineages 3 viral lineages in Guinea Multiple lineages in Liberia, but most cases attributed to a single introduction/lineage Description of distinct lineages of EBOV defined by multiple mutations, which may be potentially phenotypically important
Viral adaptation and mutation rates Tong *et al*.[278]	Describes the mutation rate in this strain of EBOV – substitution rate 1.23 x10-3 substitutions per site per year (95% CI 1.04-1.41 x10-3) This is important both for vaccine development, but also in understanding phylogenetic analysis which requires some variability in pathogen sequences
Identification of transmission routes Arias *et al*.[156], Sissoko *et al*.[279]	Identification and confirmation of new transmission routes e.g. persistence of *ebolavirus* in breast milk. Leading to infection in the baby, after no symptomatic infection in the mother.

Factors that spread and sustained the epidemic Dudas *et al*.[280]	The largest phylogenetic study including 1,610 sequences Demonstrated the heterogeneous nature of the outbreak and that it was spatially dissociated, with transmission clusters of varying sizes, durations and connectivity The outbreak occurred in areas that were susceptible to substantial outbreaks
Phylodynamics of outbreak Stadler *et al*.[281,282], Volz *et al*.[283]	Combined approaches incorporating phylogenetics with mathematical modelling to attempt to infer key epidemic parameters from sequence data: • Ro estimated • Cannot reliably infer incubation or infectious periods • Useful in identifying super-spreading and extreme variance between number of transmission per person
Development of field real-time sequencing techniques for outbreak settings Quick *et al*.[284]	Development of quick sequencing in the field – practical details This paper describes the role this might play in surveillance of future outbreaks

iv) The use of epidemiological models to assist complex decision making: A large range of models were developed during the outbreak (described above) to inform decisions such as: where to allocate scarce resources in real time[285]; how to define the likely scope of the outbreak in the absence of control measures[257]; how to describe 'worst case scenarios' that stimulated international organisations and national governments to provide the necessary support; and how to help differentiate between effective interventions and those with limited utility[286]. While these methods can be applied for many purposes, they are often limited by incomplete and poor quality data with which to build the model. The model is only as good as the underlying data, therefore, many projections are inaccurate and need to be interpreted with an understanding of the underlying data. A systematic review of mathematical models of the West Africa Ebola outbreak has made recommendations for further improvement of modelling in future outbreaks, including a degree of standardisation of techniques to enable comparisons that could then be shared rapidly[287].

4.5.2 Clinical aspects

Many of the clinical and research findings from the 2013-16 outbreak have helped confirm our prior knowledge of EVD[245], but progress has been limited. The data confirms previously documented incubation periods and the fact that people are not infectious until after they become febrile, giving additional support to the value of contact tracing, quarantine with fever surveillance, and early isolation for possible and suspect cases[214].

In line with previous outbreak emergence theories, bats have been suggested as a possible source for the 2013-16 epidemic, based on retrospective epidemiological data - interviews with contacts of the index case in Guinea indicated that the two year old child regularly played near a tree found to house a colony of insectivorous bats[288]. However, no evidence of *Ebolavirus* infection was found based on RT-PCR and serology assays in 169 specimens obtained from bats in this area[288]. Although, species of bats known to carry *Ebolavirus* were also identified close to the where the index case lived. Thus, a confirmed source of *Ebolavirus* for this outbreak remains elusive.

Finally, mortality rates appeared to decline throughout the outbreak[289]. Nonetheless, we acknowledge that this is subject to ascertainment bias of underreporting of milder cases early in the outbreak, and the difficulty in calculating a crude fatality rate from the proportion of fatal cases in the midst of the outbreak when the lag time between case reporting and case recovery or death alters the results[290]. It is thought that this finding could have been due to increased treatment-seeking, enabling better contact tracing with chains of transmission more rapidly identified and disrupted.

4.5.3 Operational aspects

Despite the numerous criticisms of the slow and inadequate response at both national and international level, once the outbreak was acknowledged and the response begun, one of the most impressive aspects of the outbreak response was the unprecedented levels of international collaboration and cooperation. The response was truly global and included a diverse range of organisations and national delegations from across the world, including NGOs, humanitarian, military, and governmental bodies. The global health community should be indebted to all those who were involved in this response, but especially to NGOs, including MSF, which first raised the alarm about the scale of this Ebola outbreak,

and a host of others that were integral in establishing and maintaining an effective response.

4.6 Lessons learned

The response to the 2013-16 Ebola outbreak in West Africa has been the subject of numerous independent assessments, with four different panels convened to review the international response, identify failures in the response and provide recommendations to strengthen future responses (Table 4.5). All four panels concluded that the slow response of the WHO was a significant factor in the outbreak not being contained. Recommendations common to all include enhanced global collaboration during outbreaks, including global governance and leadership, and the development of strong national core capacity in public health across the globe, as agreed under the International Health Regulations[248]. The IHR are negotiated global agreements for stronger health security and require all countries to develop and sustain eight core capacities in public health in order to better detect and respond to hazards, including infectious disease outbreaks. They also mandate international cooperation as a safety net when outbreaks spread across international borders, and pursuant to the Ebola outbreak an international committee was convened to assess the effectiveness of the IHR.

It is perhaps surprising to note, that many of the recommendations suggested by these four panels and other experts mirror the main conclusions from both the first 1976 Ebola outbreak and from Heymann *et al* after the Kikwit outbreak over 20 years ago[291]. These include the need for a stronger infectious disease surveillance (both national and global); improved international preparedness to provide support when similar outbreaks occur; more broad-based international health regulations; and continued and coordinated Ebola research, especially for valid diagnostic tests, better patient management procedures, and identification of the natural reservoir. We must hope going forward, that recommendations will be heeded and enacted to improve global health security.

Table 4.5: Summary of common recommendations from the four independent assessment panels

Panels	Recommendations

WHO Ebola Interim Assessment Panel[9]
 	
Harvard University and the London School of Hygiene & Tropical Medicine's Independent Panel on the Global Response to Ebola[8]
 	
US National Academy of Medicine's *Commission on a Global Health Risk Framework for the Future*[11]
 	
United Nations High-Level Panel on the *Global Response to Health Crises*[10]
	<u>Preventing outbreaks</u>:
	• WHO to re-establish leadership role as the guardian of global public health:
	  ○ WHO would lead responses, with Member states responsible for supporting WHO at the front-line by implementing core capabilities in public health (surveillance, issuing alerts, and outbreak responses)
	• Governance reforms to rebuild trust in WHO
	• WHO to develop a plan to ensure core capacities required under IHR for all countries (with World Bank) including:
	  ○ Strategies to ensure that governments invest in building core capacities to detect, report and respond rapidly to outbreaks
	  ○ New staffing approach for WHO country offices to ensure the highest-level capacity for vulnerable countries (with weak health systems and governance challenges)
	• WHO to create a Standing Emergency Committee with strengthened ability to independently identify health risks and declare public health emergencies
	• IHR review committee to consider the possibility of an intermediate level/grade alert to engage the international community at an early stage without declaring a full Public Health Emergency of International Concern
	• IHR review committee to consider incentives to encourage countries to notify public health risks to WHO (e.g. public commendations and/or financing mechanisms to mitigate adverse economic effects), and disincentives to discourage countries from instigating travel restrictions and limit trading, without scientific justification or WHO recommendation
	<u>Responding to outbreaks:</u>
	• WHO to scale back its broad remit and refocus on providing technical support
	• WHO to create a dedicated "Centre for Emergency Preparedness and Response" with strong technical expertise
	• WHO should ensure a protected budget and contingency fund to support this new Centre and allow rapid deployment of emergency response when required. This would include annual contributions from Member State, International Monetary Fund, World Bank, and other multilateral donors
	• Clear response mechanisms for coordination and escalation in health crisis, including mobilisation of the UN system for humanitarian crises, and cooperation with non-state actors including civil society organisation, the private sector, and the media
	• Consideration of formalised regional and sub-regional level arrangements to enhance prevention of and response to health crises
	• WHO and partners to ensure that appropriate community engagement is a core function of health emergency response
	<u>Research: production and sharing of data, knowledge, and technology:</u>
	• Develop a framework for research and development operations in outbreak settings to enable, accelerate, and ensure good governance. Timely roll out of successful results should also be encouraged. WHO should play a central convening role in this.
	• Develop a worldwide financing facility for outbreak-relevant diagnostics, vaccines, therapeutics, and medical and information technology (with no commercial incentives)
	<u>Governing the global system for preventing and responding to outbreaks:</u>
	• Create an independent UN Accountability Commission to oversee the new response Centre, monitor compliance with the IHR's Core Capacity requirements at a country level, and assess function in outbreak responses
	• Create a UN Security Council led Global Health committee to put global health issues at the centre of the global security agenda and expedite high-level leadership
	WHO Executive Board should mandate good governance reforms at the country-level, including establishing a freedom of information policy. They should have the capacity to challenge governments and hold them publicy accountable for protecting public health.

4.7 Conclusions

The 2013 -2016 Ebola outbreak in West Africa was described by the WHO as an "old disease in a new context"[237] due to the many unique aspects of this outbreak: it was the largest and longest Ebola outbreak in history, being larger than all others combined; it was the first with multi-continental spread, initially within Africa and then to Europe and North America; and it was the first in which there was a pipeline of clinical products that could be studied for effectiveness.

Although this outbreak was quantitatively many times larger than previous outbreaks, it was not qualitatively different – the means of amplification and transmission were the same as previously described. As a result, the control measures designed and implemented in previous outbreaks were crucial to the response, but the failure to implement these in a timely manner had devastating consequences. This outbreak specifically highlighted the crucial role of successful community engagement in outbreak response. While these established measures will remain the mainstay of control strategies, novel technologies (e.g. real-time sequencing) can provide additional benefit when used in tandem to supplement established strategies. Therefore, these old messages need reinforcing early and vocally in new epidemics.

Traditional epidemiology remains the mainstay of outbreak response. However, this outbreak was the first acute emerging infectious disease outbreak where molecular studies were undertaken in the field. These studies were not truly real-time during this outbreak, but hold the potential to be in future outbreaks, as demonstrated by the work presented in this thesis. However, molecular epidemiology provided a useful adjunct to traditional methods, particularly in understanding how and why new cases arose and the origins of these cases. Molecular epidemiology provides an additional tool alongside traditional methods, but it is not a substitute. Without traditional epidemiology and surveillance, there would be no specimens on which the molecular studies rely.

The size and length of this outbreak provided a unique opportunity to develop and test evidence-based protocols to improve both basic supportive clinical management (e.g. fluid and electrolyte treatments) and public health intervention implementation. However, woefully little progress has been made in these areas and the same questions remain pertinent now as were concluded after the Kikwit outbreak over 20 years ago[291]. Had the

key lessons from historical outbreaks been enacted, thousands of lives could have been saved. This lack of progress was a real missed opportunity to improve outcomes for future outbreaks and is perhaps one of the biggest failings in the outbreak response. We can only hope that we have now learned the lesson, and the global community will heed the recommendations suggested and translate these into practical solutions to make a difference going forward.

Although the initial outbreak response was slow and inadequate, once established, the response demonstrated unprecedented and impressive levels of international cooperation [292]. National governments and the global health community should be indebted to all the organisations and individuals who courageously took great personal risk to contribute to this effective response. This outbreak was also a salutary lesson; collaborations and structures are in place and must now be better used to enable a rapid and effective national response to outbreaks, while providing a global safety net for response when national efforts fail to prevent international spread. There is also evidence that the IHR are working better: the Emergency Committee of the IHR recommended to WHO to declare the outbreak of Zika virus a global public health emergency of international concern in February 2016[293].

While at the time of writing the 2013-16 outbreak of Ebola in West Africa has been declared over, WHO continues to stress that "Sierra Leone, Liberia and Guinea are still at risk of Ebola flare-ups, largely due to virus persistence in some survivors, and must remain on high alert and ready to respond"[224].

Acknowledgments

I would like to thank my co-authors for their comments and contribution to the publication based on the work presented in this chapter.

CHAPTER 5

ROLE OF HEALTHCARE WORKERS IN EARLY EPIDEMIC SPREAD OF EBOLA: IMPLICATIONS FOR PROPHYLACTIC COMPARED TO REACTIVE VACCINATION POLICY IN OUTBREAK CONTROL.

In this chapter, I use classical epidemiological methods to explore transmission dynamics in early outbreak settings. I learn how to construct hypothetical epidemiological transmission trees and how to exploit publicly available data to maximise information. I use this information to hypothesise about the impact of effective vaccination strategies and how this information can be used to inform intervention strategies and policy.

Of note, the work presented in this chapter was undertaken during 2015, before any vaccine trials were initiated during the West African outbreak. This chapter, the results and conclusions should be interpreted in light of this.

A brief background to Ebola vaccine progress during the West African Ebola outbreak is given in section 4.5.1.

This work has been published: Coltart CEM, Johnson AM, Whitty CJ. Role of healthcare workers in early epidemic spread of Ebola: policy implications of prophylactic compared to reactive vaccination policy in outbreak prevention and control. *BMC Med* (2015). 19; 13: 271[168].

5.1 BACKGROUND

Preventing or mitigating future epidemics is a public health priority for Africa. Existing strategies to prevent transmission of Ebola Virus Disease (EVD) are effective public health measures (e.g. infection prevention education, isolation, contact tracing, safe burial) and vaccination strategies. There is widespread agreement that an effective vaccine would be an important tool to respond to future epidemics. Several vaccines are currently being trialled, but true efficacy study data will be limited. How best to deploy these vaccines most effectively is a crucial policy question in debate and various strategies must be considered. In this paper, I set out to explore the differential impact of four vaccination strategies in preventing early epidemic transmissions and, thus, potentially limiting or aborting future epidemics.

Healthcare workers (HCWs)[‡‡] have been shown to have a high frequency of severe disease and deaths from many emerging infections including SARS and EVD[99]. For example, up to 25% of reported cases across historic EVD epidemics[294,295] and a case attack rate of 31% for physicians reported in a single EVD epidemic[296]. In the recent West African outbreak, prior to the provision of good equipment and impeccable attention to infection control procedures, the relative risk for acquiring EVD was around 100 times higher for HCWs compared to the general population[261], although this risk can be substantially reduced with the correct use of barrier precautions and personal protective equipment (to 21-32 times that of the baseline population)[262,297]. However, by the time EVD is suspected, diagnosed and precautions implemented, many exposures and infections have often already occurred, as the incubation period for EVD is relatively short (2-21 days, average 11 days)[119] and transmissibility high among severely ill patients in healthcare settings. Therefore, in these early stages of an epidemic, before the outbreak has been recognised, HCWs not using personal protective equipment will be at substantial risk and often constitute a high proportion of early cases. Therefore, there is a case to vaccinate HCWs in potentially affected countries for their own protection.

[‡‡] The definition of a HCW varies considerably. In this review, I define HCWs as 'all people engaged in the promotion, protection or improvement of the health of the population'. If taken in a literal sense, this would include all family members and volunteers caring for relatives, therefore, I restrict the definition to include only those who are health service providers with direct patient contact, training and skills to undertake this work, and an occupational status or paid salary. Healthcare management and hospital support staff are not included as they do not routinely have direct clinical contact with patients of the sort which tends to increase risk of acquiring EVD. Table 5.3 (conclusions) outlines the estimated numbers of HCWs in West Africa prior to the onset of the current epidemic.

It is likely that HCWs may provide a transmission link to both other HCWs and the general population, acting as 'super-spreaders' in early epidemics. However, to my knowledge there is no definitive evidence to support this. The importance of this transmission pathway is likely to predominate in early epidemic transmission chains and would decrease as the epidemic evolves, probably due to effective infection control measures for HCWs once the epidemic is recognised. This trend was seen by the number of HCWs infected during different phases of the West African epidemic. There are a number of reasons that HCWs might be 'super-spreaders' of EVD early in epidemics: the occupational risks of acquiring disease are high, for example close contact with patients who are highly infectious, but not known to be infectious at the time (e.g. severely sick or dead bodies); and also there is an increased risk of spreading infection through societal and cultural factors, including the fact that HCWs can be significant public figures within communities, they tend to live in larger groups, touch a wide variety of strangers (both patients and social contacts), travel more widely, and have large traditional burials. Strategies to break or reduce these key transmission chains, including vaccination strategies of HCWs, could be critical in containing an outbreak during the early stages.

Currently, there are no approved vaccines for EVD or any of the other haemorrhagic fevers. However, partly as a result of EVD having been listed as a potential bioterrorism high-priority agent[298,299], vaccine development work had begun prior to the recent outbreak. Potential vaccines exist for both *Zaire* and *Sudan* strain of *Ebolavirus*. Recent events in West Africa have increased the urgency for an Ebola vaccine and fast track clinical trials are underway for three vaccines with expedited preclinical to clinical trial phases[276,300,301]. Several more are at advanced pre-clinical stages of development. Initial data from these early trials suggest they are safe[300], and animal data from non-human primates (a reasonable model for human Ebola) suggest they are likely to be efficacious[302,303], although efficacy cannot be determined until Phase II/III trials are completed. However, due to the rapidly falling case numbers, efficacy studies will be limited and vaccine efficacy inferred from a non-outbreak setting. However, should these vaccines prove to be effective, developing strategies to maximize the impact and cost-effectiveness of vaccination is crucial to prevent or minimise the effect of future epidemics.

Whilst the ideal Ebola vaccine would be safe, well tolerated, rapidly effective, and have long duration of high levels of protection, there may well be a trade-off between different characteristics (e.g. tolerability versus speed of onset versus duration of protection). At least four possible vaccination strategies are worth considering, and these have different ideal vaccine profiles:

1) Ring vaccination[§§] of the contacts of cases. This would need to be rapidly effective; duration of protection would be less important.
2) Mass vaccination of the general population during an outbreak. Here safety and ease of use (e.g. single dose) would dominate as essential characteristics; efficacy and speed of protection would be less important.
3) Reactive vaccination of HCWs caring for patients during an epidemic. The efficacy of the vaccination would be essential given the high risk of transmission and speed of onset important; duration of protection beyond a few months and mild side effect profile would be less important given the high risks associated with infection.
4) Vaccination of HCWs in endemic counties prior to an outbreak as a prophylactic measure. Unlike other vaccine strategies, duration of protection would be essential and safety and tolerability would be much more important as most vaccinated HCWs would not encounter the virus. Speed of onset of protection would be less important.

There is no question that if an effective vaccine was developed it would be used reactively to vaccinate HCWs in an epidemic. However, I hypothesised that the magnitude of outbreaks could be significantly reduced or even aborted at an early stage with vaccination and protection of HCWs, with a prophylactic strategy likely to be more effective at preventing early epidemic transmissions compared to a reactive strategy once epidemics are detected. Therefore, this paper explores prophylactic vaccination strategy (number 4), compared to the reactive strategy (number 3).

If prophylactic vaccination of HCWs is likely to reduce transmission substantially, it would be worthwhile developing well-tolerated vaccines which have long duration of protection, irrespective of the speed of onset of protection, as epidemics are rare. If not, then speed of onset and efficacy could be the main parameters, with duration of protection beyond a

[§§] Ring vaccination= vaccination of a 'ring' of people around an infected case i.e. those potentially exposed

few months of lesser importance, leading both to different product profiles and to different ways of evaluating vaccines. In a reactive strategy, vaccines could be held in emergency stock and only deployed in epidemics. A two-dose (or more) vaccine strategy would be realistic for prophylactic vaccination of HCWs, who are a limited, easily identifiable, and a largely stable group, but would be much more challenging strategy for mass or ring vaccination. The two vaccination strategies are, therefore, substantially different. It would only be worth developing a prophylactic vaccination strategy (and target product profile) if it is likely to have significant additional benefits.

Given the current speed of development of vaccines, and discussion on Ebola vaccine policies for the future, the aim of this paper is a rapid analysis of known and provisional Ebola transmission tree data to assess the differential impact of prophylactic versus reactive vaccination strategies for HCWs in preventing early epidemic transmission. Descriptive hypothesis generation and testing of different vaccination strategies for HCWs were undertaken to see if a long-duration vaccine for prophylactic use is sufficiently attractive to be worth prioritising due to its impact on transmission of early epidemics, as compared to reactive vaccination.

5.2 Methods

In order to test the hypothesis, I carried out a review and basic analysis of data available from the current and historic Ebola epidemics. The methods and search strategy for the review are detailed in Figure 5.1 and Table 5.1. From the articles identified I then reconstructed the initial epidemic transmission trees for all available previous outbreaks of Ebola virus *Zaire* strain initially using published and grey literature, and where these were not available, press reports. The transmission trees detail the number of cases, as well as who transmitted the virus to whom. Therefore, it is possible to estimate the number of cases that resulted from HCW transmission. These transmission trees were then used to test the hypothesis that the majority of known epidemics, or early cases in epidemics, could have been be reduced or aborted by vaccination of HCWs, with a prophylactic strategy likely to be more effective than a reactive strategy. To do this I assessed the proportion of the early known EVD epidemic cases that would have been averted under various vaccination strategies if HCWs had been protected and had not contributed to transmission chains (Figure 5.2).

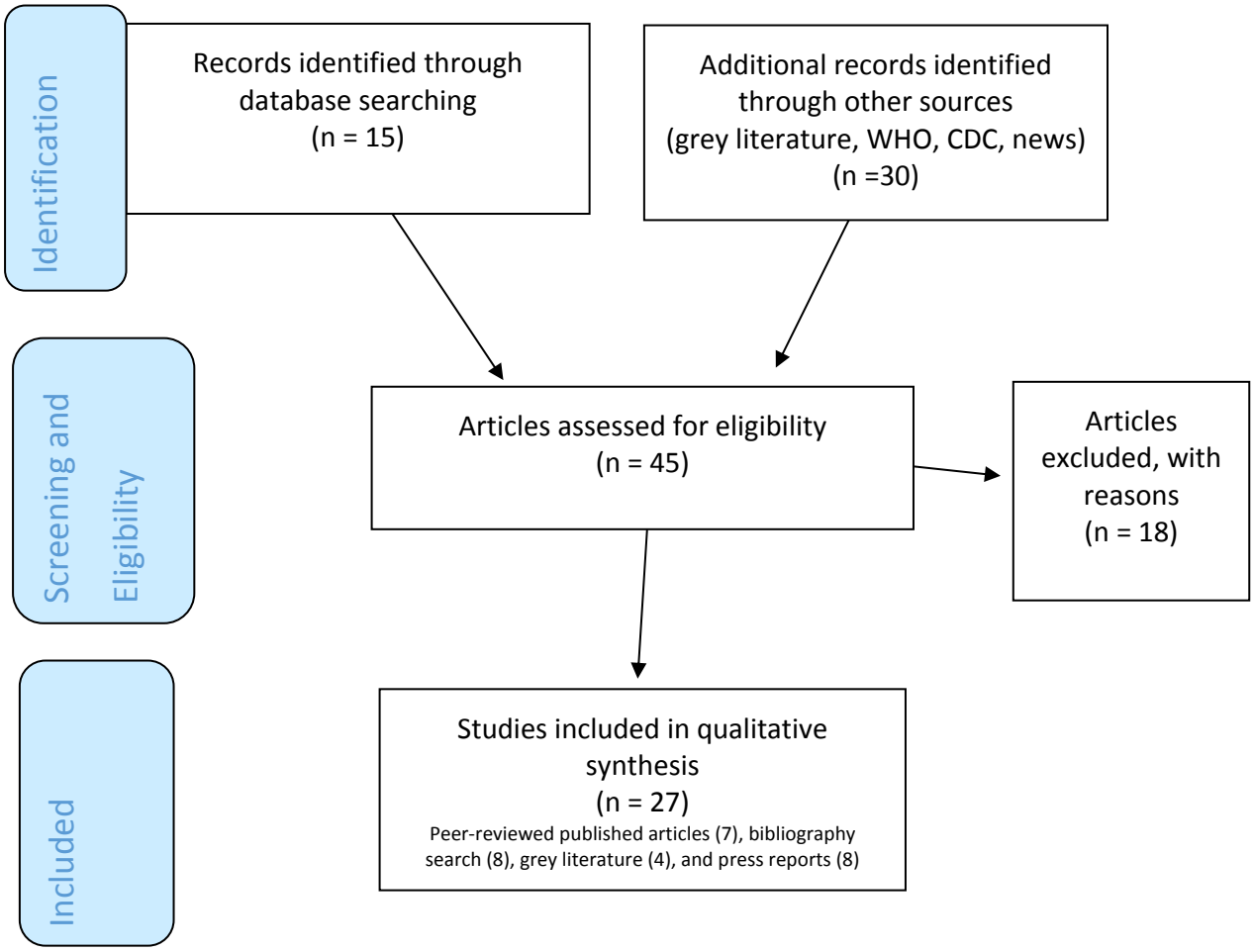
Table 5.1: Methods and search strategy for review and transmission tree reconstruction

Search strategy	A compound search strategy was developed to identify all relevant open-source articles regardless of publication status. The initial search was undertaken via PubMed using the search terms outlined below. Further relevant information was obtained by reviewing article bibliographies for relevant citations and a Google search to find open-source published articles, press articles and other grey literature including outbreak updates, WHO roadmaps, WHO situation reports and Morbidity and Mortality Weekly Reports from the Centre for Disease Control.
Search terms used	Key words: *Ebola, Ebola haemorrhagic fever, Ebolavirus, Ebola virus disease, transmission, transmission trees, epidemic, epidemic trees.* *((("hemorrhagic fever, Ebola"[MeSH Terms] OR ("hemorrhagic"[All Fields] AND "fever"[All Fields] AND "Ebola"[All Fields]) OR "Ebola hemorrhagic fever"[All Fields] OR "Ebola"[All Fields] OR "Ebolavirus"[MeSH Terms] OR "Ebolavirus"[All Fields]) OR ("hemorrhagic fever, Ebola"[MeSH Terms] OR ("hemorrhagic"[All Fields] AND "fever"[All Fields] AND "Ebola"[All Fields]) OR "Ebola hemorrhagic fever"[All Fields] OR ("Ebola"[All Fields] AND "virus"[All Fields] AND "disease"[All Fields]) OR "Ebola virus disease"[All Fields])) AND ((("transmission"[Subheading] OR "transmission"[All Fields]) AND ("trees"[MeSH Terms] OR "trees"[All Fields] OR "tree"[All Fields])) OR (("epidemics"[MeSH Terms] OR "epidemics"[All Fields] OR "epidemic"[All Fields]) AND ("trees"[MeSH Terms] OR "trees"[All Fields] OR "tree"[All Fields])))*
Inclusion criteria	Studies and articles were eligible for inclusion if they reported human-to-human transmission chains and/or gave details of occupational exposure of infected patients which could be traced to individuals in the transmission tree. There were no restrictions with regards to date of publication. The aim was to identify as many early epidemic trees as possible, and include all which could be reliably identified to minimise bias.
Data extraction	Data extraction was first undertaken using peer-reviewed literature. These data were supplemented by grey literature and press articles to add further information and fill in the gaps for the epidemic tree construction (CEMC).
Quality of included studies	The purpose of the review is to outline potential policy implications for vaccine development, as opposed to define detailed epidemic trees. Approximate numbers have been used where precise numbers were not available, but in all cases, conservative assumptions have been made to avoid overestimation of any effects identified.
Data analysis	Initial analyses of the different vaccination strategy were undertaken on Microsoft Excel (2010). Using the initial transmission trees constructed, the number of cases that both developed the disease and were averted were calculated for each vaccination strategy. All results are expressed as a percentage of averted cases by the total number of cases included in the initial transmission tree per vaccination strategy and epidemic location.

Of the 15 Ebola virus *Zaire* strain outbreaks, initial epidemic transmission trees of varying detail were reconstructed for eight outbreaks (see supplementary information). Only two studies outlined transmission trees directly (the Guinea and Nigeria epidemics of the recent West Africa outbreak), while all others were constructed using multiple data sources and linking of this information to reconstruct a 'best approximation' tree. For many of the historical outbreaks there was no detailed person-to-person transmission information to enable construction of transmission trees, only hypothetical transmission scenarios and, therefore, these were subsequently excluded. Therefore, our analysis is based on three outbreaks. Where possible, information regarding HCW and traditional healers were included in the reconstructed trees, as was any data on date of infection.

Transmission trees from later in the outbreak are available but have not been included as this study focuses on early transmission[304].

Figure 5.1: PRISMA Flow Diagram



My analysis is based on the recent outbreak in West Africa (including all nine countries in which cases have occurred, each considered as a separate transmission tree in this study based on an imported primary case) and two historical outbreaks for comparison. The data for Sierra Leone gave gross figures to enable limited inclusion in the analysis, although a detailed transmission tree cannot be constructed reliably at this time. One historical outbreak includes two separate transmission trees based on different locations; Kikwit and Mosango, the latter due to an imported case from Kikwit. The historical outbreaks were selected according to those outbreaks with transmission trees detailed enough to allow the proposed analysis (Yambuku 1976, Kikwit 1995 and Mosango 1995).

My analysis investigates four different vaccination strategies for HCWs to determine the most likely maximally effective strategy:

1. Strategy 1: Prospective prophylactic vaccination of all front-line HCWs in high-priority areas prior to an epidemic and those likely to be deployed to epidemic areas, assuming >99% vaccination coverage and >99% vaccination efficacy. In this scenario, all cases of HCW infection were prevented, as were the cases arising from transmission by HCWs.

2. Strategy 2: Prophylactic vaccination as above with only 75% coverage (>99% efficacy). As not all HCWs are vaccinated in this scenario the anticipated outcome of the vaccination strategy varied depending on which HCWs were not vaccinated. In all transmission trees the HCW chosen for vaccination were those associated with the least number of linked transmission cases, except in the situation in which a HCW was not documented to have transmitted the disease. For example, where 75% vaccine coverage is analysed, the 25% of HCWs (i.e. one in four) in the transmission tree who were not vaccinated were deliberately selected to minimize the number of cases averted. This method guards against an overestimation of the efficacy of a vaccination strategy by giving the most conservative estimates possible ('worst case scenario' with the highest onward transmission occurring). Therefore, any effect identified by a given strategy would provide an underestimation of the true value rather than overestimation. Figure 5.2 gives a visual example of two epidemic trees and details the analysis undertaken.

3. Strategy 3: Reactive rapid vaccination of all front-line HCWs once an epidemic has been identified. I assume a lag-time of 42-days from case presentation to immune protection from vaccine response, based on a 21-day logistical window from epidemic identification to initiation of a vaccination campaign (including outbreak verification, vaccine transportation time, training, and strategy refinement), followed by a 21-day lag-time for immune protection following vaccination. This is based on recent data from Chimpanzee Adenovirus vector Ebola vaccine studies showing a mean time for maximal antibody response of 28 days (63-90% ELISA positive for anti-glycoprotein IgG at 28 days depending on vaccine dose), and a mean time for T-cell response of 14 days[276]. I took the average of these, which is 21 days. To calculate the averted cases for reactive vaccination the dates of infection were analysed. Any HCW presenting more than 42 days after the date of the index

case was considered to be potentially vaccinated and protected, therefore preventing infection and any subsequent ongoing transmission further down the transmission tree.

4. Strategy 4: Extension of strategy 1 (prophylactic vaccination) to include traditional healers and front-line HCWs. Therefore, all cases of HCWs, traditional healers, and subsequent ongoing transmission from both these groups would be prevented.

5.3 Results

Figure 5.2 shows examples of epidemic trees, the analysis undertaken, and the potential impact of each vaccination strategy on transmission. It demonstrates that that the majority of early cases in those who were not HCWs could have been prevented by a prophylactic vaccination strategy, whilst a reactive vaccination strategy would have had little impact on the initial epidemic.

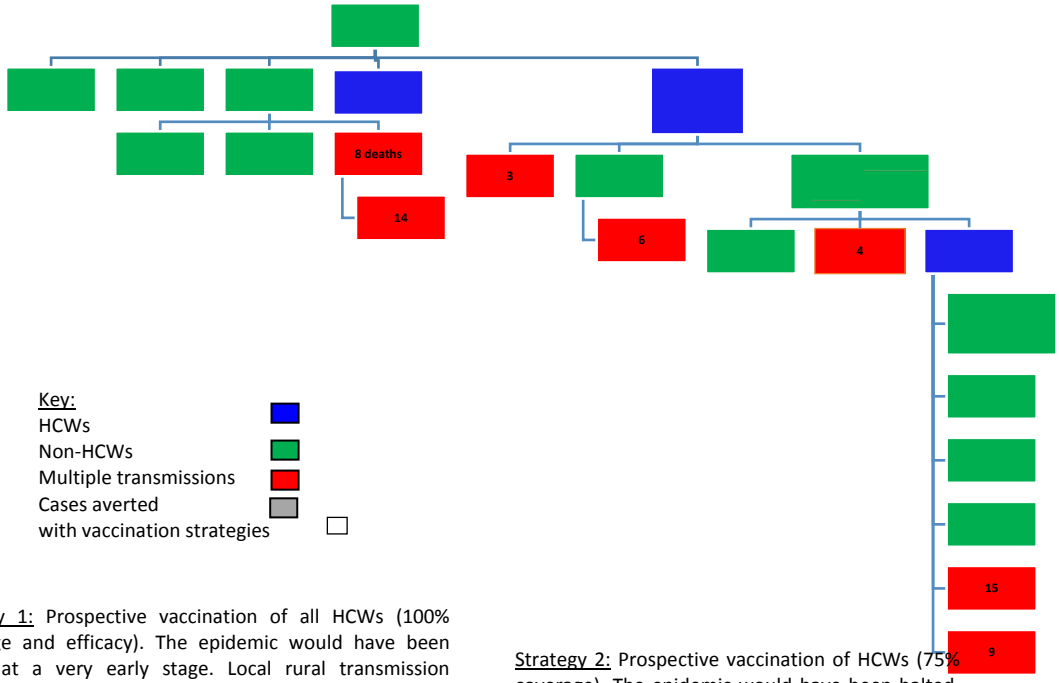
Table 5.2 shows the putative impact of the first three vaccination strategies outlined by epidemic location. It details the cases averted had HCWs been vaccinated and therefore protected:

- Strategy 1: Overall, approximately two-thirds (65%, 73/115) of early epidemic cases across three different epidemics (and 12 outbreak locations) would have been averted with prophylactic vaccination of HCWs and a vaccination coverage >99%. Across all epidemics, this strategy decreased early epidemic transmission by between 38-100%. The Sierra Leone data suggest that at least 25% (125/506) of early transmissions would have been averted with this strategy, but as this is based on gross approximate numbers with which I was unable to construct a transmission tree these figures have not been included in our overall analysis.

- Strategy 2: In this strategy I assume prophylactic vaccination of approximately 75% of HCWs. This continues to show evidence of averting over 40% (42%, 58/138) of epidemic cases. Across all epidemics the percentage of cases averted ranged from 11-74%. There will be a drop–off point in which vaccine coverage falls below critical levels, but further work would be needed to model this. These data are based on two outbreaks (West Africa and Mosango) as the other two historic outbreaks did not provide enough detailed transmission data to determine exact transmission chains with relation to HCW status.

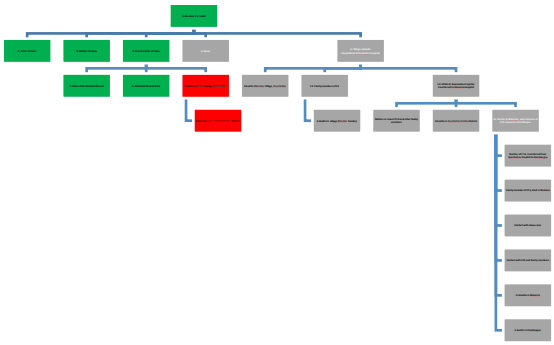
Figures 5.2: Two examples of epidemic transmission trees and the impact of four different vaccination strategies on the transmission chains

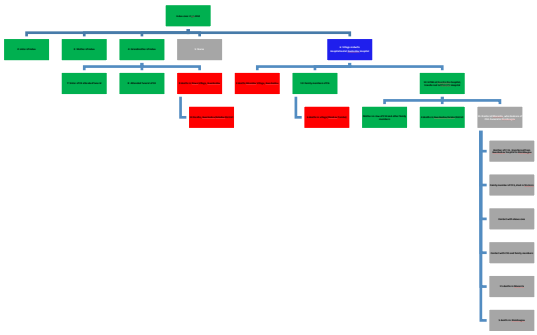
5.2a: Guinea 2014 outbreak[167]



Strategy 1: Prospective vaccination of all HCWs (100% coverage and efficacy). The epidemic would have been halted at a very early stage. Local rural transmission occurred from index case, but the vast majority of ongoing transmission would have been halted (61%).

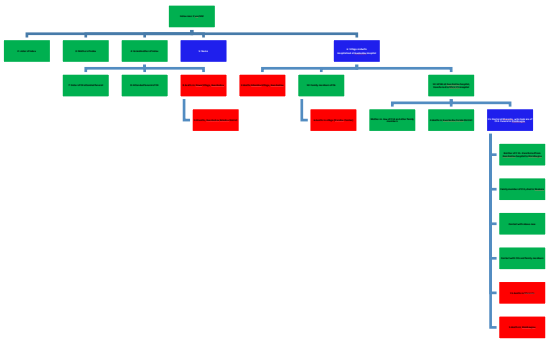
Strategy 2: Prospective vaccination of HCWs (75% coverage). The epidemic would have been halted, but less effectively than 100% coverage (36.6%).

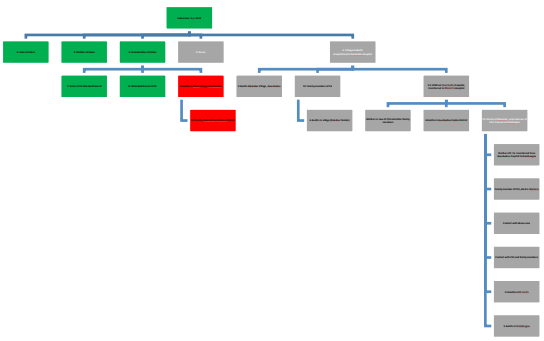




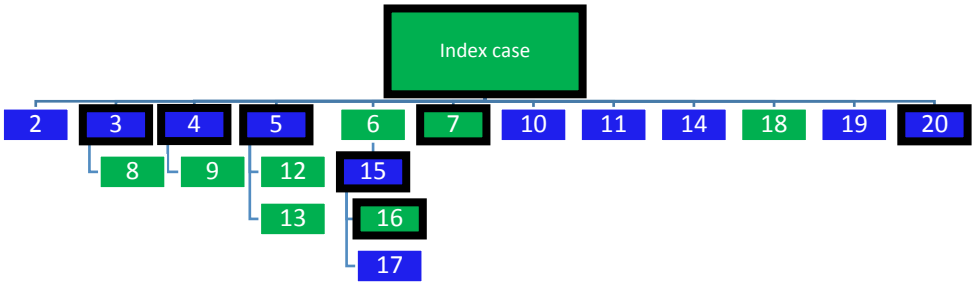
Strategy 3: Reactive vaccination. No effect in halting the onset epidemic tree with 42-day lag time. All transmission in the tree had occurred by this timepoint.

Strategy 4: Prospective vaccination of all HCWs and local traditional healers (100% coverage and efficacy). No difference between strategy 1 and 4.



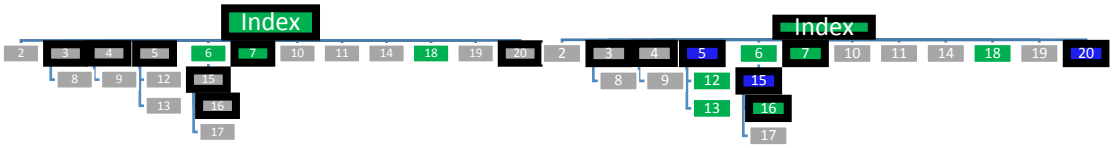


5.2b: Nigeria 2014 epidemic[225]



Strategy 1: Prospective vaccination (100%). Epidemic size decreased by 80%

Strategy 2: Prospective vaccination (75%). Epidemic size decreased by 50%)



Strategy 3: Reactive vaccination. Index case diagnosed on 23 July 2014. Therefore with a 42 day lag time before vaccination campaign and protection, all the transmission in this epidemic tree would have occurred. 0 cases prevented.

Strategy 4: Prospective vaccination (100%) for HCWs and traditional healers. No added benefit to strategy 1 alone.

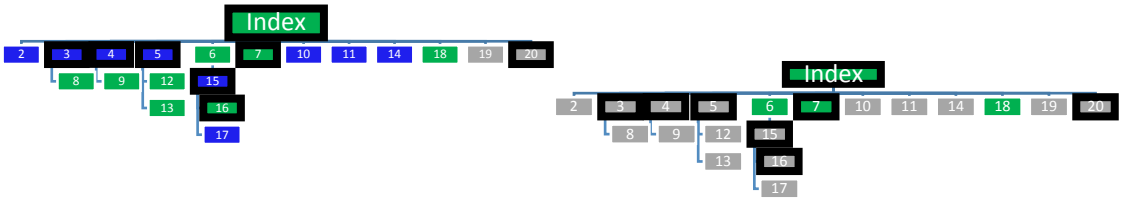


Table 5.2: Proportion of early outbreak prevented by implementing different vaccination strategies: Prospective versus reactive vaccination of healthcare workers

Epidemic	Country	Total cases	Total deaths	Cases in epidemic tree	% of initial outbreak prevented by vaccination strategy		
					Strategy 1: Vaccinate prophylactic-ally (100% coverage)	Strategy 2: Vaccinate prophylactic-ally approx. 75% of HCWs	Strategy 3: Vaccinate reactively (lag time 42 days)
2014 West Africa	Guinea[167]	3,108	2,057	71	61% (43/71)	36.6% (26/71)	0
	Liberia[305-309]	9,007	3,900	9	67% (6/9)	11% (1/9)	0
	SL[310]	11,103	3,408	NR	NR	NR	0
	Nigeria[225]	20	8	20	80% (16/20)	50%(10/20)	0
	Mali[226-228]	8	6	8	38% (3/8)	13% (1/8)	0
	USA[230]	4	1	4	75% (3/4)	50% (2/4)	0
	UK[232] and Spain[231]	2	0	2	100% (2/2)	50% (1/2)	0
	Senegal[233]	1	0	1	0	0	0
	Overall (95% CI)	23,253	9,380	115	63·5% (73/115) (0.54-0.72)	35·7% (41/115) (0.27-0.45)	
Historic outbreaks	Kikwit [294,311-314]	315	250	9	100% (9/9)	NR	NR
	Mosango[315]	23	18	23	100% (23/23)	74% (17/23)	NR
	Yambuku[100]	318	280	45	44% (20/45)	NR	NR
Total (95% CI)				**192**	**65·1%(125/192) (0.58-0.72)**	**42·0% (58/138) (0.34-0.51)**	**0·0% (0/609)**

Cases numbers accurate as of 18/02/2015
NR = not reported

- Strategy 3: A reactive vaccination strategy was assessed. This strategy was unable to prevent any early cases (0%, 0/609) and was, therefore, ineffective at mitigating epidemics based on this study data. As only initial transmission trees were used most of the data does not extend past the third wave of infections and does not detail transmission events after 42 days to fully assess the impact of this strategy on later epidemic transmission.

- Strategy 4: When the vaccination strategy included traditional healers the effect was context-dependent. It had a large effect (63-100%) in two regions of the current West Africa epidemic, namely the outbreaks in Sierra Leone and Mali, but did not appear to have an effect in the majority of locations. Based on this limited information, it is difficult to draw conclusions.

These findings suggest that prophylactic vaccination of HCWs with coverage as close to 100% as possible appears to be the most successful strategy, averting the highest proportion of early epidemic cases, and might have led to a significant reduction, or even avoidance, of epidemics by preventing the subsequent snowballing cascade which resulted in the exponential growth of cases. However, a more realistic target of 75% coverage also showed clear benefit over a reactive vaccination strategy. Finally, a reactive vaccination strategy did not prevent any cases in the first few of waves of the initial epidemic. More data would be required for formal modelling of each strategy as a matter of urgency. Regrettably, good quality data that could be used to support this analysis has not been compiled from the early phase of the current epidemic.

5.4 Conclusions

Although all Ebola epidemics are devastating for the local communities, the recent epidemic in West Africa has been on a scale not seen before in spreading to numerous other African countries as well as to other continents with global implications. Questions addressing how to prevent further outbreaks of this scale need to be answered with urgency. I hypothesized that vaccination of HCWs could have a substantial epidemic-wide effect, particularly prophylactic vaccination to prevent early transmission via HCWs, thereby markedly slowing the epidemic at an early stage and limiting the subsequent cascade of infection and death. Using historical data this study provides evidence to support this hypothesis and highlights the likely benefits of prophylactic vaccination of HCWs in endemic regions (where Ebolavirus is endemic in animal populations). Prophylactic vaccination coverage >99% would be most effective, averting nearly two-thirds of early epidemic cases studied. However, vaccine coverage of approximately 75% still confers clear benefit at around 40% of cases averted. Reactive vaccination in high-risk areas did not appear to be of value in preventing early disease transmission, although in the setting of a confirmed outbreak, reactive vaccination would be an essential humanitarian priority to protect HCWs (both international and national) and maintain the workforce key to controlling any epidemic[261,316]. Therefore, based on this data, a prophylactic vaccination campaign for all front-line HCWs with a vaccine providing long-lasting immunity should be seriously considered. It is likely to have a more profound

impact on the prevention of future outbreaks and epidemics than reactive vaccination strategies, and in some cases might stop them completely.

In real terms from the recent West African Ebola outbreak, 100% vaccination coverage equates to approximately 7,400 HCWs (physicians, nurses and midwives) (Table 5.3) across Guinea, Liberia and Sierra Leone in a population of nearly 22 million. Based on the available data, vaccinating 7,400 HCWs could have prevented the epidemic starting, or decreased the severity of this epidemic in the early stages and potentially prevented thousands of deaths. A prophylactic vaccination campaign for HCWs would be relatively easy to administer and may be more acceptable with high uptake rates than in the general population. For example, routine vaccination at medical and nursing schools, as with Hepatitis B vaccination in many countries. Whether it would be cost effective will depend on vaccine efficacy, duration of protection, turn-over of HCW staff and pricing structure. This will need to be modelled once these data become available for vaccines under development. If a vaccine successfully emerges from development, finding the most effective vaccination strategy is critical in the current low-resource setting, with a dearth of resources and infrastructure to administer population vaccination campaigns.

Many of the historical outbreaks have been so rural that formal HCWs were not involved in the initial chain of transmission, but in some rural areas traditional healers and community workers play a substantial role in community structure and healthcare provision e.g. the first Ebola outbreak in Yambuku. I have been unable to assess the role of traditional healers in this paper due to the paucity of reliable data, but there is certainly an argument for further investigation of targeted vaccination campaigns of these providers, particularly in known 'at risk' areas or reactively after the onset of an outbreak in a nearby area. This sort of targeted campaign would have prevented the spread of the current epidemic into Sierra Leone, where traditional healers play an integral part of healthcare and community structure and acted as early 'super-spreaders' of infection.

There are many ethical and practical issues in designing vaccination strategies, particularly in a crisis setting. It is important to note that numerous alternative vaccination strategies would also be valid and complimentary. Much transmission in the current epidemic occurs at a community level[304], which strategies focussed on HCWs would not directly target. This is particularly marked in later epidemic phases. This review demonstrates which of the

Table 5.3: Number of healthcare workers in West Africa[317]

Type of Healthcare workers	Number of Healthcare workers (Density per 1,000 pop.)		
	Guinea (2005)	Liberia (2008)	Sierra-Leone (2010)
Physicians	940 (0·1)	51 (0·014)	136 (0·022)
Nurses and midwives	3,839* (0·408)	978 (0·274)	1,017 (0·166)
Community and traditional healers	170 (TBA only)	142 (TBA only)	132
Others:			
Dentists	33	4	6
Pharmacists	199	269	114
Laboratory	252	115	14
Public health and environmental	67	40	159
TOTAL	**5,950**	**1,599**	**1,578**

*approximate data - inferred from previous data

possible HCW vaccination strategies are effective at halting or slowing the epidemic at an early phase, thus preserving healthcare services to cope with a community epidemic.

There are inevitably several limitations to this analysis. The main limitation is the paucity of open-access data and the inevitable incompleteness of epidemic trees. The analysis is therefore based on a small number of outbreaks with limited information. Furthermore, reported transmission events are biased towards larger events and hospital-based events, therefore, HCW association may be over represented. The definition of HCWs vary and there is lack of data regarding numbers of HCWs and the specific capacity in which they were working, making a denominator difficult to estimate. This study is not intended to be a fully detailed policy review of all vaccination strategies, which will have to wait until vaccine efficacy and duration data are available for formal modelling studies, nor evidence of a detailed epidemic tree. It is, however, an attempt to synthesise the available empirical evidence with the aim of informing those considering potential vaccine product profiles.

Several assumptions are made in this analysis, which will need to be modified as data become available. I assumed that the vaccine is entirely safe, 100% effective, and that implementing a vaccine strategy with 100% coverage is achievable. All three assumptions are unlikely to hold in reality. I also assume that outbreaks are from one single introduction in each strategy described. I believe this to be true in most, but not all epidemics (e.g. Kikwit, 1995). Furthermore, I assumed that HCWs play a key role in propagating ongoing transmission chains. Although the risks to HCWs of acquiring infection are well documented, I am not aware of literature confirming their role in

propagating further infections. I hypothesised this based on their high exposure risk, historical patterns of nosocomial spread in outbreaks of emerging infections (for example in SARS), and the likely social contact patterns of HCWs. If, however, there is limited ongoing transmission from HCWs, then vaccination programmes focussed on this target group would have little impact beyond protecting HCWs individually with some wider implications for hospital control. With the limited data included in this study, it is not possible to further explore the implications of the effect of transmission from HCWs to the wider population (high vs. low vs. no effect), as the numbers of HCWs documented are too low. However, this would be a valuable question to further explore with the wealth of data arising from the West African Ebola outbreak. Finally, I assumed there is no significant transmission from asymptomatic infection, an area of debate in the literature[318].

5.5 Summary

In conclusion, serious consideration of a prophylactic HCW Ebola vaccination strategy is needed, particularly in countries at threat from EVD epidemics. This analysis provides evidence for a substantial benefit of prophylactic compared to reactive vaccination strategies for HCWs on disease propagation, including if the prophylactic coverage is at a realistic target of 75%. The value of vaccination of HCWs is often seen solely in terms of personal protection and maintenance of the health (and morale) of the HCW workforce. However, our analysis of the limited available empirical data suggest that a prophylactic strategy could bring substantial additional benefit by preventing chains of early transmission, before the risk of epidemic spread is recognised. Therefore, there is a clear need for a vaccine with long-term efficacy and a good safety profile for use in prophylactic vaccination campaigns for HCWs based in countries with the potential for epidemic introduction. Given the geographical range of the probable bat reservoir, this is much of sub-Saharan Africa. This approach would appear to be a sufficiently effective strategy based on current data to be worthy of detailed consideration in assessing vaccine prioritization. The full economic impact resulting from a large Ebola epidemic is not limited to direct-Ebola healthcare costs alone, but also to many indirect consequences as a result of collapsed health systems, the effects of which will continue for many years. It is the global health community's responsibility to ensure an epidemic of this kind does not happen again, and prophylactic vaccination of HCWs with a safe and long-lasting vaccine, even if it was not of rapid onset, is likely to have significant potential to reduce the risk of future large epidemics.

Addendum

Since writing this chapter, vaccine clinical trials have taken place on healthy human controls, as well as part of a ring-vaccination trial during the West African outbreak. Unfortunately, this latter trial came too late to have any tangible impact on the outbreak or to accrue data on the number of cases averted. However, both trials were used to assess immunological response, tolerance and side effect profile of vaccines. The vesicular stomatitis virus vaccine (Merck) has been shown to be highly effective (70-100%) and safe[277] and will undoubtedly be used as a preventative strategy in the future. However, it is not without risks – the most notable side effect being arthralgia. Further data are needed regarding duration of immune response to be able to fully explore the optimal vaccine for different vaccine strategies.

Acknowledgements: I would like to thank my co-authors for their comments and contributions to the publication and this chapter.

CHAPTER 6

ANATOMY OF A COMMUNITY OUTBREAK OF EBOLA: A MULTI-DISCIPLINARY APPROACH

In this chapter, I use field data to undertake a descriptive analysis of a small isolated outbreak of Ebola during the West African epidemic of 2013-16. The analysis incorporates epidemiological, clinical, laboratory, sequence and behavioural data. I begin with a classical descriptive epidemiological analysis and develop epidemiological hypotheses about the drivers of infection. I then use a multidisciplinary approach to investigate the contribution of each discipline in enhancing understanding of the transmission dynamics in this isolated outbreak.

CHAPTER REDACTED AS PERMISSION HAS NOT BEEN RECEIVED TO PUBLISH THE DATA INCLUDED IN THIS CHAPTER

P150-186 REDACTED

CHAPTER 7

USING MOLECULAR DATA TO IDENTIFY THE TRANSMISSION DYNAMICS OF HIV-1 IN KWAZULU-NATAL: A COMBINED PHYLOGENETIC AND EMPIRICAL EPIDEMIOLOGICAL ANALYSIS

In this chapter I explore the discipline of phylogenetic analysis further. I work with a large HIV dataset and explore the construction and investigation of HIV phylogenies from South Africa. This work helped me develop more sophisticated linkage and analytical skills to fully exploit the potential of combining these two disciplines.

7.1 BACKGROUND

Understanding the dynamics of transmission within an epidemic is of critical importance in devising intervention strategies. One way to do this is to analyse the pattern by which a virus spreads within a population and, subsequently, to identify the characteristics of those involved in the transmission chains. Unfortunately, in HIV epidemics numerous factors limit our ability fully to understand the transmission dynamics at the population level, including incomplete sampling, the existence of large numbers of undiagnosed transmitters, and frequent stigma-related misclassifications. There is the potential significantly to improve our understanding if these limitations can be overcome, and my aim in undertaking this work was to investigate the extent to which molecular analysis is beneficial in this regard.

As set out in Chapter 2.2.2, it is HIV-1, Group M Subtype C, which is found in sub-Saharan Africa (Southern Africa). This single clade predominates due to a founder effect[j]. Within the generalised epidemic in sub-Saharan Africa there are multiple concentrated sub-epidemics[67] at both regional and local levels. Identifying these sub-epidemics and determining the transmission dynamics within regions, in the context of a larger hyper-endemic setting e.g. in KwaZulu-Natal versus South Africa, can be difficult both epidemiologically and phylogenetically[133].This is because epidemiological identification

[j] Founder effect = A situation in which a new population is founded by a small number of incoming viruses. Similar to a bottleneck, the founder effect severely reduces genetic diversity, increasing the effect of random genetic drift.

requires expensive and time-consuming longitudinal cohort studies, while distinguishing between sub-epidemics using phylogenetic studies can be difficult due to small sample sizes. Nonetheless, a detailed understanding both of the distribution of different circulating subtypes, and of the genetic diversity of the virus within Southern Africa, could be helpful in elucidating transmission patterns of the epidemic, which in turn would help to inform control programmes.

Prevention strategies targeting key groups driving HIV transmission could provide a more focussed and cost effective approach to HIV prevention. A combined phylogenetic and epidemiological approach can be used to identify these drivers of HIV transmission. For example, recent work at AC has shown the value of these combined approaches to determine the underlying HIV transmission dynamics and the source and consequences of high rates of HIV infection in young women in South Africa[91]. This demonstrated that sexual partnering between young women and older men, who might have acquired HIV from women of similar age, is a key feature of the sexual networks driving transmission[91]. Thus, prevention strategies should include interventions addressing socio-behavioural factors, such as age-disparate sexual partnering, to reduce HIV incidence.

My aim was to investigate the trends in HIV transmission dynamics within the population of the Africa Centre (AC) Demographic Surveillance Area (DSA) by integrating phylogenetic, phylodynamic and phylogeographical analyses with epidemiological analysis. The specific objectives were:

1. To describe the subtype C strains of HIV-1 circulating the AC DSA in the context of virus strains circulating in South Africa more generally;
2. To assess viral genetic diversity to define whether the AC DSA epidemic is isolated and self-sustaining, or whether it is the result of multiple introductions via migration, either from Kwa-Zulu Natal or from elsewhere in South Africa; and
3. To describe the composition and relevance of phylogenetic 'clusters'/microepidemics within the AC DSA population by adding metadata to phylogenetic analyses in order to determine fine detail transmission dynamics. This level of detail would be difficult to decipher through either epidemiological or phylogenetic analysis alone.

7.2 METHODS

The study site, study population, and sampling processes are described in Chapter 3. In brief, I used data from population-based, longitudinal surveillance conducted by the Africa Centre. This is a predominantly rural community in the uMkhanyakude district of KZN. The district is one of the most deprived in the country and is characterized by high levels of HIV infection, unemployment and circular migration. The data collected includes demographic and health data, in addition to anonymised HIV testing and HIV *pol* sequence generation where possible.

7.2.1 PHYLOGENETIC RECONSTRUCTION

I analysed a total of 2,179 HIV-1 subtype C partial *pol* gene sequences from South Africa, sampled between 2000 and 2014. They included 1,376 routinely collected samples from AC between 2010 and 2014 (AC sequences), and a convenience sample of 803 publicly available sequences from South Africa as local controls (ZA sequences). More detail is set out in Chapter 3.

Sequences were aligned using the Clustal algorithm[43,337], as implemented in the package Mega v.6.06[40]. I was provided with the aligned sequences and I then checked and manually edited these where necessary using AliView[41]. Using the software RAxML[327], I built a maximum likelihood phylogeny of the sequences, under the General Time Reversible model of nucleotide substitutions and varying substitution rates across sites (GTR + CAT). Branch support was calculated by Shimodaira–Hasegawa-like local branch support (SH-like test), as implemented in RAxML. To ensure robust results, the phylogeny was also reconstructed after removing from the alignment 38 codon positions associated with antiretroviral drug resistance[53]. This eliminated the risk of artefactual results from 'convergent evolution', i.e. clustering as a result of acquired mutations from selective pressure at drug-resistant mutation (DRM) sites, rather than inherited mutation.

7.2.2 GENETIC VARIABILITY

The pairwise genetic distance between all possible pairs of the 2,179 sequences was calculated under the General Time Reversible model (GTR) using the phylogenetic package HyPhy[338]. The distances were assessed according to the origin of the two sequences in the pairs: Africa Centre (AC:AC), South Africa (ZA:ZA) and mixed (AC:ZA) sequence

combinations. These calculations were used to determine the distribution/extent of genetic variation between and within populations.

7.2.3 Data linkage – demographic, epidemiological and clinical

Each of the AC sequences was linked to clinical and demographic data gathered by the demographic surveillance surveys in the area, as described in Chapter 3. The linkage was carried out using Stata 14 (Stata Corp LT, College Station, TX) with a unique individual identifier number assigned to each sample in order to pseudo-anonymise it. Sequence data were linked to the longitudinal demographic and health data at the time closest to the sample date of the sequence, unless otherwise stated. The available information included in this study is summarised in Box 7.1. No metadata were available for the ZA sequences.

Box 7.1: Demographic, epidemiological and clinical variables included in the analyses:

Demographic/Epidemiological	Geographical	Clinical
1. Age 2. Gender 3. Maximum educational level • No education • Primary school only • Secondary school • Tertiary education 4. Employment status • Full-time/ Part-time/ Unemployed 5. Marital status 6. Household wealth level* • Most deprived to least deprived quintiles based on local population[141,142]	1. Household dwelling setting • Urban • Periurban • Rural 2. HIV prevalence of residence* • High prevalence (>75th centile by year) • Medium (25th-75th) • Low (<25th) • External (outside DSA)	1. HIV testing information • Last negative • First positive 2. HIV treatment status Ever had ART treatment **Behavioural** 1. Ever reported multiple concurrent sexual partners 2. Condom use

*based on a wealth asset score as described in Chapter 3 (3.1.2a)

From the demographic data, it was possible to define other variables relevant to HIV acquisition risk. These included estimated seroconversion date (using mid-point imputation) and whether the household was located in high, medium, or low prevalence areas (see 3.1.2a).

7.2.4 Identification of transmission clusters

Putative transmission clusters were identified using the software 'ClusterPicker'. First, all phylogenetic clusters of at least two sequences, with a maximum genetic distance of 4.5%

and a SH-like local branch support of at least 70%, were extracted from the maximum likelihood phylogeny. This was repeated for all genetic distances between 1.5% and 4.5% at 0.5% intervals. The phylogeny of the putative clusters was reconstructed after removing the DRM sites (as above). Subsequent analyses were restricted to those clusters that remained monophyletic in all cases.

The putative clusters were annotated with metadata and analysis of cluster composition was undertaken, for example in terms of gender, population origins (AC versus ZA sequences), and geolocation of residential home (according to local prevalence data).

7.2.5 Baseline characteristics

Firstly, I assessed the baseline characteristics of individuals in this study. I produced summary statistics to include factors suggested to be predictors of HIV acquisition, e.g. gender, age, employment status, education, wealth, HIV prevalence of residence, circumcision, and number of sexual partners. I compared the characteristics of the overall sequence population to all HIV positive people in the AC DSA to investigate whether the sequence population is representative of the underlying population. I then went on to compare the characteristics of those sequences that cluster, compared to those that do not cluster. The populations are compared using tests of association - Chi-squared test for categorical, Kruskal Wallis test for ordinal, and t-test for continuous data. All statistical tests were undertaken in Stata 14. I also examined whether any of these variables were correlated using Pearson Correlation Coefficient ($p>0.8$ indicates correlation).

7.2.6 Further analysis for any large cluster

Further analysis was undertaken on any distinct large cluster that was identified (>20 sequences). This included reconstructing a dated phylogeny using BEAST v1.8.0[340], which allowed estimation of the evolutionary rate and determination of the time of the most common ancestor (tMRCA). BEAST bioinformatics software limits analysis to a maximum of 250-300 sequences; thus, given the large sample size, it was not possible to undertake this analysis for the whole sample. The criteria used to carry out this analysis in BEAST matched other HIV phylodynamic transmission work, i.e. an uncorrelated log-normal relaxed clock model, with the GTR + γ model of nucleotide substitution (6 rate categories to take into account that some positions evolve faster, e.g. third codon positions, than others), and the Bayesian skyline plot coalescent model[340-342]. The Bayesian Markov Chain

Monte Carlo (MCMC) analysis was set to 20 million generations, sampling every 10,000 steps. Convergence of the estimates was assessed by Effective Sampling Size (ESS) in Tracer[343], after removal of a 10% burning, with an ESS > 200 being deemed as satisfactory. BEAST2 v2.4.7 was then used to estimate the effective reproduction number (R_e) over time. R_e estimates the average number of secondary cases per infectious case in a population made up of both susceptible and non-susceptible hosts (i.e. the number of secondary infections produced by a typical infective case). This was calculated using an uncorrelated log-normal relaxed clock model, with the GTR + γ model of nucleotide substitution, and the Birth-Death Skyline Serial model.

7.2.7 Supplementary analysis

Finally, I investigated if, with combined epidemiological and phylogenetic data, it is possible to answer questions that are not easily answered by one modality alone. To do this, I addressed the question of directionality of infection. I annotated the clusters with all available linked metadata to determine further information about the direction of transmission and to see if the epidemiological metadata supported and/or enhanced the information from the sequence data.

To do this I used a smaller dataset from an initial pilot of this study, which used a subset of the sequences described above (i.e. 800 AC sequences rather than the 1,376 sequences described above from sample years 2010 and 2011 only, plus the same 803 ZA control sequences). All phylogenetic reconstruction and analysis methods undertaken were identical to those outlined above. I restricted this analysis to clusters identified at the 1.5% genetic distance level to try to target the identification of new infections, thereby reducing the possibility of un-observed third parties/intermediate transmitters. I excluded clusters in which there were no associated metadata for at least two sequences (including ZA sequences where the only metadata are the year of sampling). Of note is the fact that this sub-study was undertaken early in my PhD using the smaller dataset, which is all I had access to at that time. Given that I only received the larger dataset in the recent past, this analysis was not repeated on the larger dataset, but the opportunity to do this remains.

7.3 Results

A total of 2,179 HIV-1 partial *pol* sequences were genotyped and used in my analysis. 2,175 of the sequences were confirmed as subtype C, as expected. During the phylogenetic reconstruction there was the intentional addition of one subtype B sequence to allow rooting of the tree (outgroup rooting, Chapter 2). Three further recombinant sequences were identified, recombinant of HIV-1, group M, sub-types: D, A1 (AC); C, A1 (AC); and A1,C (ZA) (example shown in Appendix 7.1). Data presented in this study include previously published data from 2011 and 2012 AC surveillance rounds[145,146].

Metadata were only available for the AC sequences (n=1,376). All 1,376 AC sequences were linked to the demographic data (100% linkage). This represents a sampling coverage of 14.6% of the HIV positive population within the AC DSA in 2016 (n=9,405). However, if I only consider potential transmitters at this particular time, i.e. exclude those we know to be virally suppressed and, therefore, at very low risk or transmission (n=2,568), then this increases the sampling fraction of the transmitting population to 20.1%. This latter scenario is not an entirely realistic assumption, as viral load status can change rapidly and, although suppressed at a point in time, subjects may not have been supressed at a previous point in time and, therefore, may have contributed to the transmission chain.

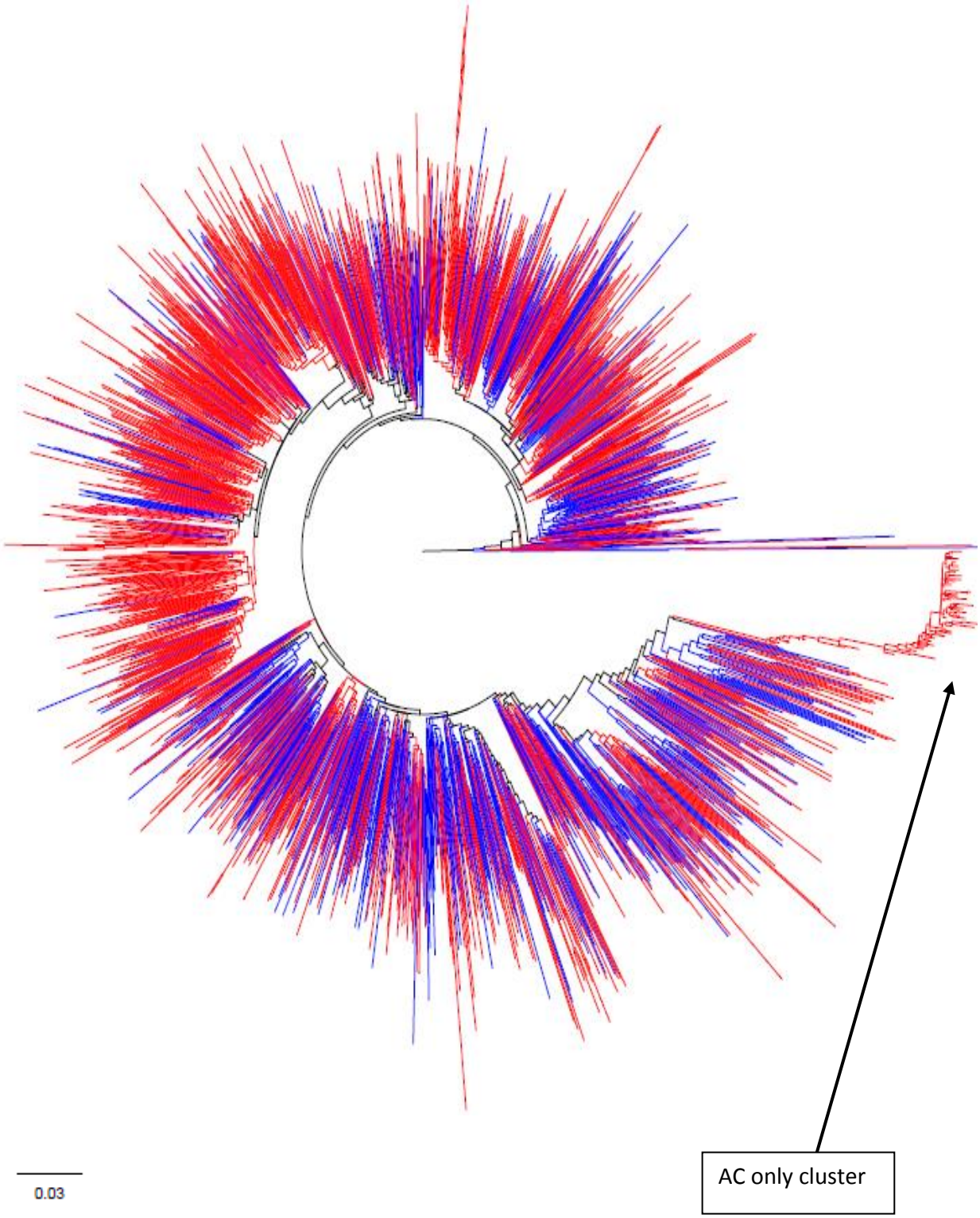
7.3.1 Phylogeny

A maximum likelihood phylogenetic tree of 2,179 sequences was reconstructed (1,376 AC sequences and 803 ZA sequences). The phylogeny produced is shown in Figure 7.1. AC sequences are coloured red and ZA sequences blue. AC sequences are largely interspersed throughout the tree with mixing between AC and ZA sequences, although some regions of the tree have a higher density of AC sequences than others.

The integration of AC and ZA sequences throughout the tree suggests substantial mixing between these populations. Furthermore, this suggests that the localised epidemic within the AC population is not an isolated outbreak. However, there is one example of a single monophyletic group of 75 AC only sequences (described in more detail below).

Figure 7.1: Maximum likelihood Phylogeny of 2,179 Africa Centre (AC) and South African (ZA) sequences. The tree is mid-point rooted. Branch support not shown for clarity. Red sequences=AC, Blue sequences= ZA sequences. Unit=nucleotide substitutions per site.

The tree shows AC and ZA sequences interspersed throughout the tree, except for one large AC only cluster (marked).



7.3.2 Genetic variability of sequences

I investigated the pairwise genetic distance as a proxy for genetic variability within a sequence population (e.g. AC or ZA) in order to assess whether the AC epidemic was an isolated, monophyletic outbreak (small mean population genetic distance,) or whether it arose as a result of multiple introductions from outside the DSA (larger mean population genetic distance, which might be similar to the ZA mean population genetic distance). The pairwise genetic distance between all possible pairs of the 2,179 sequences gave 2,372,931 combinations ($n!/r!\,(n-r)!$, where n=total number of sequences and r= number of sample points in each combination (i.e. two-way pairwise combinations)). Table 7.1 shows the summary statistics for the different genetic populations based on geographical origin of the sequences (AC vs ZA combinations).

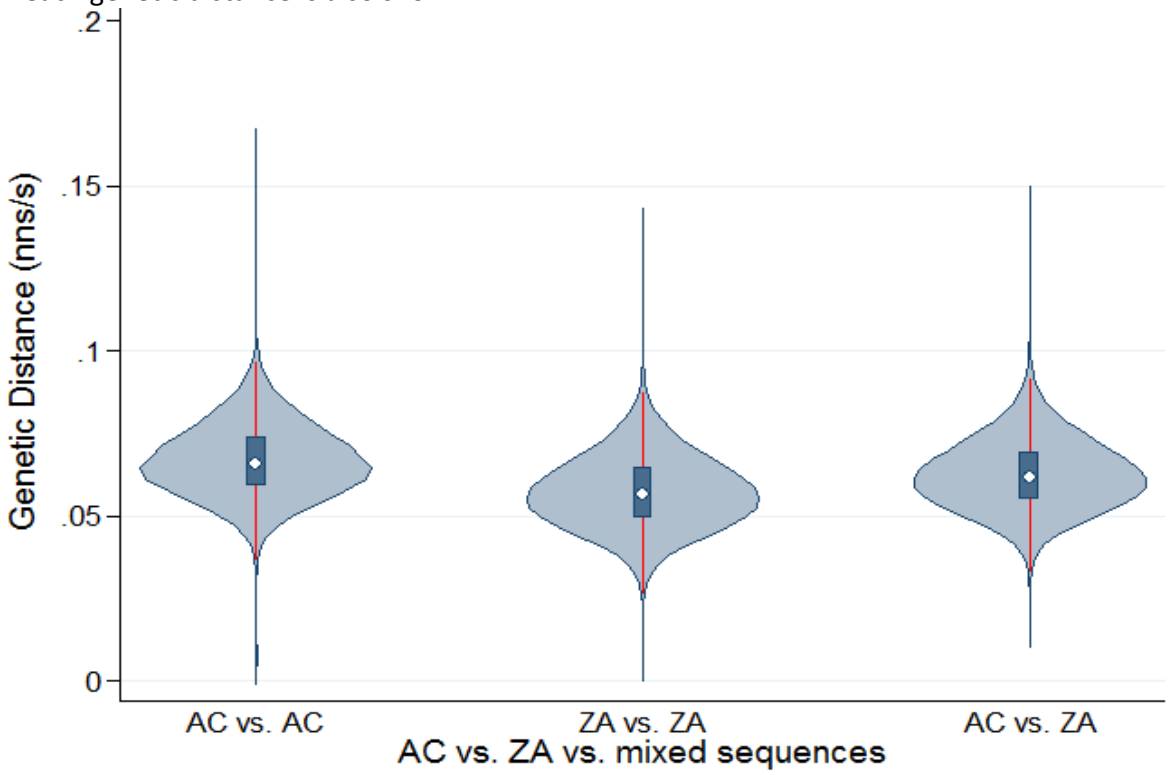
The violin plot demonstrates the population distribution of viral pairwise genetic distances for the three populations (Figure 7.2). It shows that the frequency is approximately normally distributed for all three populations, and that they overlap. The distributions appear to show that there is no difference in genetic relatedness between the populations, further confirming the high level of mixing in the area. This mixing, in all probability, leads to constant reintroduction and viral flow between populations, suggesting a role for migration in perpetuating the epidemic, and raising questions about transmission dynamics within certain sub-groups of the population e.g. migrators. Moreover, this might imply that the epidemic in the AC DSA population is not a self-sustaining standalone epidemic.

Table 7.1: Pairwise genetic distance characteristics between the three viral populations according to origin.

	AC vs AC	ZA vs ZA	AC vs ZA	Overall
No. of observations	946, 000 (39.9%)	322, 003 (13.6%)	1,104,928 (46.6%)	2,372,931 (100%)
Mean (nucleotide substitutions per site (nss))	0.066	0.057	0.063	0.063
SD (nss)	0.012	0.011	0.012	0.012
Range (nss)	0-0.166	0.001-0.142	0.011-0.150	0-0.166

AC vs AC is the pairwise genetic distance between all combinations of AC sequences, ZA vs ZA between all combinations of ZA sequences, and AC vs ZA between mixed pairs of one AC and one ZA. The lower the genetic distance the more similar the sequences.

Figure 7.2: Violin plots of genetic distance distribution between the three populations (AC vs AC; ZA vs ZA; AC vs ZA). White dot shows population mean, blue box is 95% CI of mean, red line shows IQR, and blue spoke shows total range. The probability distribution for each genetic distance is also shown.



Despite the overlap of the three frequency distributions, a t-test to compare the means of the populations provides evidence against the null hypothesis, suggesting that they are three separate populations ($p<0.01$ for all combinations). However, even though this might imply little interaction between the populations, given that this is a poor statistical test owing to the fact that the same observation contributes multiple data points (through pairwise combinations with all other sequences), the results should be interpreted with caution. Moreover, the statistical significance ($p<0.05$) is likely due, at least in part, to the very large number of observations, leading to a tiny difference between the populations being significant on statistical testing. Statistical significance does not always imply a meaningful difference in practice. In this case, a difference in the order of 0.01-0.003 substitutions per site, with visually very similar genetic distributions, does not represent a meaningful difference between the populations.

7.3.3 Linkage to epidemiological data

Metadata were available for all of the 1,376 sequences from AC and linkage was achieved for all samples (100%).

7.3.4 Characteristics of Clusters

I identified all putative transmission clusters within the phylogenetic tree using different genetic distances to define a 'cluster'. A total of 878/2,179 (40.3%) HIV-1 subtype C sequences were linked to at least one other sequence in the database, forming 327 putative transmission clusters (maximum intra-cluster genetic distance 4.5% and branch support >70%). The number of putative transmission clusters decreased with decreasing genetic distance to 91 putative clusters involving 221 sequences at a genetic distance of 1.5% (branch support 70%). Figure 7.3 and Table 7.2 show a summary of the cluster characteristics. These both show that varying the thresholds affects the size and distribution of phylogenetic clusters.

The clusters identified were scattered throughout the phylogenetic tree. The 59.7% of sequences with no linkage to other sequences in the database are indicative of infections acquired elsewhere, or from individuals whose partners were not diagnosed, were not included in the database, or had too distant a connection to be identified.

A discrete large cluster emerged at a genetic distance of 1.8%, suggesting recent transmission (with 59 sequences). It rises to a cluster size of 64 sequences at a genetic distance of 3.5%, before reaching 75 sequences at genetic distance of 4.0%. This cluster is described in detail below. Other than this large cluster, cluster sizes were small, ranging from two to eight sequences, which may reflect lack of sampling of recent infections.

Figure 7.3: Number of putative transmission clusters by genetic distance (branch support >70%)

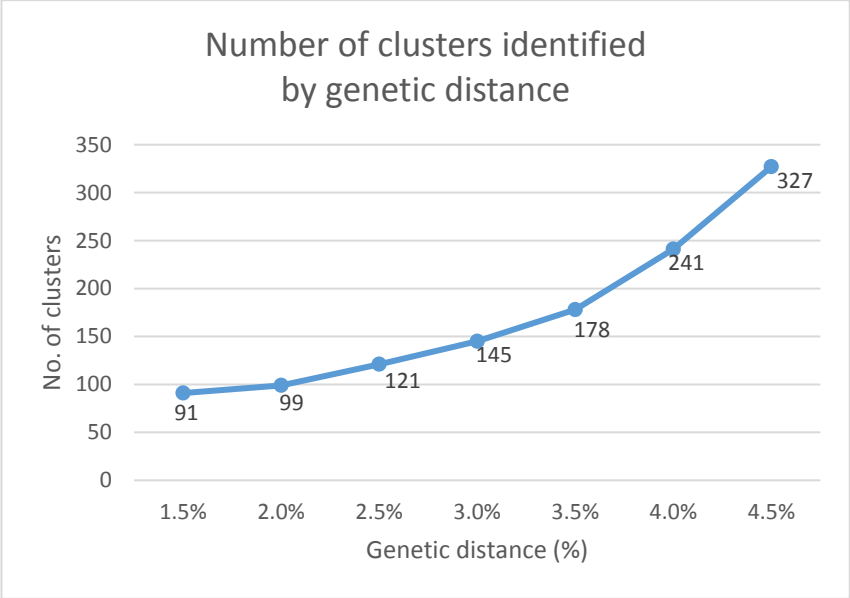


Table 7.2 Putative cluster characteristics

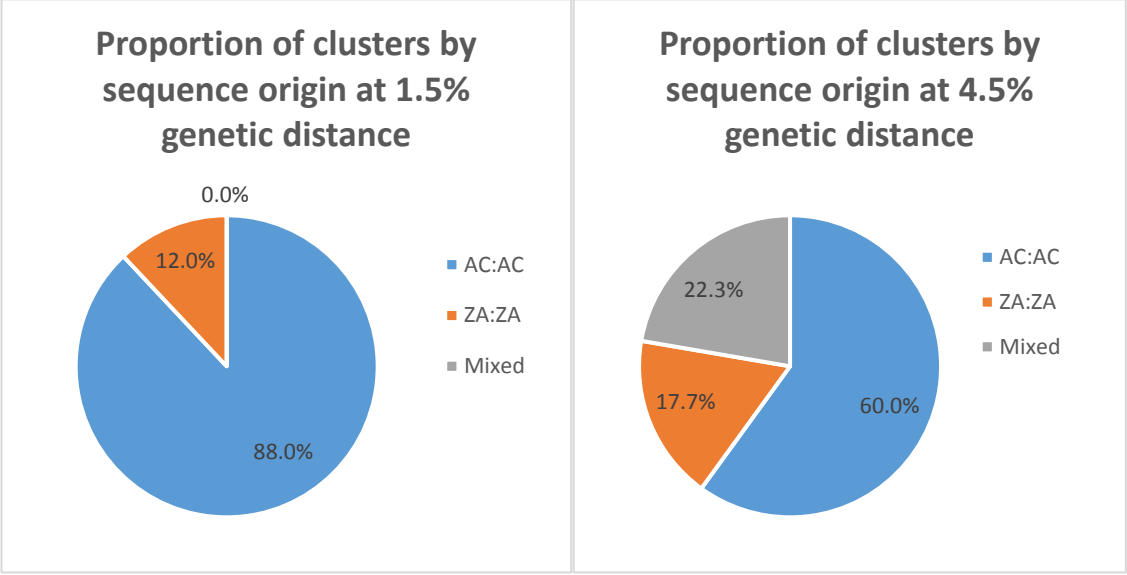
Cluster characteristics	1.5% GD	4.5% GD
Cluster size = 2	77%	71%
Cluster size = 3	12%	19%
Cluster size = 4	8%	6%
Cluster size = 5	1%	2%
Cluster size = 6	0%	1%
Cluster size = 7	1%	0%
Cluster size = 8	1%	1%
Max. cluster size	8	75
No. of clusters	91	327
Sequences clustering	221	878

Within the putative transmission clusters identified, I went on to assess the characteristics and traits of participants. I used a genetic distance of 1.5% to do this, as this is most likely to identify recent infections and true linked pairs given the variation between the two sequences was minimal. At this genetic distance the largest cluster size was 8, with 77% of clusters being pairs and 12% triplets. Assessing the geographical origin of these clusters, 88% were comprised of sequences from AC only and 12% from ZA only (Figure 7.4). There were no mixed clusters between AC and ZA sequences. This might be explained by the higher proportion of AC sequences in this study group, and by the fact that they are more densely sampled. In contrast, a genetic distance of 4.5% gives a broader picture of transmission, but with decreased certainty. The clusters identified by the 4.5% distance were likely to include older infections (i.e. a longer time between infection and sampling)

and to be incomplete. The largest cluster size was the big cluster of 75, and the second largest cluster size was 8, with 71.2% of clusters being pairs and 19.2% being triplets. The composition in terms of sequence origin changed as follows: 60% were AC only clusters, 17.6% were ZA only clusters, and 22.3% were mixed clusters.

With the exception of the single large cluster, by relaxing the cluster definition threshold criteria from a genetic distance of 1.5% to 4.5%, the cluster size (i.e. proportion of sequences clustering in pairs, triplets etc.) did not appear to change (Table 7.2). However, using the relaxed criteria, there were more mixed and ZA:ZA clusters. This might be expected as the relaxed criteria include older infections, given that the ZA sequences are from a different year range and historically older than the AC sequences. Furthermore, the sampling coverage from outside the AC DSA is much lower in this cohort and, therefore, it is likely that a higher proportion of ZA sequences were not sampled and are missing in this population. This could lead to larger genetic distances within clusters and, therefore, these were only identified with the more relaxed criteria.

Figures 7.4 A&B: Pie charts to show proportion of transmission within putative clusters between different sequence populations (AC only, ZA only, and mixed clusters) at different genetic distances (GD) levels (A – 1.5% GD and B – 4.5% GD).



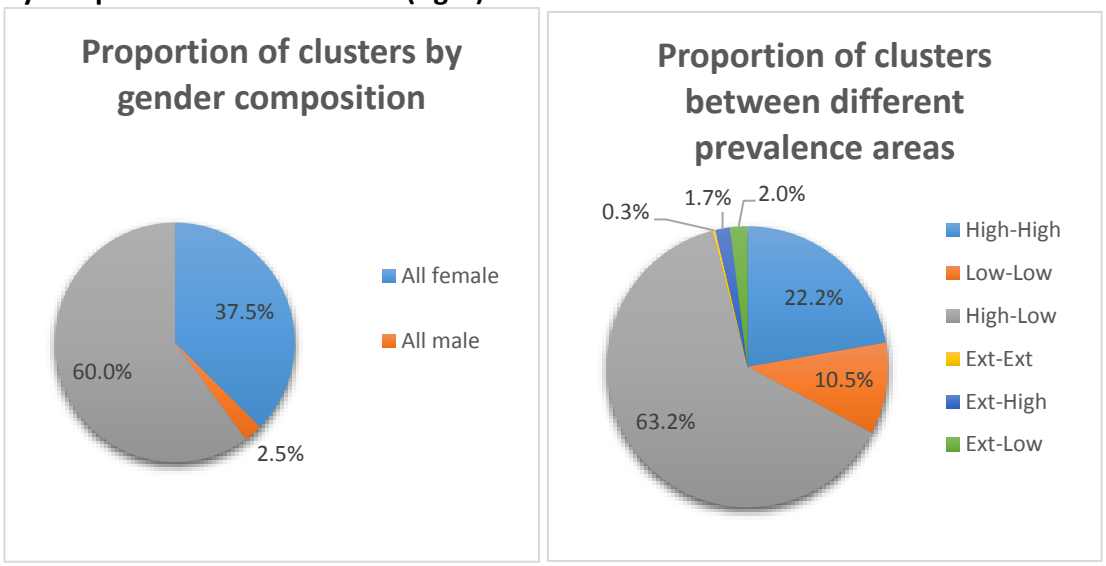
As this analysis was targeted at general transmission dynamics, and not only at recent infections, the remainder of the analysis was restricted to the 327 clusters identified at 4.5% genetic distance. There were 878 individuals within these 327 clusters, of whom 669 were AC sequences and 209 ZA sequences. This further analysis was restricted to AC-only

clusters where linked metadata were available. As I had no metadata for the ZA sequences, the 71 AC sequences that linked only to ZA sequences were excluded, as I could not determine anything meaningful about the cluster composition with metadata concerning only one sequence in the cluster. Therefore, for this section of the analysis n=598 AC sequences.

In addition, I investigated the proportion of linked cases between genders (Figure 7.5a). Most linked cases were of mixed gender (60%). However, 37.5% of clusters were composed of females only (80 pairs, 18 triplets, 2 quadruplets, 2 quintuplets, and one cluster each of six and eight females). These female-only clusters are most likely to indicate the presence of one or several unobserved intermediate male(s) and this may also reflect the low participation rates, particularly in men, observed in the sampling strategy, resulting in the inclusion of a higher proportion of females in the sequenced cohort (70% vs 30%). The proportion of male-only clusters was low at 2.5% (six pairs of males and one triplet cluster). This could represent homosexual transmission, but is more likely to indicate an intermediate unsampled female partner.

Finally, I went on to investigate transmission between different geographical areas determined by HIV prevalence (defined in Box 7.1 and explained in Chapter 3). Two hypotheses I tested were: firstly, that the epidemic within the AC DSA is continually seeded from outside the area; and/or, related to the first hypothesis, that transmission within the AC DSA might see residents of higher prevalence areas seeding, and thus sustaining, infections within the lower prevalence areas. Unfortunately, it was difficult to draw any meaningful conclusions regarding the first hypothesis as the proportion of clusters which include sequences from people not resident within the DSA (external sequences) was very low, being 4% of the total. With respect to the second hypothesis, although there is evidence of transmission within high prevalence areas occurring in 22.2% of clusters and transmission within low prevalence areas in 10.5% of the clusters (Figure 7.5b), it is not possible to tell anything about the direction of transmission from these clusters. Therefore, I could not determine whether or not the high prevalence areas were seeding the lower prevalence areas. Given the higher proportion of High-High and Low-Low clusters observed, as opposed to clusters linking to sequences in areas of differential prevalence (High-Low or vice versa) or to external sequences, this might suggest that the majority of transmissions within the AC DSA are actually local.

Figures 7.5 A&B- A: Composition of cluster by gender (left), and B: Cluster composition by HIV prevalence of residence (right). NB Clusters identified at 1.5% GD



*Female-to-Female does not imply female homosexual transmission, but is likely to indicate an unsampled male in the transmission chain.

7.3.5 Baseline characteristics and epidemiology

Table 7.3 highlights the characteristics of the 1,376 AC sequences with respect to baseline demographic characteristics, sexual behaviour traits, and ART exposure. The characteristics within different sub-groups, represented within the sequenced population, were then compared - sequences that cluster compared to those that do not cluster, and the characteristics of the large cluster compared to all other sequences are outlined. Furthermore, this sequence population is compared to the baseline characteristics of all HIV positive individuals within the AC population. I also undertook a sensitivity analysis to compare this population to a control group: all known individuals in the AC population who tested negative, as a proxy for those at risk of HIV acquisition (results presented in Appendix 7.2). The Pearson Correlation Coefficient did not show any correlations between the variables (data shown in Appendix 7.3).

Of the overall sequenced population, approximately 70% (952) were females and approximately 30% (424) males. This is similar to the participation rates reported for AC surveillance surveys. The mean age was 33.5 years (SD 11.7). Residential geolocation data were available for 1,321 cases. 30.3% of sequences were from those resident in high prevalence areas, 48.4% from medium, and 16.1% from low prevalence areas.

Table 7.3: Baseline characteristics of sequence population compared to sub-groups and controls

	All Sequences	Sequences that cluster	Non-Cluster sequences	p-value: Clustered vs Non-clustered	Big cluster	p value: All vs Big cluster	All Positive	p-value: All vs Pos
Eligible individuals	1,376	669	707		75		11,912	
BASIC DEMOGRAPHIC INFORMATION								
Gender								
Male (%)	30	29.3	30.9		57.3		26.7	
Female (%)	70	70.7	69.1	p=0.24	42.7	**p<0.001**	73.3	**p<0.001**
Age (years)								
Mean (years)	33.5	32.7	34.2		35.0		36.8	
SD	11.7	11.7	11.7	**p=0.04**	8.4	**p<0.001**	12.1	**p<0.001**
<20 (%)	8.3	9.6	7.2		2.7		3.7	
20-24 (%)	17.1	18.4	16.0		5.3		11.1	
25-29 (%)	18.5	20.0	17.2		18.7		16.6	
30-34 (%)	15.4	16.4	14.8		25.3		17.7	
35-39 (%)	14.0	10.8	16.7		22.7		14.5	
40-44 (%)	9.7	7.8	11.2		9.3		11.5	
45-49 (%)	6.3	6.4	6.2		10.7		8.9	
>50 (%)	10.7	10.6	10.7	p=0.05	5.3	**p<0.001**	16.0	**p<0.001**
Wealth quintiles								
Most deprived (%)	18.6	17.9	19.2		21.3		18.8	
2nd most deprived (%)	19.9	19.3	20.5		13.3		19.8	
Middle quintile (%)	21.8	21.7	22.0		18.7		19.7	
2nd least deprived (%)	20.5	21.5	19.7		20.0		18.5	
Least deprived (%)	14.1	15.7	12.7	**p=0.03**	24.0	**p<0.001**	15.2	**p=0.47**
Missing (%)	5.1	3.9	5.9		2.7		8.0	
Maximum Education								
None (%)	5.3	5.7	4.9		4.0		6.3	
Primary (year 1-7)(%)	31.9	33.6	30.4		29.3		25.5	
Secondary (8-12)(%)	43.6	42.3	44.7		50.7		36.8	
Tertiary (>12)(%)	0.9	1.1	0.8	p=0.41	2.7	**p=0.04**	1.2	**p<0.001**
Missing(%)	18.3	17.3	19.2		13.3		30.2	
In employment								
Yes, FT (%)	17.7	17.6	17.7		45.3		24.4	
Yes, PT (%)	7.9	7.9	7.9		4.0		4.3	
No (%)	51.4	53.1	50.0	p=0.78	40.0	**p<0.001**	49.4	p=0.06
Missing (%)	23.0	21.3	24.4		10.6		21.9	
Marital Status								
Never married (%)	72.1	75.6	69.1		66.7		80.1	
Married (%)	7.3	7.0	6.8		9.3		3.8	
Engaged (%)	6.9	8.2	6.4		10.7		0.2	
Divorced/Separated/ Widowed (%)	4.0	3.7	4.4	p=0.48	0.0	p=0.1	0.4	**p<0.001**
Missing (%)	9.7	5.4	13.3		13.3		15.5	
Residence HIV prevalence								
High (%)	30.3	32.3	28.5		21.3		28.4	
Mid (%)	48.8	46.8	50.6		49.3		47.5	
Low (%)	16.1	17.0	15.3		28.0		18.3	
External (%)	4.7	3.9	5.3	p=0.97	1.3		0.0	
Missing (%)	0.1	0.0	0.3		0.0	p=0.05	5.8	**p<0.001**
HIV prev median	25.9	25.7	25.9		23.1		25.3	
HIV prevalence IQR	22.1-30.8	22-31.5	22.4-30.3	p=0.81	18.5-27.2	p=0.06	20.9-30.2	**p<0.001**

Fig 7.3 continued

		All Sequences	Sequences that cluster	Non-Cluster sequences	p-value C vs N-C	Big cluster	p value: All vs Big cluster	All Positive	p-value: All vs Pos
Eligible individuals		1,376	669	707		75		11,912	
CLINCIAL VARIABLES									
ART treatment									
	Yes (%)	13.8	8.3	18.4		9.3		33.6	
	No (%)	36.1	37.4	35.1	**p=<0.001**	20.0	p=0.13	15.3	**p<0.001**
	Missing (%)	50.1	54.3	46.5		70.7		51.2	
SEXUAL BEHAVIOUR									
Multiple partners ever reported									
	Yes (%)	9.5	10.0	9.0		12.0		7.3	
	No (%)	67.2	70.3	64.7	p=0.87	77.3	p=0.7	33.6	**p<0.001**
	Missing (%)	23.3	19.7	26.3		10.7		59.1	
Circumcised	Men only								
	Yes (%)	8.8	10.8	7.3		14.3		8.3	
	No (%)	79.5	78.5	80.4	p=0.22	69.0	p=0.14	45.3	p=0.11
	Missing (%)	11.7	10.7	12.3		16.7		45.8	
Condom use									
	Always (%)	12.9	12.3	13.4		2.7		5.6	
	Sometimes (%)	20.2	24.2	16.8		57.3		9.1	
	Never (%)	14.0	15.8	12.5	p=0.13	18.7	p=0.06	5.5	**p<0.001**
	Missing (%)	52.9	47.7	57.3		21.3		79.8	

It is likely that the highly significant findings are due to the very large sample size, but values shown for completeness.

With respect to risk factors, the clustered population differed from the non-clustered population across three demographic and behavioural variables: the clustered population was slightly younger (32.7 vs 34.2 years, p=0.004), wealthier (p=0.03), and were taking less ART (8.3% vs 18.4%, p<0.001) than the un-clustered population. The ART figures were extremely low for both clustered and non-clustered, and might indicate a bias towards the sequencing of new infections, undiagnosed infections, or those failing treatment.

Comparing the control populations to the overall sequenced population, the group of all HIV positive individuals within the DSA appeared to be statistically different to the sequenced population across many of the variables tested. This might be expected given the very large sample size for this population (n=11,912), meaning that a small difference would lead to a significant change. Overall, the HIV positive population within the DSA had a slightly higher proportion of females (73.3% vs 70%), an older population (36.8 vs 33.5 years), was from a more deprived wealth quintile, had less education, and a larger number had reported multiple sexual partners in a prior 12 month period (p<0.001 for all variables), than the overall sequenced population.

7.3.6 LARGE CLUSTER ANALYSIS

a) Characteristics of the large cluster

The baseline characteristics of the large cluster (n=75) were significantly different to those of all other sequences in a number of ways (see Table 7.3 above). There was a significantly higher proportion of males compared to females (57.3% vs 42.7%, p<0.001), and although there was a statistical difference in age, it was so small that it is not practically meaningful (34.2 vs 33.5 years, p<0.001). The large cluster had more people in the two extremes of the wealth quintiles, particularly the most wealthy quintile (32.9% vs. 14.1%, p<0.001), with lower proportions in the middle categories. This was partly driven by gender, with males disproportionately falling into the most wealthy quintile (39.5% of males, compared to 21.9% of females). Furthermore, a significantly increased number were in full-time employment in this large cluster compared to the overall sequenced population (45.3% vs. 17.6%, p<0.001) and they were better educated, with a higher proportion having secondary and tertiary education as their maximal educational attainment (p=0.04). Again, the employment rates were higher for males in the large cluster (53.5%, compared to 34.3% in females). It appears that a higher proportion did not answer the question regarding ART treatment exposure (70.7% vs. 53.6%), which is to be expected for individuals who do not know they are HIV positive, as they do not consider it relevant. This might suggest a higher proportion of new/undiagnosed infections in this cluster.

b) Phylogenetic analysis

The distinct AC-only large cluster in the phylogenetic tree (highlighted in Figure 7.1) was very different to any other region of the tree due to the cluster size and the shorter intra-cluster branch lengths compared to other clusters in the tree. Therefore, it warranted further investigation. Firstly, I undertook several tests to ensure that this finding was genuine, that my results were robust and to exclude hypotheses of an artefactual result, technical errors or sample contamination. My investigations included the following:

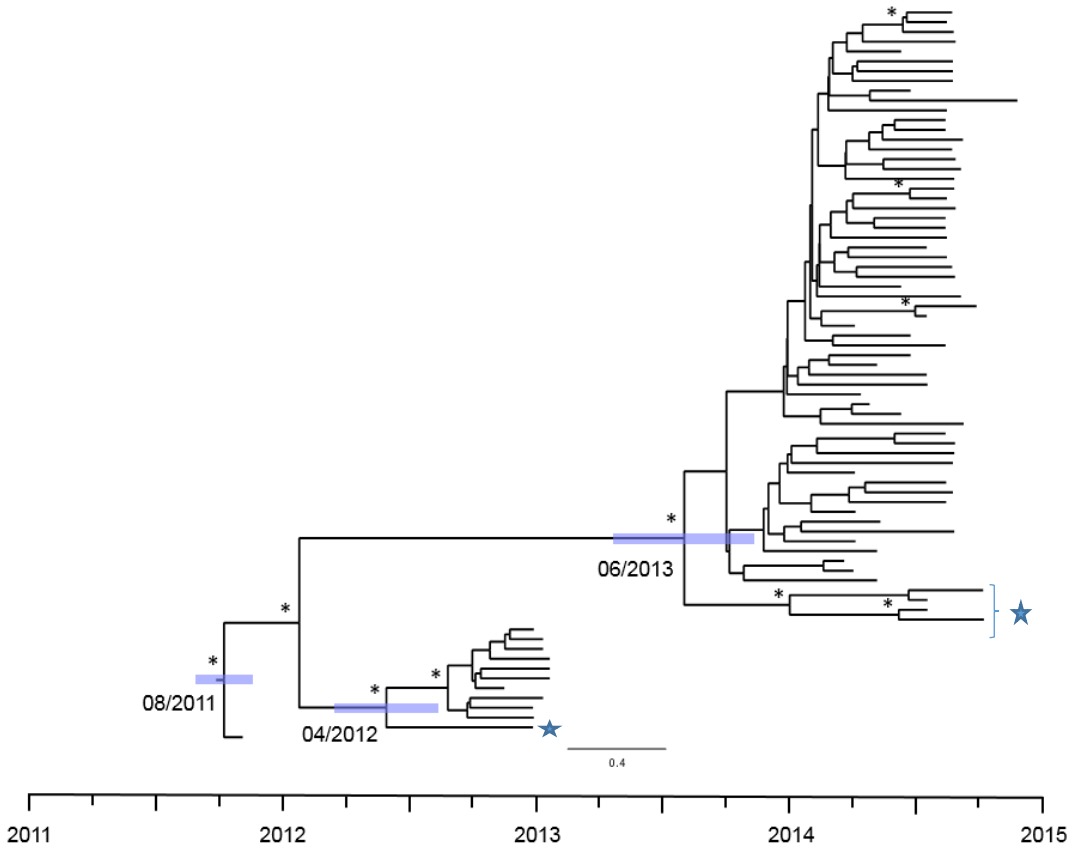
1. Rechecking the alignment of these sequences ;
2. Rerunning the tree, again using RAxML[327], as well as two other phylogenetic software tools – FastTree v2.1.5[329] and PhyML v3[344] - to ensure replication of finding, and to exclude technical error;
3. Investigating branch length and genetic distances to exclude contamination – none of the sequences were identical, which would have been the case had there been laboratory contamination or mislabelling;

4. Rechecking that the drug resistant mutation sites had been correctly deleted to prevent artefactual clustering from acquired rather than inherited mutations;
5. Rerunning the cluster with recent controls from surrounding Southern African countries – for each of the 75 sequences in the cluster, I used the Basic Local Alignment Search Tool (BLAST)[345] to find the 10 closest publicly available sequences. We restricted the sequences identified to human primary clinical samples and to countries neighbouring South Africa. After removing duplicates, I reran a phylogenetic analysis with the 75 AC sequences along with 56 newly-identified closely related Southern African sequences. The tree topology still showed a distinct cluster of 75 sequences from AC only, with no other sequences interspersed.

All my testing replicated the initial result, with an identical large cluster of sequences from the AC only. This confirmed that the cluster identified was genuine and that the result was robust.

In order to further investigate this large cluster, I constructed a dated phylogeny using BEAST 2[340] to enable an estimation of the tMRCA. Figure 7.6 presents the dated phylogeny showing that this cluster emerged from a common single ancestral source in approximately August 2011 (confidence intervals between July and September). However, there appear to be two distinct (non-overlapping) outbursts of infections within this cluster – one in 2012 and one in 2013. The 2014 sampled infections (a larger cluster emerging from July 2013) are clearly separated from the older ones (smaller cluster emerging in April 2012), suggesting a more recent surge in transmission, perhaps initiated by a single individual. The boxes around the emergence dates show the 95% confidence interval window for the emergence event (date= mean estimate). I am mindful that some caution is needed in interpreting this result, as no sequences were available from 2013 sampling at AC, and this could lead to sampling bias.

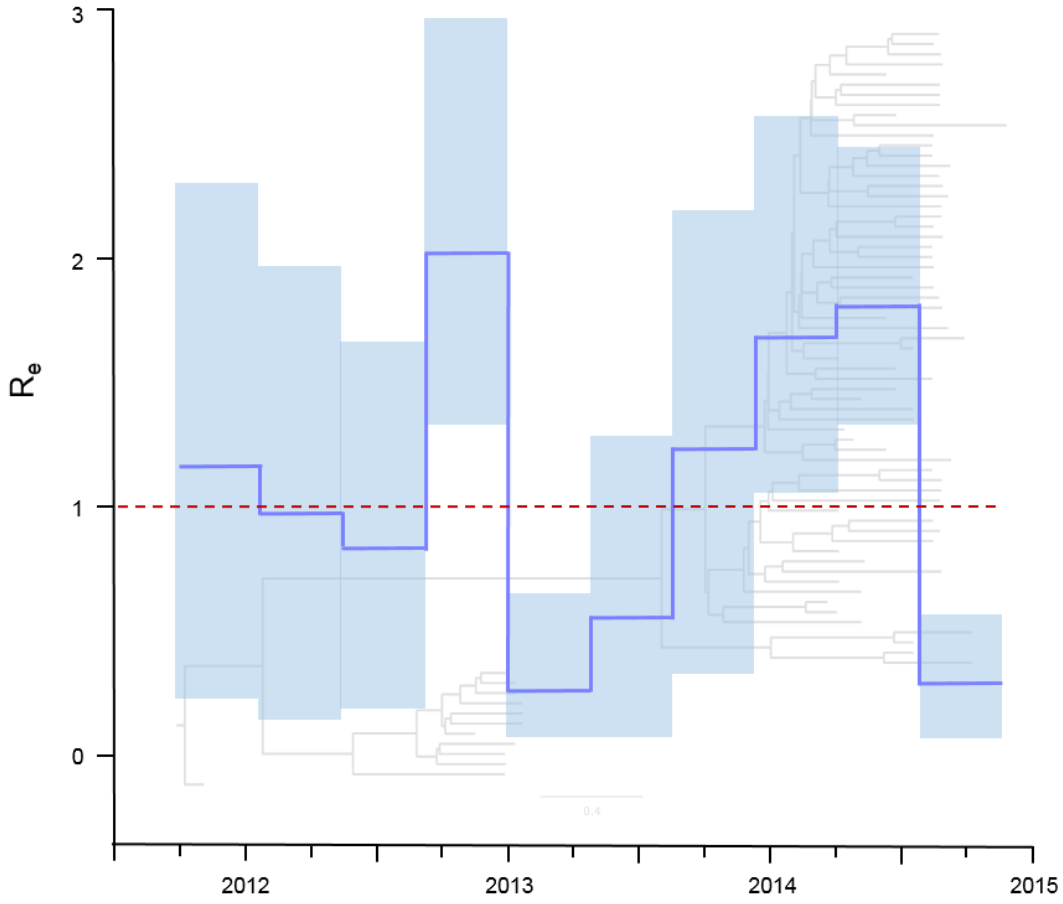
Figure 7.6: Dated phylogeny of large cluster



Legend: Bayesian data phylogeny of large cluster. Mean date of node to within 0.4 of a year. Bars = 95% CI for emergence event. * Posterior Probability (branch support) > 0.90. **Potential source as earliest identifiable case within cluster. ★ = Sequence/s closest to the origin of the outbreak.

In addition, Figure 7.7 shows estimates of the reproduction number (Re) of this cluster over time, overlaid onto the phylogenetic tree (dark blue line: mean estimate; light blue boxes: confidence intervals). This shows an increase in R_e to above 1 corresponding to both outbursts of infection in 2012 and 2013. There is a sudden drop-off at the end of 2012 suggesting that this outburst died out or there was an intervention dramatically reducing ongoing transmissions. However, I am not aware of any new intervention being undertaken in this area at this time. A second increase occurs in late 2013 to R_e >2. This increase corresponds to an observed increase in incidence. The apparent drop-off at the end of 2014/2015 is likely to be artefactual, as I did not access to sequences sampled after that point in time. This lack of data leads to the default of assuming no further infections, which is not necessarily true. Of note is the fact that there were no sequences available for the 2013 year of sampling. It is hard to predict where these sequences would fall within the phylogeny, although it is unlikely to change the overall conclusions drawn or challenge the shape of the tree.

Figure 7.7: Plot of the reproduction number over time, overlaid onto the phylogeny



Legend: boxes =CI, line is mean.

c) Epidemiological analysis

One hypothesis is that this emerging cluster is related to a new mine within the DSA, which opened in 2008. The Somkele mine is an open cast coal mine, which employs 989 people. Most of the miners live locally within the DSA (80%) and the mine provides transport for local employees.

To explore this hypothesis further, we mapped the residences of the 75 individuals whose samples made up the large cluster, which revealed that the sequences group into three main geographies (Figure 7.8): one group around the main road in the previously identified 'hotspot' region, another group in the centre of the DSA immediately adjacent to the mine site, and then several small groups in rural areas. Previous HIV incidence (and prevalence) rates in the area adjacent to the mine were low (in 2009), so it is unusual to see so many new infections - 33 of the sequences (almost 50%) are from homes immediately adjacent to the mine. Epidemiological research undertaken by the AC earlier this year suggested

new incident infections in this area, with a greatly increased hazard risk (HR 16) of contracting HIV[131]. However, the extent of the outbreak and evidence to suggest a link to the mine was not identified by this work.

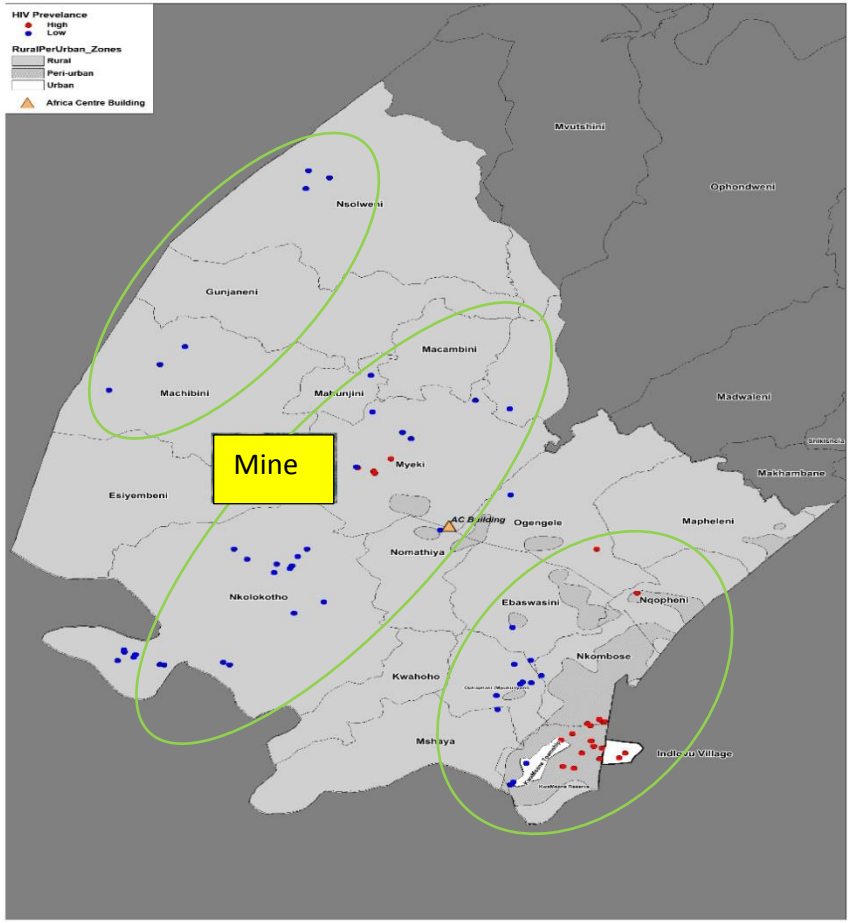
Furthermore, linking the sequences to epidemiological data also provides evidence to support my hypothesis that this emergent cluster is related to the new mining community. The baseline demographics of this cluster are included in Table 7.2 and discussed above (section 7.3.6a). The sequences in this large cluster are systematically different to the rest of the sequences in this study. The key difference is the high proportion of males in this large cluster (57.3%). This finding contrasts with what one might expect given the baseline characteristics of the AC population: the sequenced population contains only 30% males which, as previously discussed, is likely due to differential participation rates; and HIV infections, within AC DSA, are disproportionately seen in women (approximate ratio 2:1). These differential characteristics are compatible with the large cluster being related to the mine's male workforce (see discussion). Additional circumstantial evidence concerning the large cluster, which supports my hypothesis includes: that the range of ages is compatible with the ages of mineworkers; that this cohort has a far higher rate of employment than the rest of the sequence population (17.6% vs 45.3%); and that they are better educated (which may be required to be employed in the mine) and overall more wealthy.

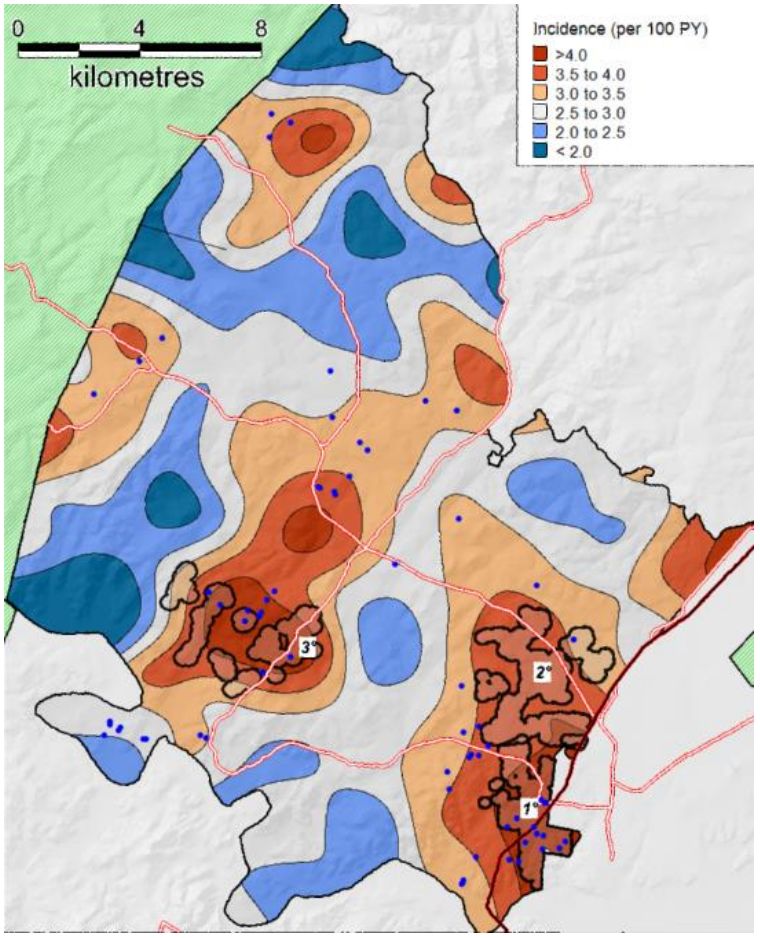
From the dated phylogeny, it is possible to identify the closest sampled sequence/s to the origin case of this large cluster (labelled in Figure 7.6). This sequence might not itself be the true origin, but, nonetheless, it may provide information about the origins and mechanisms of transmission. Investigating these earliest sequences identified one sequence from a male in his early 30s who lives near the mine in a peri-urban region. Most notably he reported >15 sexual partners in the 12 months prior to his participation in the 2013 surveillance round, which coincides with the timing observed in the growth of the large cluster (Figure 7.6). This is the sort of high-risk behaviour that could explain the rapid emergence of a large outbreak, although this inference remains speculative. In addition, this high-risk behaviour might provide a link to the sex worker community, which could further propagate the outbreak.

Figures 7.8 (a) & (b): Maps geolocating the large cluster sequences

Both maps incorporate a small random error into geographical position of each participant, but do give an accurate distribution.

7.8a) The clusters are also plotted showing infections relative to high (red dots) or low (blue dots) prevalence areas and type of area (urban/periruban/rural). This map also shows the three geographical 'groups' of infection: scattered rural cases (west border of DSA), around mine (central group), and peri-urban/urban group along the main road. **7.8 b)** The cluster of 75 infections superimposed onto the incidence map from 2017, showing the areas of high and low incidence[131].





7.3.7 Investigation of direction of transmission within clusters

Finally, I investigated whether it is possible to determine the direction of transmission within small clusters when combining epidemiological and phylogenetic data.
To do this I used the 11 AC-only clusters of 23 individuals, identified at a genetic distance of 1.5% in our pilot study dataset. Figure 7.9 shows the clusters along with relevant linked metadata. I chose to exclude cluster 8 from the analysis as this sequence did not link to full metadata.

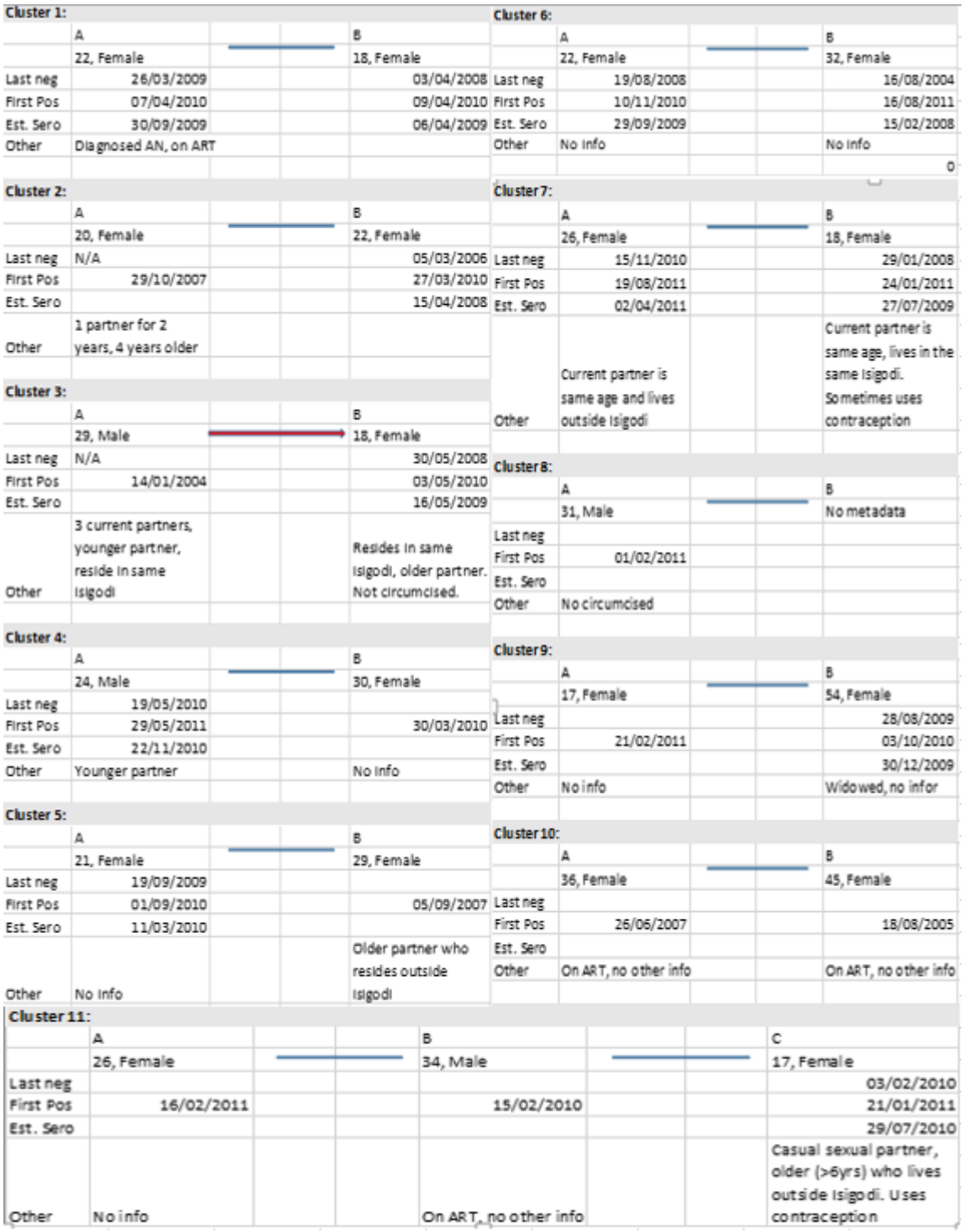
Of the 10 clusters, one (cluster 3) contains sufficient data to indicate the direction of transmission, as shown by the dates set out in Figure 7.9 below. In addition, this increases the probability of this being a true direct transmission event. Unfortunately, the other clusters mostly appeared to be missing either data or unsampled intermediates, meaning that none gave clear indications of direction of transmission.

This work demonstrates that by combining molecular and epidemiological analysis it is possible to enhance one's understanding of transmission dynamics. Although this was only the case in one cluster out of 10, the result is likely to be an underrepresentation of the full potential of this method, due to the small sample size, the sampling strategy used, and the fact that so many clusters were links between females with a missing male as the probable connecting link.

In order to infer the direction of infection with any degree of certainty from consensus sequence data, linked epidemiological data are required. The key pieces of epidemiological data in this example were the estimated date of seroconversion and the window period between the last negative and first positive tests. Other epidemiological data provided supporting evidence once the seroconversion date had indicated a likely direction: partner age, circumcision status, and geographical location all provided supporting evidence.

In terms of drawing more general conclusions about the utility of epidemiological data in enhancing molecular analysis, it is reasonable to expect that seroconversion date, when indicated within a narrow window between last negative and first positive test, is the most likely piece of data to allow direction to be inferred. However, it is also important to remember that there is an underlying assumption that the data are accurate, and this is not always the case, particularly with respect to sexual behavioural data.

Figure 7.9: Annotated transmission clusters



Cluster 1:

	A	B
	22, Female	18, Female
Last neg	26/03/2009	03/04/2008
First Pos	07/04/2010	09/04/2010
Est. Sero	30/09/2009	06/04/2009
Other	Diagnosed AN, on ART	

Cluster 2:

	A	B
	20, Female	22, Female
Last neg	N/A	05/03/2006
First Pos	29/10/2007	27/03/2010
Est. Sero		15/04/2008
Other	1 partner for 2 years, 4 years older	

Cluster 3:

	A	B
	29, Male	18, Female
Last neg	N/A	30/05/2008
First Pos	14/01/2004	03/05/2010
Est. Sero		16/05/2009
Other	3 current partners, younger partner, reside in same isigodi	Resides in same isigodi, older partner. Not circumcised.

Cluster 4:

	A	B
	24, Male	30, Female
Last neg	19/05/2010	
First Pos	29/05/2011	30/03/2010
Est. Sero	22/11/2010	
Other	Younger partner	No info

Cluster 5:

	A	B
	21, Female	29, Female
Last neg	19/09/2009	
First Pos	01/09/2010	05/09/2007
Est. Sero	11/03/2010	
Other	No info	Older partner who resides outside isigodi

Cluster 6:

	A	B
	22, Female	32, Female
Last neg	19/08/2008	16/08/2004
First Pos	10/11/2010	16/08/2011
Est. Sero	29/09/2009	15/02/2008
Other	No info	No info

Cluster 7:

	A	B
	26, Female	18, Female
Last neg	15/11/2010	29/01/2008
First Pos	19/08/2011	24/01/2011
Est. Sero	02/04/2011	27/07/2009
Other	Current partner is same age and lives outside Isigodi	Current partner is same age, lives in the same Isigodi. Sometimes uses contraception

Cluster 8:

	A	B
	31, Male	No metadata
Last neg		
First Pos	01/02/2011	
Est. Sero		
Other	No circumcised	

Cluster 9:

	A	B
	17, Female	54, Female
Last neg		28/08/2009
First Pos	21/02/2011	03/10/2010
Est. Sero		30/12/2009
Other	No info	Widowed, no infor

Cluster 10:

	A	B
	36, Female	45, Female
Last neg		
First Pos	26/06/2007	18/08/2005
Est. Sero		
Other	On ART, no other info	On ART, no other info

Cluster 11:

	A	B	C
	26, Female	34, Male	17, Female
Last neg			03/02/2010
First Pos	16/02/2011	15/02/2010	21/01/2011
Est. Sero			29/07/2010
Other	No info	On ART, no other info	Casual sexual partner, older (>6yrs) who lives outside Isigodi. Uses contraception

7.4 Discussion

These analyses have demonstrated the value of combining epidemiological data with molecular data in three key areas. Firstly, it enables a high-level understanding of the AC epidemic in the context of the generalised South African epidemic. This has shown that there is substantial mixing between the AC and ZA populations; i.e. that the AC epidemic is not an isolated epidemic, but is propagated by frequent reintroductions of the virus from other ZA locations. Secondly, it has allowed a better understanding of the dynamics of transmission events within the AC DSA, such the probable role of the new Somkele mine in propagating new infections within the AC DSA. In doing so, it has revealed a new emergent cluster[§§§], which provides essential knowledge that can help public health authorities to minimise its impact. To my knowledge, this is one of the first examples of molecular methods alone identifying a new cluster in an existing outbreak. Finally, it has shown that sequence data and epidemiological data together can be more effective than either method alone when analysing the details of transmission dynamics. Two examples of this are the increased weight of evidence provided by both methods in combination to confirm the likelihood of linked infections and the identification of likely direction of transmission, as described above. Each of these three key areas are discussed in more detail below, starting with the emergent cluster.

This newly identified large cluster, henceforth ”New Cluster”, identified by my phylogenetic analysis was an unexpected result. On detection of the New Cluster, I discussed the findings and implications with the AC team. Its existence corroborates recent, as yet unpublished, epidemiological research by Tanser *et al*.[131], which geo-located high-risk areas of incident cases over a 10-year period (Figure 7.8(b) above). Their work revealed three areas of high incidence: two of them are adjacent areas in close proximity to the national road, which borders the DSA, comprising urban and peri-urban settings; and a third distinct area in the south-central area of the DSA, near the recent coal mine development. The first two areas have both been documented in previous work[130], but the third area was not a high-risk area when previous spatial analysis was undertaken in

[§§§] The terminology used to describe this large cluster is controversial. In the context of a generalised outbreak with an overall prevalence >30% and incidence >9%, HIV is endemic. Therefore, with an ongoing outbreak and such high background rates of infection, it is difficult to define and identify new “outbreaks” in these settings. However, incidence may increase in specific areas, as in this scenario. Identifying and understanding these new emergent “outbreaks” or clusters is important in curbing the outbreak and preventing further expansion.

2009[130], suggesting a recent increase in incident cases in this area. Pursuant to the work of Tanser *et al*. and to the New Cluster identified in my work, the area around the coal mine is being reclassifed as a high-incidence area.

The results of my combined phylogenetic and epidemiological analysis confirm that the increased incidence of infection detected by the spatial mapping conducted by Tanser *et al*. is a real, new and ongoing single HIV outbreak. In addition, my work suggests that the three high-risk areas may, in fact, be interlinked, as there are genetically similar cases across the three areas. The spatial and epidemiological data did not suggest such a link, but instead pointed to isolated hotspots of infection, again demonstrating the value of combined analysis.

My work also provides more detailed information about the New Cluster including its minimum size, a timeline of its rapid emergence, its persistence from 2011-2014 (when the study ended), and the background setting associated with it (including a description of the people involved). This cluster is systematically different to the rest of the sequence population, particularly in relation to gender (more males), employment (a substantial higher proportion in full-time employment) and wealth (more wealthy). These are the results one would expect if the cluster were related to mine workers, with gender and employment being the crucial linked metadata in this scenario. To my knowledge, this is one of the first examples in which phylogenetic techniques alone have identified a newly emerging cluster/outbreak. As such, it highlights an important role for using this modality in routine surveillance work and an approach to informing treatment and prevention strategies in the future.

In considering changes in recent years, which may have contributed to the emergence of this New Cluster, the most obvious and plausible link is the opening of the Somkele coal mine. This hypothesis fits with previous work highlighting an increased incidence of HIV around mines and other industrial developments[346]. There are many factors which may contribute to this increased risk, including infrastructure improvements and changes in the social dynamics of the local area, which in turn drive behavioural changes. Specific examples include:

- New opportunities for employment, which promote migration to the area leading to an increase in population density, particularly for single or ‘far from home’ men.

Of note is that the mine is now the largest employer in the region, having taken over from AC approximately 5 years ago;

- Better road networks linking the mine to local towns and national highways, which likely results in increased mobility of the local population. Both mobility and better access to transport/transport links are documented risk factors for HIV[130,347];
- Daily transport services for mine workers to and from local urban areas - these high-density urban communities are known hubs for HIV incidence and are reported to play a disproportionate role in seeding/reseeding the epidemic[131,347];
- Increased wealth and new opportunities for those working in the mine, resulting in increased access to transactional sex and sex workers. Furthermore, sex workers are likely to be drawn to the area given these same dynamics;
- Disruption to pre-existing communities and family structures as a result of the construction of the mine, as forcible resettlement occurred. This may increase risk factors for HIV acquisition.

With respect to a potential increase in transactional sex, although this point is speculative, the literature does document high concentrations of sex workers around mines and other industrial developments[348,349]. It is also interesting that one individual, whose sequence data are one of the closest to the origin of the New Cluster, reported >15 sexual partners in the 12 months prior to the 2013 surveillance. Even taking into account that such self-reporting might not be totally accurate, it is indicative of potentially higher risk sexual behaviour, and it would be reasonable to infer the possibility of partnering with sex workers, although this inference remains speculative.

These heterogeneities in incidence, together with the emerging clusters, suggest that evidence-based interventions targeting the most vulnerable populations in areas of greatest HIV incidence could be a powerful and cost-effective prevention approach in the AC DSA area. Such an approach could provide valuable information about where to focus new prevention programmes or where to intensify and/or supplement existing programmes, particularly in hyper-endemic settings[67]. Furthermore, given the scale of the mining industry in Southern Africa and the body of work (including this study) highlighting the association between increased HIV incidence and mining, regulatory policy/legislation should now incorporate health and safety clauses to consider prevention and education strategies for both employees and the local communities. The mining industry is declining

in the area, but these finding and principles are true to any industrial or economic development and thus, would require the same regulation. The advantages of a healthy workforce are far reaching, and the positive impacts of these programmes would cover other aspects of infectious disease transmission, including tuberculosis. Additionally, local/national prevention strategies should consider the prioritization of intervention campaigns to focus on areas identified as “high risk”, as they are likely to seed infection both inside and outside that area. The high levels of mobility seen in these populations mean that interventions may result in benefits that extend beyond the intervention site alone. Strategies could involve: treatment as prevention; messaging campaigns; community engagement to encourage testing and linkage to care; micro finance incentives; pre-exposure prophylaxis; occupational health programmes for mining companies; and targeting of men and vulnerable groups, including sex workers.

This work was undertaken in 2017, approximately 5 to 6 years after the emergence of the New Cluster. Even though the area in question was covered by the AC surveillance programme, the New Cluster was not detected by research surveillance methodologies and was only recently suggested by retrospective epidemiological research work. However, had real-time or near real-time sequence analysis been implemented earlier, this New Cluster could have been identified much sooner. This would have afforded the opportunity to intervene at that time, with a probable reduction in the growth of the cluster. The wealth of data produced by AC research provides a unique opportunity to undertake such work, but it is important to note that the vast majority of South Africa does not have equivalent sampling or data collection. Thus, setting up routine surveillance and real-time sequence analysis is a priority for the future.

As sequence data from 2015 onwards are not available, it is not possible to confirm whether or not this outbreak continues to grow, although the Re estimates suggest it was still expanding at the end of 2014. As this is a new finding and no intervention has been implemented yet, it would be reasonable to hypothesise that the cluster has continued to grow. Moreover, findings from phylogenetic population studies in which only a proportion of the population is sampled often underestimate the size of an outbreak. Given both of these factors, it is possible that the true size of the outbreak is larger. Timely sequencing and analysis of new samples is needed to confirm the subsequent evolution of the New Cluster. In addition, the local and/or national authorities should be notified to allow

available interventions to be put in place to limit the size and impact of ongoing transmission.

At the time of writing, further information on the New Cluster and the mining community was not available. However, my finding has prompted a rapid qualitative assessment of the Somkele mine setting by the AC Social Science team to identify particular behaviours associated with increased risk and the availability of transactional sex in this and surrounding areas. The work is currently ongoing and will help contextualise these results.

With increasingly affordable and portable sequencing techniques, storage capacities and improved analytical methods, real-time phylogenetic analysis will become more accessible over time. Pilot studies to incorporate real-time phylogenetics into routine surveillance protocols have been undertaken in a few high resource settings, for example the New York Department of State public health programme for HIV monitoring[350], which is trialling an automated system to continuously build phylogenetic trees incorporating new infections. In addition, the system detects new emergent and expanding clusters for further investigation, although to date it has not been used to target interventions. While phylogenetics alone can highlight these clusters, it is of limited use in informing intervention strategies and in affecting the outcome of these interventions without the addition of relevant epidemiological data. It is, however, a powerful tool to assess the outcome of trials of interventions dedicated to reduce transmission. Although epidemiological analysis alone may also lead to the same conclusions, this often takes longer and requires a larger volume of data to infer linked infections. Furthermore, in the case of HIV these data are often poorly reported or unreliable, meaning that epidemiological analysis alone may not be sufficient. This is particularly true in low-resource settings where data collection is often paper-based and where a higher proportion of participants may be illiterate.

As set out above, a goal of my work was to understand the AC epidemic in the context of the wider South African epidemic, and I found that there was substantial mixing between the AC and ZA populations. These findings contradict the hypothesis that the AC epidemic is a monophyletic and self-sustaining epidemic. If it was monophyletic, one would expect the genetic diversity of AC:AC sequences to be smaller than the genetic diversity found in both ZA:ZA and AC:ZA sequences, as the viruses would be closely related. Moreover, in

such a scenario one would also expect the ZA:ZA diversity to be higher than that shown in the AC:AC sequences. Factors serving to increase the diversity of the ZA:ZA sequences include multiple introductions, wide temporal (2000-2012) and geographical variation in sampling, and the potential for a higher rate of mixing between populations. In contrast, the following factors would act to limit the diversity shown by the AC:AC sequences: the relatively smaller region, the partially isolated nature of the AC DSA, and shorter sampling frame (2010-2014). However, as discussed previously, the diversity of the AC:AC sequences was not materially different from the ZA:ZA or AC:ZA sequences and was, in fact, slightly higher.

One question this raises, referring back to Figure 7.2, is why the AC:AC genetic diversity should be slightly higher than that of the ZA:ZA samples. This result appears counterintuitive given the small size of the AC DSA. However, one should bear in mind that the difference in size between the AC and ZA sequence populations is likely to serve to increase the observed diversity of the AC population relative to that of the ZA population (n=1,376 and 803 respectively). Given the similar distributions observed, as well as the fact that the mixed population (AC vs ZA population group) is also virtually identical to both the other two comparison groups, it seems probable that, essentially, we are sampling one population which includes both AC and ZA sequences. It is worth noting that an alternative explanation for these results, based on anecdotal suggestions from experts at AC, is that the subtype C virus genotype does not allow sufficient resolution to be able to distinguish between national and community transmission, when compared to other viral strains However, this view is not supported by the published literature[133].

For all of these reasons, this work suggests an important role for migration in both seeding and propagating the AC epidemic. A recent study (from the Treatment as Prevention trial data, as yet unpublished) suggests that for inhabitants of the AC DSA, a high proportion of sexual partnerships are with people from outside the DSA (>12%)[133]. This indicates a plausible mechanism for the continued mixing between AC and ZA virus populations.

The final key finding was the knowledge enhancement produced by linking sequence data with epidemiological data. This study demonstrated that significant advances in inference of transmission events can be made using phylogenetics, and it may be possible to infer information not previously available through either traditional or molecular epidemiology

when studied alone. In one cluster, the analysis was able to determine the likely direction of transmission, and it offered additional evidence to confirm the likelihood of this being a true transmission link. With further linkage and increased sample size this method is likely to have a significantly higher impact. There is much work being undertaken in this field[66,351], and with the emergence of Next Generation Sequencing, which will allow deep sequencing and inference of time of infection, it is likely that it will be possible to infer who infected whom in the future (discussed in Chapter 10). As discussed in the introduction of this chapter, this knowledge enhancement aligns with other recent literature supporting the role of these combined techniques to enable identification of key groups driving HIV transmission and thus, inform prevention strategies e.g. with respect to messaging campaigns to highlight age–disparate relationships between young women and men in their 30s[91]. These targeted campaigns to supplement existing programmes are required to change community norms, particularly as men are a challenging group to reach for HIV testing and ART. The use of these methods to identify drivers of transmission and high-risk groups is discussed further in Chapter 8.

The AC provides a unique opportunity to be able to access detailed, longitudinal, epidemiological and sequence data, which is rarely available in one place. Although there are inherent limitations with the sampling strategy, compared to other studies this sample provides both a relatively unbiased and representative sample, which is a strength. It includes people who do not know they are infected, as well as those who do, so the surveillance samples provide a more accurate representation of the infected population than studies arising from those who attend clinics (which is often the sampling frame for HIV phylogenetic work and may be biased by overrepresenting the sickest cases)[352]. Furthermore, it includes a random subset of the underlying population.

Limitations:

The study has many limitations, including limitations inherent in longitudinal demographic surveillance data and/or phylogenetic studies. One of the most striking limitations is the volume of missing data. In the epidemiological dataset it was not uncommon for many variables to have missing data for approximately 20% of participants (e.g. employment status, educational level). However, for sexual behavioural data and treatment-related behavioural data, the rate of missing data rises to in excess of 50% (e.g. condom use and ART treatment). This makes comparisons between different groups difficult – either the

desired analyses are not possible to undertake given the dearth of data or the unreliability of the results. Although this is a common limitation in large demographic datasets, the overall value of these datasets remains high as they provide an unrivalled source of data to enable tracking of population-level characteristics over time.

Missing and unsampled data are also a limitation as regards the phylogenetic analysis. The identification of 'clusters' is dependent on the sampling strategy. For example, one explanation for sequences that do not cluster is that some members of the transmission chain might not have not been sampled, rather than because the sequences are 'different' from clustered cases. Furthermore, the sequence population covered nearly 15% of all positive cases within the target population (of the DSA), which may not be sufficient to allow conclusions to be drawn which are representative of the target population.

In addition, one of the striking findings of my analysis is the number of putative transmission clusters involving two females. This is most likely caused by an unsampled missing male intermediary in the transmission chain, rather than female-to-female transmission, as this transmission pathway represents the lowest form of sexual risk (and would likely involve contact with infected blood). This type of interaction is unlikely to be disclosed given the traditional value system in KZN. Therefore, in all probability, this finding reflects the sampling strategy and the lower participation rate for men, and the higher participation of females (30% vs. 70%). Men are known to be underrepresented by the sampling strategy for numerous reasons including that: they are less likely to consent; are often not present during surveillance due to migration for employment purposes; are more likely to be diagnosed later due to the fact they are less likely to seek healthcare and screening services given the stigma associated with being seen in clinics; and that they do not engage in routine healthcare services e.g. maternity testing for all females. Furthermore, transmission from male to female partners is more common, with a probability of transmission[****] per unprotected sexual act two to three times higher than in female to male transmission[353], and this may in part explain the observed higher proportion of females being infected.

[****] In reality the probability of transmission (β) for sexually transmitted infections is more complicated ($R_o = \beta cD$). It also involves estimations of the effective rate of partner change/sexual mixing (c) and the duration of infectiousness (D) (Heathcote, 1984).

The time of sampling is another critical issue, as only virus from patients who are not virally suppressed can be genotyped and included in the molecular analyses. Most infections arise from individuals with higher viral loads, meaning that these are the most important individuals to capture in sampling. However, an HIV-positive individual may transmit the virus to another individual at an earlier time, prior to starting treatment, while at the (later) time of sampling that individual might have received ART and, therefore, be virally suppressed. This could lead to historical clusters that previously fuelled the epidemic, but that cannot now be identified. In 2012, 39% of the HIV positive population were taking ART (64% of those eligible for treatment), but this proportion is likely to have increased significantly given the recent emphasis on home-based HIV testing and linkage to care in this area (47.5% linkage to care)[354,355]. The differential health-seeking behaviours and ART adherence mean that certain groups are likely to be taking ART and virally suppressed, and, therefore, not included in the genotyping and phylogenetic analysis. Consequently, it is impossible to sample fully the population, resulting unavoidably in missing data and incomplete transmission chains. Sequence sample size directly affects the number of linkages observed, further biasing the results. Therefore, one of the main sources of bias in this study is the sampling strategy. One potential way to mitigate this issue would be to introduce proviral DNA sequencing to obtain the integrated cellular DNA from those who are virally suppressed, or to develop reliable inference methods/models to assess the proportion of missing links in identified transmission chains.

CONCLUSIONS:

Integration of molecular and epidemiological analysis offers a unique insight into the transmission dynamics of infectious disease outbreaks at both a large scale contextual level and at a detailed fine scale level. The detailed investigation of clusters utilising integrated analysis may inform prevention strategies by revealing both unknown outbreaks/clusters of infection and behavioural patterns, confirming linked transmissions, and potentially aiding in determining direction of transmission. These complementary methods provide the potential to uncover information that would otherwise be unknown through traditional epidemiological or phylogenetic investigation alone. The detection of unknown clusters, or correlates of infection, provides essential information for the design and implementation of successful intervention strategies. Ultimately, the integration of these two methods could lead to sophisticated techniques for predicting transmission

dynamics in infectious outbreaks. Further, it might also influence when, how, and towards whom, targeted interventions are implemented, particularly to supplement population-level prevention strategies in hyper-endemic settings.

Historically, it is likely that many of the HIV infections in the AC DSA were acquired by those working at mines in other locations, prior to those individuals travelling back to the AC area. However, following the recent development of the Somkele mine in the DSA, that dynamic has likely changed, and new environments, opportunities and behaviours, all of which may propagate the transmission of HIV, have come to the DSA itself. However, the reasons underlying the emergence of the New Cluster remain speculative. Further work is needed to confirm both my hypotheses, and the implications for infection control and prevention going forwards.

This New Cluster reflects a concentrated sub-epidemic within a generalised hyper-endemic setting. Strategies to target and address the local factors, which cause high levels of transmission are vital to achieve reductions in incidence of HIV. Mathematical modelling has shown that maximum reductions can be achieved by incorporating interventions which specifically target high-risk populations (or risk spaces)[356]. Therefore, earlier identification of these sub-epidemics and a better understanding of the local factors driving transmission in these areas are needed both to understand the factors driving national level HIV dynamics and to inform national prevention strategies.

Acknowledgements:
The sequencing and alignments were overseen by Tulio de Oliveria at the Africa Centre. The phylogenetics work presented was supported by Stéphane Hué of the London School of Hygiene and Tropical Medicine.

CHAPTER 8

DETERMINING PATTERNS OF MIGRATION AND HIV INCIDENCE IN RURAL KWAZULU-NATAL: A COMBINED MOLECULAR AND CLASSICAL EPIDEMIOLOGICAL APPROACH.

In this chapter, I build on the work presented in the previous chapter by exploring both transmission events in migrators within the Africa Centre population, and transmission between areas with different levels of HIV prevalence.

8.1 INTRODUCTION

As discussed in Chapter 7, migration has been characterised as a key risk factor in the spread of HIV[125,137,357]. However, fully understanding the link between migration and incidence of HIV requires a nuanced approach. It is likely that it is the behaviour of some migrants, or short-term `commuters', when away from home (e.g. risky sexual practices), rather than mobility itself, which puts those people at risk of infection[125]. Given that there are many different reasons for migrating, categorising all migrants as `at risk' overlooks heterogeneity in migration flows and patterns, as well as in behavioural practices. In addition, it is important to differentiate between the characteristics of the different areas between which migrants move, rather than characterising such areas in general as 'hotspots'.

Migration in southern Africa is often driven by employment opportunities, which results in circular temporary labour migration between rural and urban areas[137,358,359]. This is the predominant type of migration within Kwa-Zulu Natal (KZN)[125,137,359]. This circular labour migration involves travelling to find work for variable periods and then returning home. The nature of the work is often casual and temporary, with a lack of accommodation and/or facilities for families. Circular labour migration often involves young adults, with reports of more than 60% of males between the ages of 30 and 39 years old in South Africa undertaking temporary migration, and it is associated with rapid HIV spread[358]. A

significant proportion of HIV-infected migrants are unaware of their HIV positive status and are not on ART and, therefore, may contribute to new HIV transmissions[360,361]. However, there is a lack of systematic data on HIV prevalence and incidence among migrants in both the Africa Centre (AC) population and South Africa more generally.

One hypothesis is that migrators have riskier sexual behaviours when away from 'home'. Workers often migrate to employment without their partners or families, in part, because there is rarely accommodation for family members in the areas around industrial sites to which they travel. Their employment provides increased financial means, which may create new opportunities for sex. Additionally, the research literature indicates that areas of high employment (e.g. industrial sites) have associations with the sex industry[346,348,349]. These factors may contribute towards riskier sexual behaviours, for example encouraging casual and/or commercial sex. In this way, migration may increase the risk of HIV acquisition.

Furthermore, migrants moving for employment may move to high-population density areas with a high prevalence of HIV. Any sexual encounter in these areas has a higher probability of being with an infected person. Recent reports suggest that 40% of new infections within the AC Demographic Surveillance Area (DSA) occur within 8% of that area – specifically, in an urban/peri-urban area adjacent to the main highway between Durban and Mozambique[131]. It is reasonable to hypothesise that migrants may become infected in the area to which they travel for work, and subsequently, take the infection back 'home' to lower prevalence areas when they return to visit family. This, in turn, could seed new clusters of infection in the areas from which migrators originate.

Phylogenetic analysis can aid the study of risk associated with migration by using sequence data to determine the ancestral history of sequences, thus allowing the potential geographical origins of outbreaks and transmission events to be inferred. Furthermore, the pairwise genetic distance between viral sequences can be used as a proxy for genetic variation within a population. For example, the smaller the mean population genetic distance, the more similar the sequences within that population and the more likely that the outbreak is localised; and conversely, the larger the mean population genetic distance, the more variability, and the higher the probability of distant sequences contributing to the population pool, for example via migration events.

The types and patterns of migration in sub-Saharan Africa are poorly documented. Furthermore, the role of migrants in propagating the HIV epidemic is unknown. This limits our understanding and our ability to implement preventative measures within the vulnerable migrant population. However, as described in Chapter 7, a combined phylogenetic and epidemiological analysis may begin to answer some of these questions.

There are many definitions of 'migration' and 'mobility,' and it is often difficult to classify individuals accurately. For the purposes of my work and with the aim of being consistent, migration is defined herein as any change in residence, including both moving within the DSA (internal migration) or moves which start or end outside the DSA (external migration).

8.2 Hypotheses

In this chapter I build on the work presented in the previous chapter by exploring both transmission events in migrators within the AC population, and transmission between different HIV prevalence areas. I hypothesise that:

1. Migrators have a higher risk of HIV incidence when compared to non-migrators, and that migration type influences risk, with those migrating externally at higher risk than those who migrate internally.
2. External migrators are likely to be a key source of new infections into the DSA, contributing to the AC epidemic.
3. Background prevalence of HIV in the area of residence influences the risk of infection: the higher the prevalence the higher the risk.
4. HIV sequences in the non-migrators group are likely to have, on average, a lower viral genetic diversity than the migrator group, since infections remain local and fuelled by closely related viruses, unless the epidemic is fuelled by multiple external introductions. In those who migrate, viruses infecting the external migrators' group have a higher mean population genetic distance than that of only internal migrators.

8.3 Objectives

I address these hypotheses by exploring the role of migration in the AC HIV epidemic. I undertake two complementary studies to document migration flows both within and out of the AC population: i) a classical epidemiological study (Study 1, set out in section 8.4

below); and ii) a novel molecular and bioinformatics studies (Study 2, set out in section 8.5 below). These studies use a common cohort and research definitions, to address three main objectives:

1. To determine if the risk of HIV acquisition is influenced by migration type. To do this, I explore two categories of migration:
    - A summary variable dividing the population into three mutually exclusive migratory groups and;
    - A time-varying categorisation to explore if recent migration (within the last year) influences risk, by the type of recent migration.
2. To investigate if location of residence (as determined by background HIV prevalence) affects HIV acquisition by:
    - Estimating the variation in the hazard ratio of HIV seroconversion by exposure to varying HIV prevalence areas (by classical epidemiological analysis); and
    - Determining the rates of transmission between different areas (by molecular analysis): To what extent is the AC epidemic due to multiple viral introductions by external sources, endemic transmission within the area, or both?
3. To assess the variation in genetic distance between HIV positive migrators and non-migrators as a proxy for population genetic variation.

8.4 **Study 1:** Classical epidemiological analysis of HIV incidence in relation to migration states

8.4.1 Methods

A retrospective cohort study and survival analysis was conducted using data from population-based, longitudinal surveillance conducted by the Africa Centre (AC) for Health and Population Studies in a predominantly rural community in the uMkhanyakude district of KZN. The district is one of the most deprived in the country and is characterized by high levels of circular migration and HIV infection, with an HIV prevalence of approximately 25%[128]. AC demographic and health data collection, in addition to anonymised HIV testing methods, are described in detail in Chapter 3.

Study cohort

Data were available from January 2004 to September 2016. I created a retrospective open cohort, into which individuals could join and exit during the study period. Inclusion criteria

were residents within the AC DSA population who were: (i) questioned as part of routine demographic and HIV surveillance between 1st January 2005 and 31st December 2015; (ii) aged between 16 and 90 years; (iii) had a negative HIV test at baseline; (iv) had at least one subsequent HIV test; and (v) had linked historical migration information.

I excluded individuals with zero or one HIV tests or only positive HIV tests. I also excluded any person-time for whom: (i) the window period between last negative and first positive HIV test >5 years (i.e. was so long that it was not possible to accurately impute an estimated seroconversion date); and (ii) residency or migration history was unknown. Figure 8.1 shows the steps taken to obtain the dataset used in this study.

Individuals entered the cohort on the date of their first HIV negative test (after 1st January 2005) and follow-up time was set from this point. They left the cohort if they had a primary outcome event (the date of estimated HIV seroconversion, see below) or they were right-censored at the date of their last HIV negative test prior to departure from the DSA (with no subsequent data) or death[††††]. If they had not had a negative HIV test within one year of the study end date (i.e. between 31st December 2014 and 30th December 2016), they were also censored at the date of their last known negative test, as we do not know their HIV status after this. Figure 8.2 shows graphical examples of the end-point and censoring undertaken in this analysis.

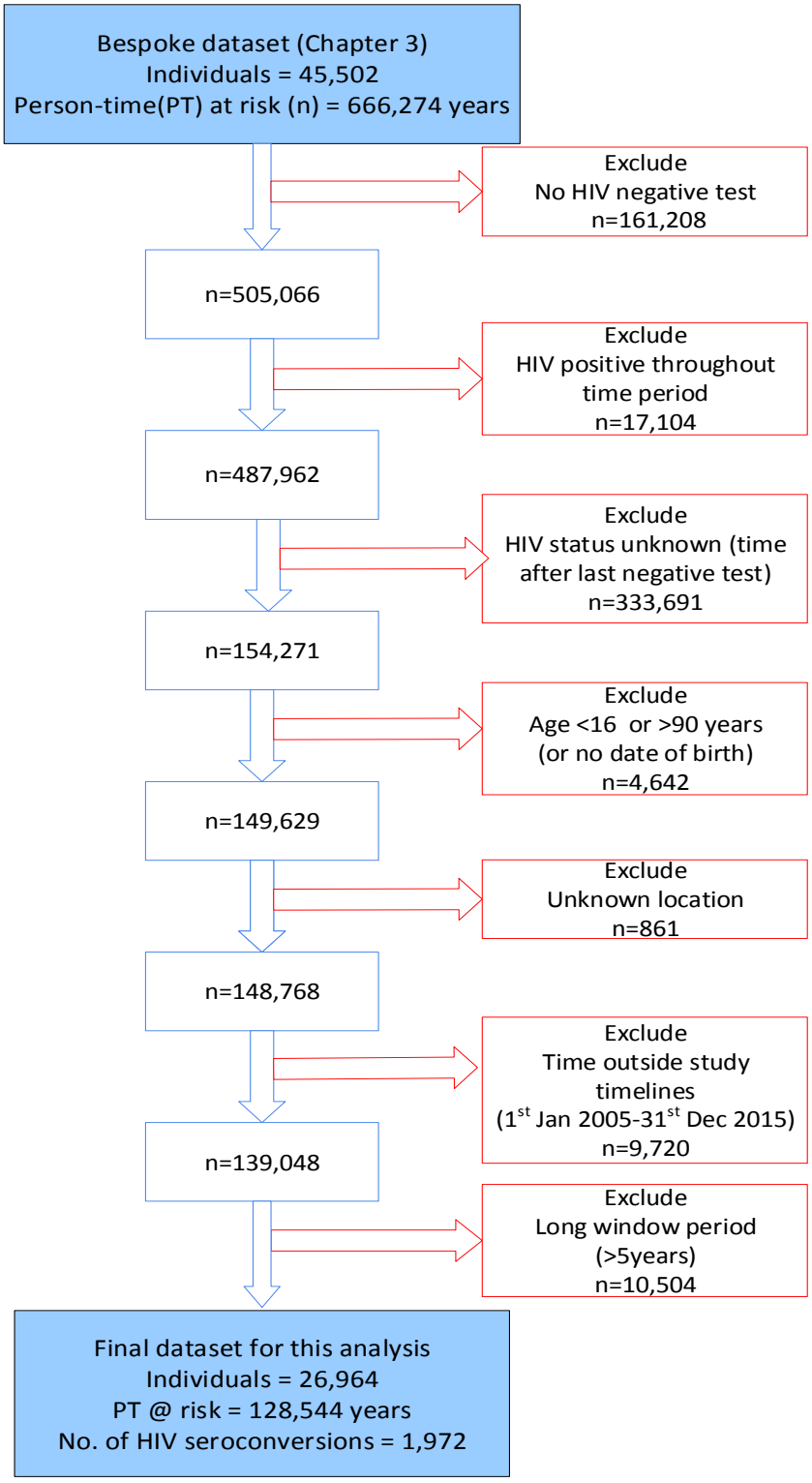
The follow-up time was split into contribution of person-time to time periods, defined by changes in calendar year and migration status.

Definition of outcome

The primary outcome was HIV seroconversion. A seroconversion date was imputed using the mid-point between the date of last negative HIV test and first positive HIV test, as per previous publications[362].

[††††] As individuals with both HIV and long window periods are excluded, the study cohort will all have a negative HIV test within the last 5 years, and often within the last 2-3 years. Therefore, I have no reason to believe that a high proportion of the deaths in this group will be likely attributable to HIV causes, and thus death is considered as a censoring, not as an end-point event.

Figure 8.1: Steps to obtain dataset for analysis



Definition of exposures

The primary exposure was migration status. I examined only the migrations which preceded HIV seroconversion. I categorised migration *status* into the following groups:

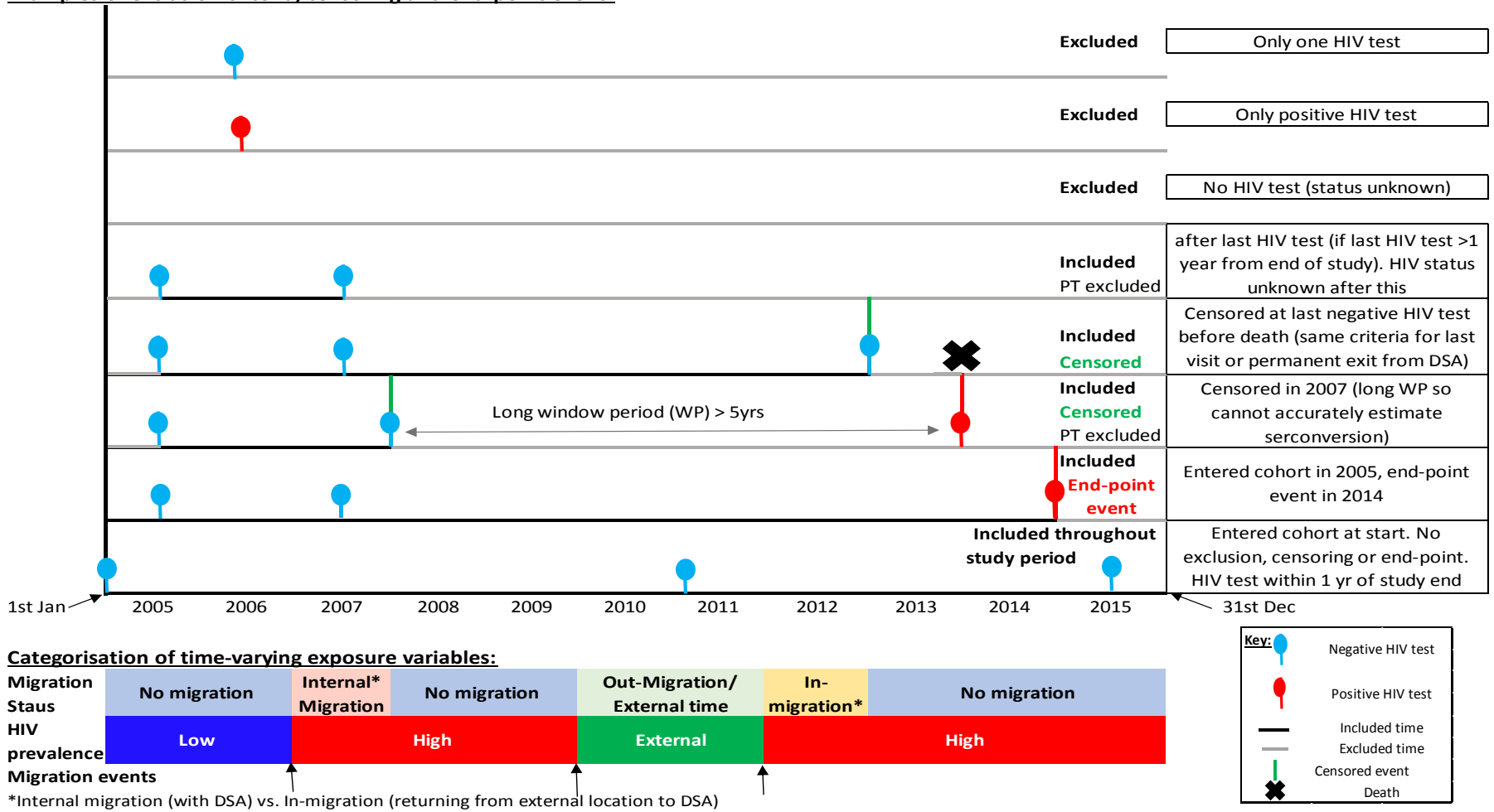
1. No migration -those resident in the DSA who had not moved household;
2. Internal migration – moved within the DSA only (i.e. moving between households and/or between areas of different HIV prevalence); and
3. External migration/ external time –those who moved into or out of the DSA and spent time outside the DSA. This group was sub-divided into:
    a. Out migration/External time - a move from within the DSA to an external location and time spent in an external location; and
    b. In migration – returned from an external location back into the DSA.

For each time-period, individuals were classified as migrant or non-migrant, along with the associated sub-classification. Therefore, individuals could contribute to more than one category over time, but only one category at any specific point in time.

After a migration event, an individual was considered as within the 'migrant period' for a period of one year from the date of the move (both within and outside the DSA). Thus, if a participant moved to a different household within the DSA on 20th January 2006, they would be classified as an internal migrant between 20th January 2006 and 20thJanuary 2007, at which time they would return to a non-migrant status, unless they moved again. For a participant who out-migrated on 26th January 2007 and in-migrated back to the DSA on 13th April 2009, they would be classified as an out-migration/external time between 26th January 2007 and 13th April 2009, and then an in-migrator from 13th April 2009 until 13th April 2010. New migrants arriving for the first time in the DSA were not included as they do not meet the inclusion criteria for the study. Out-migration with no subsequent in-migration event was characterized as "loss to follow up" and censored. Figure 8.2 shows different examples of these categorisations.

Furthermore, I collapsed this migration status into a summary variable (migration *type*) for baseline characteristic and an initial unadjusted analysis. I divided the population into those who had never migrated versus those who had ever migrated (across the whole period of observation). Of those who had ever migrated, I sub-divided them into those who had 'ever externally migrated' versus those who had 'only internally migrated'. In the situation in which an individual had both externally and internally migrated, they were

Figure 8.2: Graphical representation of cohort categorisation, exclusion, end-point event and censoring



categorised as 'ever externally migrated' as it was felt that the people who had left the DSA were systematically different from those who moved household within the DSA – which was very common.

I also considered the locations of residence by background HIV prevalence across all person-time at risk. This exposure may vary by time. For each time-period, individuals were classified into four categories by the background HIV prevalence of their residential home (high, medium or low prevalence) or external location. For all buildings in the surveillance area, the prevalence each year has been calculated between 2003 and 2012 (as described in Chapters 3, Figure 3.4). I calculated the median and interquartile range by year, and categorised homes into three categories: low prevalence for those under the 25^{th} centile, medium prevalence for the middle 50% of households, and high prevalence for those with a prevalence > 75^{th} centile. The prevalence data has not been updated since 2012, therefore these data were not available. Between 2003 and 2012, there was a general slight upwards trend in prevalence across the DSA, but the data did fluctuate. In the absence of an alternative, given the data had not varied systematically, I used 2012 prevalence information for the time-periods from 2012-2015. As with migration status, a person may contribute to more than one category (if they move), but can only contribute to one area at a specific point in time.

I considered as potential confounders all factors associated with HIV acquisition risk documented in the literature, for which I was able to obtain data in the AC datasets. As socio-demographic confounders I considered: age; gender; maximum education (primary (0-7 years), secondary (8-12 years), or tertiary level (>12 years or education beyond high school); and household asset-based wealth quintile (as described in Chapters 3 & 7)[363,364]. As potential behavioural confounders or mediators I considered: ever reported >1 sexual partner in a 12 month period (as a proxy for high-risk sexual behaviour); and marital status (never married, engaged, married, and previously married).

Furthermore, it is likely that the risk of HIV incidence varied over time within the study period. Firstly, the background prevalence data show this with a general upward trend over time. Moreover, the initiation of treatment and prevention campaigns likely had an impact. For instance, when the study started, the availability of ART was lower than it is today. Therefore, I also controlled for calendar year.

Statistical analysis

All analyses were undertaken using Stata 14 (Stata Corp LT, College Station, TX, USA). All tests of significance used $p<0.05$ as the threshold of statistical significance, unless otherwise stated.

Firstly, I described the baseline characteristics of those included in this cohort- comparing the general population to different migration *types* (never migrated vs ever migrated; and ever externally migrated vs only internally migrated). I used chi-squared tests to evaluate differences in characteristics between the groups for categorical variables, and Kruskal-Wallis tests for ordinal variables. Those variables with missing data >10% were not included in the multivariable analysis.

I explored the probability of HIV acquisition between different migratory *types* (never migrated, ever externally migrated, only internally migrated) by calculating unadjusted incidence rates per 100 person-years at risk and constructing a Kaplan-Meier survival curve (unadjusted probabilities of seroconversion by time to event). However, this categorisation of migration assumes that if an individual has ever had a migration event, the risk remains constant thereafter, whereas the effect is likely to wane over time. Therefore, to explore the time-varying nature of migration events and limit the bias of potentially over-counting exposure time, I repeat this incidence analysis with migration categorised by migration *state* (no recent migration, recent internal migration, recent out-migration/external time, recent in-migration). This allows individuals to move between migration categories over the study period and explores the effect of recent migration on HIV acquisition risk.

I used Cox proportional hazard modelling to estimate the Hazard Ratios (HR) for HIV incidence. I estimated univariable and multivariable HRs, using a forward stepwise approach to confounder inclusion. All models included the primary and secondary exposures - migration status and background HIV prevalence of residence. Variables associated with incidence in the univariable analysis ($p<0.2$ in the log-rank test) were included in the final model (age, gender and calendar year). This took into account the open nature of the cohort with constant new enrolments and loss-to-follow-up due to death or out-migration, and it maintains the time-varying nature of my primary exposure. This was a complete case analysis where only individuals with data for all variables were

included. I also explored whether there were significant differences between sub-groups within a category, using a test of equality (Wald test). The proportional hazard assumption was tested to check if the ratio of hazards between groups (within one variable) was constant over time (an assumption for Cox regression). Furthermore, model fit was assessed by comparing the likelihood ratio tests (LRTs) between models. LRT was also used to see if the inclusion of confounders improved model performance.

In addition to the main analyses outlined above, a number of further sensitivity analyses were undertaken to assess the definitions of outcome, exposure and censoring criteria used. This included:

1. Sex-specific stratified analysis: HIV incidence rates differ by age and sex, and reports suggest there are sex-specific patterns of both migration and HIV acquisition in the AC DSA[134,91], therefore, the analysis was stratified by sex. Given the *a priori* knowledge of differential patterns of both migration and HIV acquisition across both sex and age groups, I also considered effect modification of each of age, gender, and HIV prevalence on the association between recent migration and HIV acquisition, by interacting the primary exposure (migration) and these variables in the survival analysis. I used AIC[‡‡‡‡] to explore if the expanded (interacted) model provided a better fit in each case;
2. Recent migration classification defined as up to 24 months, rather than 12 months, of a movement;
3. Using random imputation (imputed uniformly at random between the date of last negative HIV test and first positive HIV test) rather than mid-point imputation to estimate the date of seroconversion. A recent publication suggests that random imputation is an effective means of estimating seroconversion data in this population[365]; and
4. The window period for estimated seroconversion (5 years vs 2.5 years).

[‡‡‡‡] Akaike's information criterion (AIC) is an estimator of the relative quality of statistical models for a specific dataset compared to other models. It tells you nothing about the absolute quality. It balances each model's 'goodness of fit' with its complexity, to select the most parsimonious model. Thus, it can be used to select the most appropriate model. A decrease in AIC greater than two points suggests an improved model.

Ethics

Ethics approval for AC surveillance was granted by the Biomedical Research Ethics Committee, University of KwaZulu-Natal. Informed consent is required separately for the main questionnaire, the sexual behaviour questionnaire, and for anonymised HIV sero-testing. This analysis was exempted from additional ethical review by University College London Ethics Review Committee since it involved only anonymized data

8.4.2 Results

Baseline characteristics

Between 1st January 2005 and 31st December 2015, 26,964 people, contributing 128,544 person-time years of follow-up, met the inclusion criteria for the epidemiological analysis of HIV incidence in relation to migration states. Baseline characteristics of the participants are provided in Table 8.1. This information is presented for the overall AC population, as well as for four sub-groups of the population according to migration type: non-migrators compared to those who have ever migrated in the follow-up period; and amongst those who have ever migrated migrated, external migrators compared to (only) internal migrators.

This cohort includes more women than men, with a ratio of 1.4:1. The median age of participants was 23 (IQR: 18-46). At baseline, the majority of the cohort were less wealthy (in the two most deprived wealth quintiles within the area) and had primary or secondary school education (figures presented in Table 8.1). There is a higher proportion of men in the migrators cohort compared to non-migrators. This appears to be due to the high proportion of men in the external migrators group (47.4%), as internal migrators have the lowest proportion of men across all the cohorts (35.3%). Migrators are significantly younger than non-migrators. Non-migrators are predominantly spread between the age groups at two extremes of age (<20 and >41 years) with smaller proportions of people in the intermediate categories. All migrating groups have substantially more people in the lowest age categories (<20 years) than all the older age categories. Migrators are more likely than non-migrators to have both primary and secondary education as their highest educational attainment, but less likely to have had either no education or tertiary level education ($p<0.01$). The trends seen across differences in wealth between migrators and non-migrators are approximately equal, although the result is statistically significant, likely due to the large sample size($p=0.01$). Background HIV prevalence of area of residence also

appears broadly similar between the groups (although given the sample size, statistical analysis suggests a significant difference across the groups, but this does not appear to be epidemiologically relevant). Finally, migrators report more risky sexual behaviour, with a significantly high proportion reporting multiple partners in a 12 month period, largely driven by the external migrators cohort. However, given the high proportion of 'missingness' in these data, the patterns are difficult to interpret accurately.

36.8% (n=9,919) of individuals enrolled into the cohort in 2005 when it opened. In 2006-07, a further 25% (n=6,744) of the cohort enrolled. Between 2008-15, there was an annual downward trend in the number of people enrolling (1,924-938, respectively), with the exception of 2009, when only 914 individuals joined the cohort.

Continued overleaf.

Table 8.1: Baseline characteristics of participants

	All	Non-migrators	Ever Migrated	p-value	Migrators		
					Ever externally migrated	Only internally migrated	p-value
Eligible individuals	26,964	10,545	16,419		13,005	3,414	
		39.1%	60.9%		79.2%	20.8%	
Gender (%)							
Male	41.7	36.7	44.9		47.4	35.3	
Female	58.3	63.3	55.1	p<0.001	52.6	64.7	p<0.001
Age (%)*							
<20	40.0	33.1	59.8		62.9	48.2	
20-25	17.4	5.3	15.5		17.3	8.6	
26-40	13.2	12.3	11.9		10.7	16.4	
41-60	15.4	25.9	8.4		6.4	16.4	
>60	13.9	23.4	4.4	p<0.001	2.8	10.5	p<0.001
Missing	0.0	0.0	0.0		0.0	0.0	
Wealth quintiles (%)							
Most deprived	25.0	23.6	25.9		24.7	30.5	
2nd most deprived	27.6	28.1	27.2		26.9	28.4	
Middle quintile	18.2	19.9	17.2		17.0	17.8	
2nd least deprived	13.1	14.6	12.1		11.7	13.5	
Least deprived	11.9	13.6	10.8	p=0.01	11.4	8.6	p<0.001
Missing	4.2	0.2	6.8		8.2	1.3	
Maximum Education (%)							
None	18.3	28.0	12.1		8.8	25.0	
Primary (1-7)	40.3	36.4	42.8		44.1	37.6	
Secondary (8-12)	24.6	17.7	29.0		31.6	19.2	
Tertiary (>12)	1.1	1.7	0.8	p<0.001	0.8	0.8	p<0.001
Missing	15.7	16.2	15.3		14.8	17.3	
Residence HIV prevalence (%)							
Low prevalence areas	28.1	28.3	27.9		27.2	30.5	
Medium prevalence areas	50.6	53.0	49.1		49.0	49.2	
High prevalence areas	17.5	18.8	16.7		15.8	20.2	
External	3.82	0.0	6.3	p<0.001	7.9	0.1	p<0.001
Residency data missing	0.0	0.0	0.0		0.0	0.0	
Multiple partners ever reported (%)							
Yes	6.3	3.5	8.1		9.0	4.4	
No	37.3	32.4	40.5	p<0.001	40.4	41.1	p<0.001
Missing	56.4	64.1	51.4		50.6	54.5	

*The age categories are divided according to approximate quintiles based on person-time contribution, rather than on individuals: <20- 24.4%, 20-25- 18.8%, 26-40- 17.4%, 41-60- 21.3%, >60- 17.1%. This is due to the baseline characteristics on enrolment into the cohort being skewed towards younger ages, as people within the surveillance area 'age into' the cohort disproportionately more than others entering the cohort at a later age, giving a larger proportion in the younger age categories.

It is likely that the highly significant findings are due to the very large sample size, but values shown for completeness.

Investigating the general patterns of external migration within the study population showed that of those who migrate externally, 68.6% (n=8,923) have only one migration event, 24.5% two migration events (n=3,186) and 6.8% (n=894) three or more migration events. 28% (n=7,538) of the overall study population were categorized as external migrants when information was last available on them. The median time spent as an external migrant was 465 days (IQR 365-701). These figures decreased with an increasing number of migrations: 363 days (IQR: 203-714) in those with more than 3 migration events and 188 days (111-353) in those with 5 or more migration events. The median time migrators spent back within the AC DSA between multiple migration events was 842 days (IQR: 351-1348). As expected, this time also decreased with increasing number of migration events: 618 days (IQR: 337-1043) in those with 3 or more migration events and 327 days (IQR: 203-669) in those with five or more migration events.

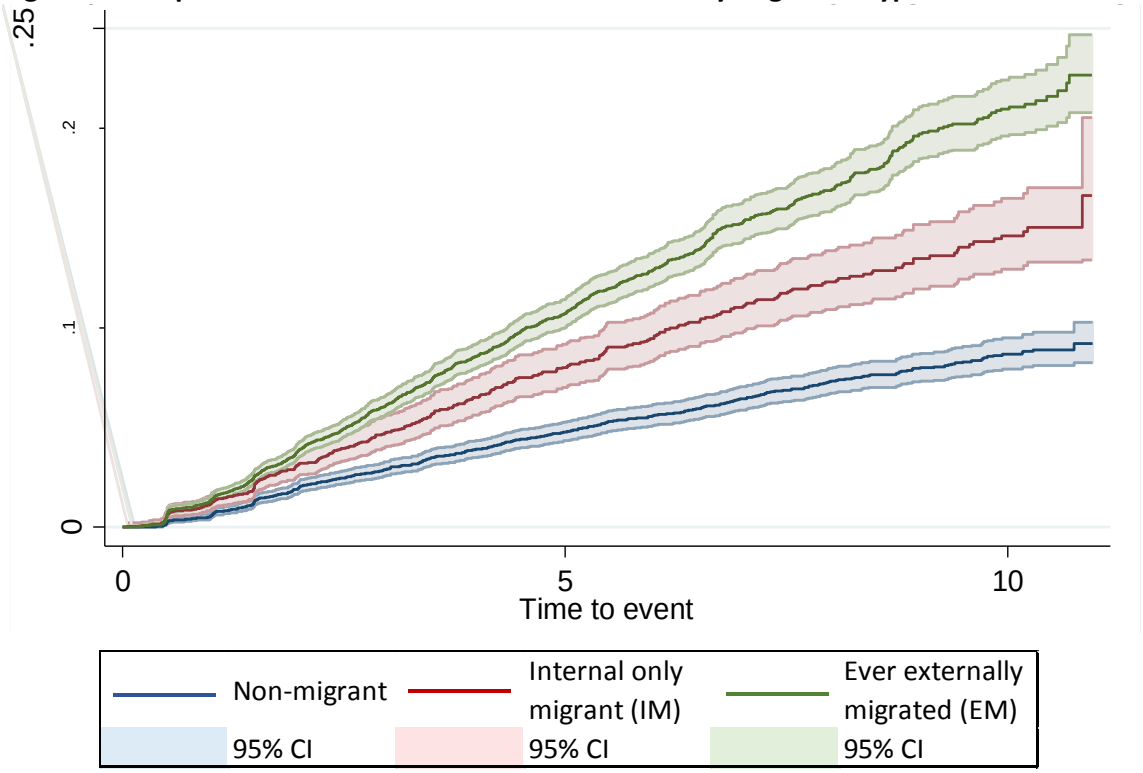
The median window period between last negative and first positive HIV test was 724.5 days (IQR: 373-1138) and there was no significant difference between the window period for all categories of migration.

Survival analysis

During follow-up 1,972 HIV seroconversions were observed. Figure 8.3 shows the Kaplan-Meier survival curve for unadjusted probability of HIV seroconversion in non-migrators compared to those who only migrated internally and ever external migrators (time to event analysis), where each participant falls into one of the three mutually exclusive categories.

Overall unadjusted HIV incidence rate per 100 person-years (py) at risk was 1.53 (95% CI: 1.47-1.60). Crude incidence rates (per 100 py at risk) by migration *type* showed that non-migrators have a lower unadjusted HIV incidence rate per 100 py at risk (0.93 (CI: 0.85-1.00)) than migrators, with those who have ever externally migrated being at highest risk compared to those who have only internally migrated (2.22 (CI: 2.09-2.35P and 1.60 (CI: 1.43-1.80), respectively).

Figure 8.3: Kaplan-Meier curve for HIV seroconversion by migration type



Time 0= Enrollment

Furthermore, I went on to calculate crude incidence rates by migration *state*, to account for the fact that the effect of migration may vary over time and may not continue indefinitely after the migration event. This classification of migration is used henceforth in the analysis as it incorporates the time-varying component of the exposure, limits the bias due to over-counting exposure time, and allows exploration of the role of recent migration events on HIV incidence. Both classifications for migration confirmed that incidence was significantly higher in all migration states when compared to no (recent) migration. The incidence rates across different migration *states* are shown in Table 8.2.

Table 8.2: Incidence rates according to time-updated status for each explanatory variables

	No. of seroconversions	Person-time (years)	Incidence rate (per 100py)	95% CI
Recent migration/mobility (1 year)	(n=1972)	(n=128,544)		
None	1,587	112,542	1.41	[1.34-1.48]
Internal migration	84	3,809	2.21	[1.78-2.73]
In-migration (from external)§	78	3,264	2.39	[1.91-2.98]
Out-migration/External time at risk ǂ	223	8,929	2.50	[2.19-2.85]
Background HIV prevalence				
Low	405	36,005	1.12	[1.02-1.24]
Mid	909	61,665	1.47	[1.38-1.57]
High	435	21,921	1.98	[1.80-2.17]
External*	223	8,929	2.50	[2.19-2.84]
Age categories (yrs)				
<20	551	30,604	1.80	[1.66-1.96]
20-25	798	22,818	3.50	[3.50-3.26]
26-40	423	21,692	1.95	[1.77-2.14]
41-60	170	30,479	0.56	[0.48-0.64]
>60	30	22,950	0.13	[0.09-0.19]
Gender				
Male	407	46,454	0.88	[0.80-0.97]
Female	1,565	82,090	1.91	[1.81-2.00]

§ Best estimate of HIV infection to be in AC within one year of an external migration event
ǂ Best estimate of HIV infection to be outside AC during an external migration event
* HIV prevalence in external locations is unknown. This category is the sum or all person-time spent in this state

In survival analysis (Table 8.3) univariable model (model 1), I found that any recent migration event is associated with a significantly higher risk of HIV acquisition than no recent migration ($p<0.05$). However, the risk is not significantly different between the different categories of migration (internal migration (HR 1.57) vs. in-migration (1.62) vs. out-migration (HR 1.69)) (on trend testing for differences between migration categories - $p>0.5$ for all combinations). The background HIV prevalence of an individual's current residence is positively associated with risk of HIV acquisition i.e. it significantly increases with increased HIV prevalence (trend test to explore differences within category: $p<0.05$ for all combinations) and those living in external areas have a significantly higher risk than those within the DSA (HR 2.12, $p>0.05$)(model 1). The effects of both recent migration and background HIV prevalence are independent of one-another (model 2), as combining the two in multivariable models does not alter the HR, except for the out-migration group, whose HR increases and becomes significantly different to the other migration categories ($p=0.02$).

Next, I assessed whether the variation in risk due to migration status and location might be attributable to other factors reported in the literature to be associated with both migration and HIV acquisition. I found strong evidence that the association between migration status and HIV seroconversion was driven by age (model 3). Controlling for age decreases the HR for all categories of migration, particularly for recent in-migration. Both internal migration and recent in-migration become not significantly different to those with no recent migration. However, external out-migration remains at significantly higher risk than other categories of migration. Furthermore, the Likelihood Ratio Test (LRT) shows a significant difference between the model with and without age (models 3&2, respectively)($p<0.0001$). This suggests the more complex model (including age) may explain the variability observed in the association between HIV acquisition and recent migration, and age is likely to be a confounder of this effect.

Controlling for age had less effect on the increased risk associated with HIV area incidence risk patterns (i.e. the positive association remained). Thus, internal background HIV prevalence is an independent risk factor for HIV acquisition, which is not substantially affected by other variables (as shown by the lack of change in HRs across models (models 4-6).

Additionally, these data suggest that HIV acquisition is significantly higher in the 20-25 year old age group than any other group (HR 1.80, $p<0.05$). Those <20 years and between 26 and 40 years did not have significantly different risks of HIV acquisition, but all age groups >40 years were at decreasing risk, with risk falling as age increased. This effect was found across all models and was statistically significant ($p<0.05$).

Including gender in the model does not affect the trends in associations seen between migration and HIV, but it suggests that women were at substantially higher risk than men (HR 2.29, $p<0.001$)(model 5). Both models including gender (models 5&6) indicate some deviation from the proportional hazards (PH) assumption. This suggests that the inference of the association between HIV acquisition and gender varies over time. I explored the nature of the deviation from PH for the effect of gender, by the graphical representation of the proportional hazard test which shows that over study time (from enrolment) the difference in hazard of HIV between men and women is greater initially, then appears to even out(Appendix 8.1). Furthermore, the very large dataset means that even trivial

Table 8.3: Univariable and multivariable analyses using Cox regression model – Hazard ratio calculations for HIV acquisition by risk factors

	Univariate analysis		ST cox regression models (multivariate, adjusted HR)				
	Model 1 Crude HR (HR [95% CI])	p-value (LRT)**	**Model 2** Mig+HIV prev (HR [95% CI])	**Model 3** M2 + Age (HR [95% CI])	**Model 4** M2+Gender (HR [95% CI])	**Model 5** M3 + Gender (HR [95% CI])	**Model 6** M5 + year (HR [95% CI])
Recent migration/mobility (1 year)							
None	1.00		1.00	1.00	1.00	1.00	1.00
Internal migration	1.57 [1.26-1.96]		1.55 [1.24-1.93]	1.18 [0.95-1.48]	1.53 [1.22-1.90]	1.21 [0.90-1.40]	1.13 [0.90-1.40]
In-migration (from external)	1.62 [1.29-2.03]		1.62 [1.29-2.03]	0.93 [0.74-1.17]	1.73 [1.37-2.18]	0.92 [0.73-1.16]	0.93 [0.74-1.16]
Out-migration/External time	1.69 [1.47-1.94]	p<0.001	2.20 [1.86-2.59]	1.23 [1.04-1.45]	2.43 [2.06-2.86]	1.30 [1.10-1.53]	1.30 [1.10-1.54]
Background HIV prevalence							
Low	1.00		1.00	1.00	1.00	1.00	1.00
Mid	1.32 [1.17-1.48]		1.32 [1.17-1.48]	1.25 [1.11-1.41]	1.33 [1.18-1.50]	1.29 [1.14-1.45]	1.29 [1.14-1.45]
High	1.76 [1.54-2.01]		1.75 [1.53-2.00]	1.54 [1.35-1.77]	1.81 [1.58-2.08]	1.65 [1.44-1.88]	1.65 [1.44-1.89]
External*	2.12 [1.80-2.50]	p<0.001	omitted	omitted	omitted	omitted	omitted
Age categories (yrs)							
<20	1.00			1.00		1.00	1.00
20-25	1.80 [1.60-2.03]			1.80 [1.59-2.03]		1.73 [1.53-1.96]	1.71 [1.50-1.95]
26-40	1.00 [0.86-1.15]			0.99 [0.86-1.15]		0.90 [0.78-1.03]	0.90 [0.78-1.04]
41-60	0.28 [0.24-0.34]			0.29 [0.24-0.34]		0.24 [0.20-0.29]	0.24 [0.20-0.29]
>60	0.07 [0.05-0.10]	p<0.001		0.07 [0.05-0.10]		0.06 [0.04-0.08]	0.06 [0.04-0.08]
Gender							
Male	1.00				1.00	1.00	1.00
Female	2.19 [1.96-2.44]	p<0.001			2.29 [2.05-2.55]	2.87 [2.57-3.21]	2.87 [2.57-3.21]
Calander year							
2005-07	1.00						1.00
2008-11	0.67 [0.58-0.77]						0.88 [0.76-1.02]
2012-15	0.84 [0.73-0.98]	p<0.001					0.98 [0.84-1.15]
Model parameters							
No. of individuals (n)	26,964		26,957	26,957	26,957	26,957	26,957
Total person-years@risk	128,544		128,520	128,520	128,520	128,520	128,520
No. of events (HIV seroconversion)	1,972		1,972	1,972	1,972	1,972	1,972
Proportional Hazard assumption test			p=0.95	p=0.1	p<0.001	p<0.001	p<0.001

***Values omitted from analysis due to collinearity with External migration variable**

**** Log-rank test of equality (included in the multivariate analysis if p<0.2)**

departures from proportional hazards will have a small p-value when testing the assumption. Therefore, as the deviation observed was only modest, I have retained the PH model.

Finally, controlling for calendar year appeared to have no effect on the associations observed and there was no significant difference in HR across the year categories during the study period.

Owing to data constraints, I could not include in the model all the variables that may affect HIV acquisition. The amount of 'missingness' for sexual behavioural data (i.e. number of sexual partners) was too high to yield any meaningful result (>65% missing). This was also the case for maximum educational level (>20% missing). However, given the well-documented association of HIV acquisition with migration and with multiple sexual partners, I did a sensitivity analysis including multiple sexual partners. Univariable analysis showed an increased risk of HIV acquisition in those who reported multiple sexual partners compared to those who did not (HR 1.70(1.46-1.97)). When multiple sexual partners data were included in the model, the general trends described above did not change. Furthermore, univariable analysis for education showed a significantly increased risk in those whose maximal educational level was at primary and secondary school level when compared to no education (HR 3.40 (2.75-4.19) and 4.21 (3.50-5.35) respectively), while those who had tertiary level education had a significantly lower risk (HR 0.6 (0.24-1.37)).

While wealth quintile data were not missing a proportion substantial enough to warrant exclusion from the model, it was not possible to calculate a wealth asset score for those individuals who were categorised as 'External'. Therefore, as the information for this subgroup was 100% missing, there was no way to build a model for imputation. Thus, this was not included in the main multivariable model, as it would not be possible to interpret the results. A separate univariable analysis suggested that the risk of HIV acquisition across the wealth quintiles did not vary substantially (HR range: 0.80-1.13, p>0.05 between all quintiles). Furthermore, these additional variables have been the focus of previous work at AC and therefore, are not analysed further here.

Finally, I undertook sensitivity analyses, repeating the above analysis for: (i) a dataset using randomly imputed estimated seroconversion date, rather than mid-point imputation; (ii) a

sex-stratified approach; and (iii) a migration status definition of 2 years post movement rather than 1 year. The model for each is shown in Table 8.4. The main conclusions are:

1. The HRs and trends do not broadly change when compared to the primary analysis. Restricting the window period definition and using a different method to impute seroconversion estimates does not significantly change the main conclusions i.e. external out-migration remains the greatest risk, although it becomes non-significant with random seroconversion imputation and when a shorter window period is used. The findings for expanding the migration definition to 2 years has no impact on the results obtained.

2. Sex-stratified analysis shows that for females migration does not significantly impact the risk of HIV acquisition. However, for men there is an increased risk in those how externally migrate. Furthermore, there is a general difference in hazard profile with age between sexes. The 20-25 age category has the highest risk for both sexes.

3. Age-stratified analysis shows that migration does not impact the risk of HIV acquisition in the young strata, although it does for the older strata, where any migration event increases risk. This might reflect the younger people being at high-risk anyway (by virtue of their age), and as the risk decreases with age, the effect a migration event impacts on risk. However, this is speculative without further analysis.

4. I then used the sex-stratified models to consider whether differential patterns of both migration and HIV acquisition across both sex and age groups might lead to modification of the associations seen. I considered different variations of interactive models (with the primary and secondary exposure) using AIC to compare the models and draw conclusion about the best fit (lowest score) (Appendix 8.2). The best model is the uninteracted model used in the main part of this thesis, thus there was no evidence of effect modification of the association between HIV acquisition and migration by age across both men and women.

Table 8.4: Sensitivity analysis for primary analysis

	Original model	Sensitivity models 1	2	3	4	5	6	7
		Random seroconversion	2 year migration	Shorter window period	Sex-stratified: Male	Sex-stratified: Female	Age-stratified: Young	Age-stratified: Older
	(HR [95% CI])	(HR [95% CI])	(HR [95% CI])	(HR [95% CI])	(HR [95% CI])	(HR [95% CI])	(HR [95% CI])	(HR [95% CI])
Recent migration/mobility (1 year)								
None	1.00	1	1	1	1	1	1	1.00
Internal migration	1.21 [0.90-1.40]	0.95 [0.75-1.20]	1.05 [0.90-1.22]	0.81 [0.59-1.11]	1.30 [0.79-2.15]	1.08 [0.84-1.37]	1.14 [0.87-1.46]	1.70 [1.12-2.59]
In-migration (from external)	0.92 [0.73-1.16]	0.97 [0.78-1.22]	0.91 [0.74-1.12]	0.53 [0.36-0.77]	1.04 [0.65-1.66]	0.86 [0.66-1.12]	0.94 [0.72-1.21]	1.81 [1.13-2.90]
Out-migration/External time	1.30 [1.10-1.53]	1.11 [0.93-1.32]	1.30 [1.10-1.54]	1.06 [0.84-1.32]	1.81 [1.27-2.57]	1.12 [0.92-1.35]	1.20 [0.98-1.46]	3.75 [2.75-5.11]
Background HIV prevalence								
Low	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mid	1.29 [1.14-1.45]	1.31 [1.17-1.48]	1.29 [1.14-1.45]	1.35 [1.17-1.56]	1.36 [1.02-1.81]	1.27 [1.11-1.44]	1.25 [1.07-1.43]	1.42 [1.16-1.75]
High	1.65 [1.44-1.88]	1.70 [1.49-1.95]	1.65 [1.44-1.89]	1.73 [1.46-2.05]	1.99 [1.46-2.72]	1.55 [1.33-1.81]	1.74 [1.44-1.99]	1.71 [1.34-2.19]
External*	omitted	omitted	omitted	omitted	omitted	omitted	omitted	omitted
Age categories (yrs)								
<20	1.00	1.00	1.00	1.00	1.00	1.00	n/a	n/a
20-25	1.73 [1.53-1.96]	1.64 [1.45-1.86]	1.72 [1.52-1.95]	1.88 [1.61-2.19]	3.46 [2.54-4.74]	1.49 [1.30-1.71]	n/a n/a	n/a n/a
26-40	0.90 [0.78-1.03]	0.87 [0.75-0.99]	0.90 [0.78-1.03]	0.94 [0.79-1.12]	3.16 [2.28-4.39]	0.65 [0.56-0.76]	n/a n/a	n/a n/a
41-60	0.24 [0.20-0.29]	0.22 [0.18-0.27]	0.24 [0.20-0.28]	0.25 [0.20-0.31]	1.02 [0.67-1.55]	0.17 [0.14-0.21]	n/a n/a	n/a n/a
>60	0.06 [0.04-0.08]	0.05 [0.03-0.07]	0.06 [0.04-0.08]	0.07 [0.05-0.11]	0.46 [0.24-0.88]	0.03 [0.02-0.05]	n/a n/a	n/a n/a
Gender								
Male	1.00	1.00	1.00	1.00	n/a	n/a	1.00	1.00
Female	2.87 [2.57-3.21]	2.90 [2.60-3.24]	2.87 [2.58-3.21]	3.21 [2.79-3.69]	n/a	n/a	4.36 [3.76-5.04]	1.17 [0.98-1.40]
Model parameters								
No. of individuals (n)	26,957	26,959	26,962	26,959	11,247	15,710	16,008	11,949
Total person-years@risk	128,520	128,470	128,489	128,493	46,447	82,074	53,409	75,111
No. of events (HIV seroconversion)	1,972	1,950	1,972	1,950	407	1,565	1,349	623

All parameters remain identical to the original model other than the specific parameter indicated in the sensitivity analysis

8.5 **Study 2:** Molecular analysis of genetic diversity between different migrant groups and determination of patterns of transmission within the AC epidemic from ancestral state reconstruction

8.5.1 Methods

i. Genetic diversity

In this analysis, I use pairwise genetic distance calculations as a proxy for the distribution of genetic variation between different migratory populations (objective 4).

The sample collection, sample processing and sequencing is described in Chapters 3 and 7. The pairwise genetic distances between all combinations of the AC sequences (n=1,376) were calculated based on the number of nucleotide substitutions per site, under the General Time Reversible model (GTR) using the phylogenetic package HyPhy[338]. The distances were assessed according to summary migration state of the two sequences in each pairwise combination and thus divided into the following populations if both sequences in the pair fell into the same migration state: Non-migrators, migrators, ever externally migrated, only internally migrated, and mixed populations where the pairwise comparison included two sequences in different migration states (e.g. between internal and external migrants). These calculations were used as a proxy for the distribution of genetic variation between the populations.

ii. Phylogenetic/Bioinformatics analysis - Ancestral state reconstruction analysis by migration location

The work presented in this section was undertaken in collaboration with Anna Zhukova and Olivier Gascuel at the Institut Pasteur. In this analysis we explore the hypothesis previously presented in Chapter 7, suggesting that the AC epidemic is homogenous (and embedded) within the broader South African (ZA) epidemic, and that it is seeded by multiple external introductions which lead to new clusters in the DSA, predominantly within the high-density urban areas (with high HIV prevalence), prior to subsequent spread to low prevalence areas (objective 3b). Therefore, we investigate rates and patterns of transmission between three different geographical categories –'External' to AC, or 'High' or 'Low' prevalence AC areas (defined below). This allows statistical inference to be made

of how infection has spread between these areas and the most likely source of the infections recorded in these areas.

We used the algorithm 'PASTML' (Prediction of Ancestral State Tree reconstruction using Maximum-Likelihood), developed by Sota Ishikawa at the Institut Pasteur[366,367], which uses a maximum likelihood approach to reconstruct ancestral states on a virus phylogeny, annotated with metadata defining extrinsic 'states' (e.g. risk groups or geographical locations) on the tips. This allowed inference of the most likely 'state' at the internal nodes of a tree, e.g. in this study the most likely geographical location of ancestors of each strain.

PASTML finds a compromise between the two major ancestral character reconstruction methods: joint[368] and marginal[369] posterior probabilities of a character state at each tree node. The joint method predicts the most likely unique ancestral character state per internal node. On the other hand, the marginal method proposes all possibilities of the character evolution, including ones with very low probabilities. PASTML finds a balance between the two methods by removing unlikely states, but keeping alternative highly likely scenarios.

This analysis used the RAxML phylogenetic tree described in Chapter 7 (rooted on the non-C outgroup), together with linked character 'state' metadata. The 'state' for this analysis was the geographical location of residence at the time of sequence sampling: divided into 'External' i.e. resident outside the DSA, or internal, which is further divided into 'High' or 'Low' prevalence areas according to the background HIV prevalence of the area of residence in the calendar year of sample collection, above or below the median respectively. The methods for geographical mapping and prevalence calculations within the DSA are described in Chapters 3 and 7[130].

To ensure a balanced distribution of each of the 'states' (Low, High, External), and avoid sampling biases, a randomly sampled sub-group was extracted from the total of 1,376 AC sequences to include 250 'Low', 250 'High' and 250 'External' tips. This was undertaken by collaborators using Python code to randomly select the sequences[370]. The initial tree was then pruned to contain only the selected 750 tips using ETE3[371]. This new subtree was input to PASTML to infer the most likely ancestral states of internal nodes (and their probabilities). When PASTML found more than one possible state for some of the internal

nodes, we label such nodes 'ambiguous'. A subtrees with reconstructed ancestral states was produced, as well as summarised maps, in Cytoscape[372]. In the summarised maps, we: (1) recursively merged together child and parent nodes that were in the same state, the size of the merged nodes corresponds to the number of the tips it contains and is shown on the label; and (2) kept one representative per sister tip node (i.e. nodes that were in the same state (from step 1)) - the number of merged tips included is shown (Figure 8.5). The "representative tips" from step (2) typically represent multiple independent transmissions. This resampling reconstruction of ancestral states and summarising, as described above, was repeated 10 times. As each sequence is unique to one individual (no duplicates), each internal node in the tree represents one transmission event linking the sequences, where the internal node corresponds to the parent's virus, and the tip node/s correspond to the sampled sequences (child/ren). As the sampling is incomplete, some of the transmissions and tree branches are missing.

Finally, the relative rates of transmission between each of the states are calculated based on the total number of transmissions between each state across all 10 of the reconstructed sub-trees.

8.5.2 Results

i. Viral Genetic distance

The pairwise genetic distance between all combinations of AC sequences (n=1,376) gave 915,981 different pairwise combinations. The population of non-migrators has a lower mean population genetic distance when compared to migrators (in aggregate and between different sub-categories of migrators)(Table 8.5). This is to be expected, as it is likely that the pool of circulating virus within the DSA is genetically less diverse than the pool to which migrators are exposed if infected outside the DSA. However, external migrators have a very slightly lower population mean genetic distance than internal only migrants (6.70% vs 6.71%). This difference is statistically significant, although this is due to the very large sample sizes, as discussed in Chapter 7. However, in practice, this 0.01% difference is not relevant and it is likely that these two populations are the same external and internal migrants infecting each other.

Table 8.5: Genetic distance (%) between different migration status populations

	All	Non-migrators	Ever migrated	Ever external migrant	Internal only migrant
Observations (n)	915,981	81,003	451,725	269,745	23,220
Mean GD (%)	6.65	6.55	6.70	6.70	6.71
SD (%)	1.16	1	1.21	1.19	1.29
Range (%)	0-16.7	0-16.5	0-16.3	0-14.6	0-16.3

ii. Statistical model incorporating phylogenetic tree with some epidemiological parameters - Ancestral state reconstruction analysis by migration status

A representative sub-tree generated from a random sample of 750 sequences from 'Low' prevalence, 'High' prevalence and 'External' areas (250 of each) with ancestral state reconstruction is shown in Figure 8.4. Overall, towards the root of the trees (the top), the majority of nodes correspond to external transmissions (green). This suggests that the AC outbreak originated predominantly from external introductions. Many of these 'Externals' also link to 'High' nodes (red), suggesting that the outbreak was initially spread by people from 'External' and 'High' prevalence areas. Progressing down the tree (i.e. to more recent infections), transmission begins to move from 'High' to 'Low' prevalence areas. The people in 'Low' prevalence areas generally transmit to other people in 'Low' prevalence areas and/or are located at the tips of the tree. However, there is one cluster from 'Low' prevalence areas where one 'Low' tip spread the infection to a few people in 'High' prevalence areas, as well as many in 'Low' prevalence areas (blue cluster to the bottom left of the tree). Overall, this suggests a general movement of infections from 'External' sources to 'High' prevalence areas, and subsequently to 'Low' prevalence areas.

The findings above are also presented in Figure 8.5, which shows a summary representation of a reconstructed tree. All 10 sub-trees show very similar results. The parts of the tree that are transmitting to the same state (e.g. 'External' to 'External') are merged together and the number shown represents the number of tips in the merged node. At the root of the tree, the state is 'External', which generally then spreads to 'High' nodes. Of note, there are two distinct 'High' clusters, and it would be interesting to further analyse these with additional metadata, e.g. more geographical resolution. From the 'High' nodes, infection spreads predominantly to 'Low' states. However, there are also situations of direct 'External'-to-'Low' transmission. Figure 8.6a shows the same pattern. It shows the

distribution of states at different levels of the tree, with the tips on the left (level 0 - most recent), moving to the origin on the right. The tips were sampled to include equal proportions of 'Low:High:External' (level 0), but moving towards the root, we see the 'Lows' disappear, subsequently followed by the 'Highs', leaving just 'External' nodes as the source of infection. Figure 8.6b shows the distribution of states across time, with people originating from an 'External' area being the highest proportion throughout the epidemic (except for a few recent years), with increasing numbers of people from 'High' prevalence areas more recently, and a sustained low level of people from 'Low' prevalence areas.

From these trees, we can calculate the rates of transmission between different geographical areas. We used the data from all 10 sub-trees and calculated an average transmission rate across the trees (Table 8.6). This shows that people from 'External' areas are the greatest source of infection overall. Infections generally stay within the same 'state', e.g. the highest rate of transmission occurs between two 'External' people (78.4%), as expected, followed by 'High-High' (38.6%). Very few 'External' people were infected from people in 'High' or 'Low' states. This suggests more viral imports than exports into DSA – centrifugal movement rather than centripetal. However, 'External' people also infect a substantial proportion of 'High' and 'Low' people (28.6% and 23.5% respectively), reinforcing this notion. People from 'High' prevalence areas are mainly infected by others from the same category (i.e. other 'High' people)(38.6%), but people from 'Low' prevalence areas are infected by people from 'External' areas more than others within 'Low' areas (29.0% vs. 23.5%, respectively)($p<0.001$). A significant number of ambiguous states were also identified.

Table 8.6: Rates of transmission between different migration states (mean proportion of total transmission by type of child node origin [range])

		Child node					
	From \ To	**Low**		**High**		**External**	
Parent node	**Low**	23.5%	[15.6-31.7]	6.5%	[2.8-10.5]	2.2%	[0.4-4.5]
	High	16.1%	[9.5-24.7]	38.6%	[28.6-51.0]	6.6%	[3.8-13.9]
	External	29.0%	[15.1-32.9]	28.6%	[23.6-37.8]	78.4%	[71.4-83.0]
	Ambiguous	31.3%	[22.3-36.8]	26.3%	[17.6-35.4]	12.7%	[8.1-19.2]

Figure 8.4: PASTML sub-tree reconstruction with ancestral states

This shows an example of a sub-tree which demonstrates that infection into the AC DSA originally came from 'External' sources (green nodes at the top of the tree) and then was spread largely by people from 'External' and 'High' prevalence areas (green and red), before seeding infection to people in 'Low' prevalence areas (blue).

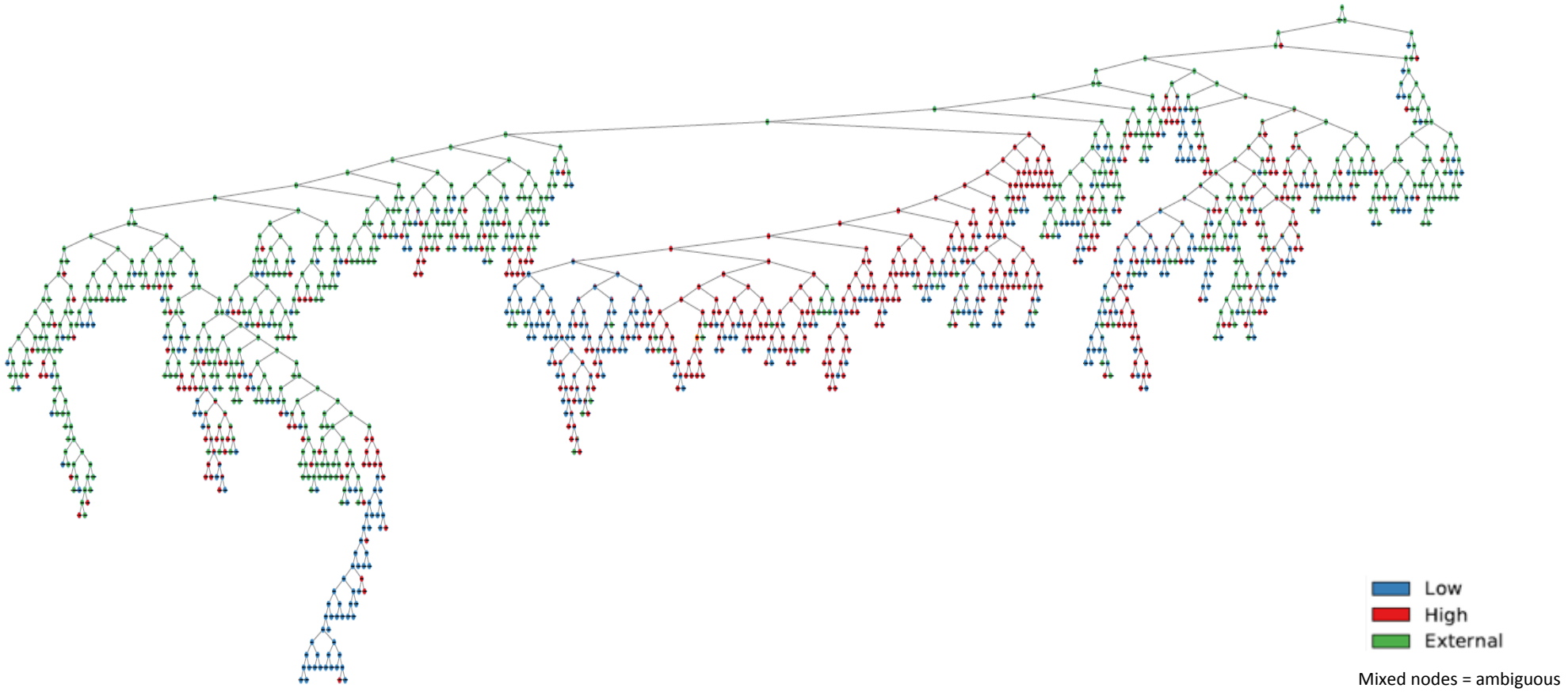


Figure 8.5: Novel representation of summary PASTML sub-tree reconstruction with ancestral states data

Depicts a collapsed representation of a tree where parts of the tree that are in the same state are merged into one. The number on the node is the number of tips (sampled sequences), with arrows corresponding to the number of transmissions to child nodes in the same state. The size of the nodes and arrows are proportioned by connectedness. The root state is External (green), then transmissions mainly spread to High (red), and from there predominantly to Low (blue). The numbers on the edges indicate the numbers or independent configurations of this type (e.g. a red tip with an edge 55 means 55 red tips like this).

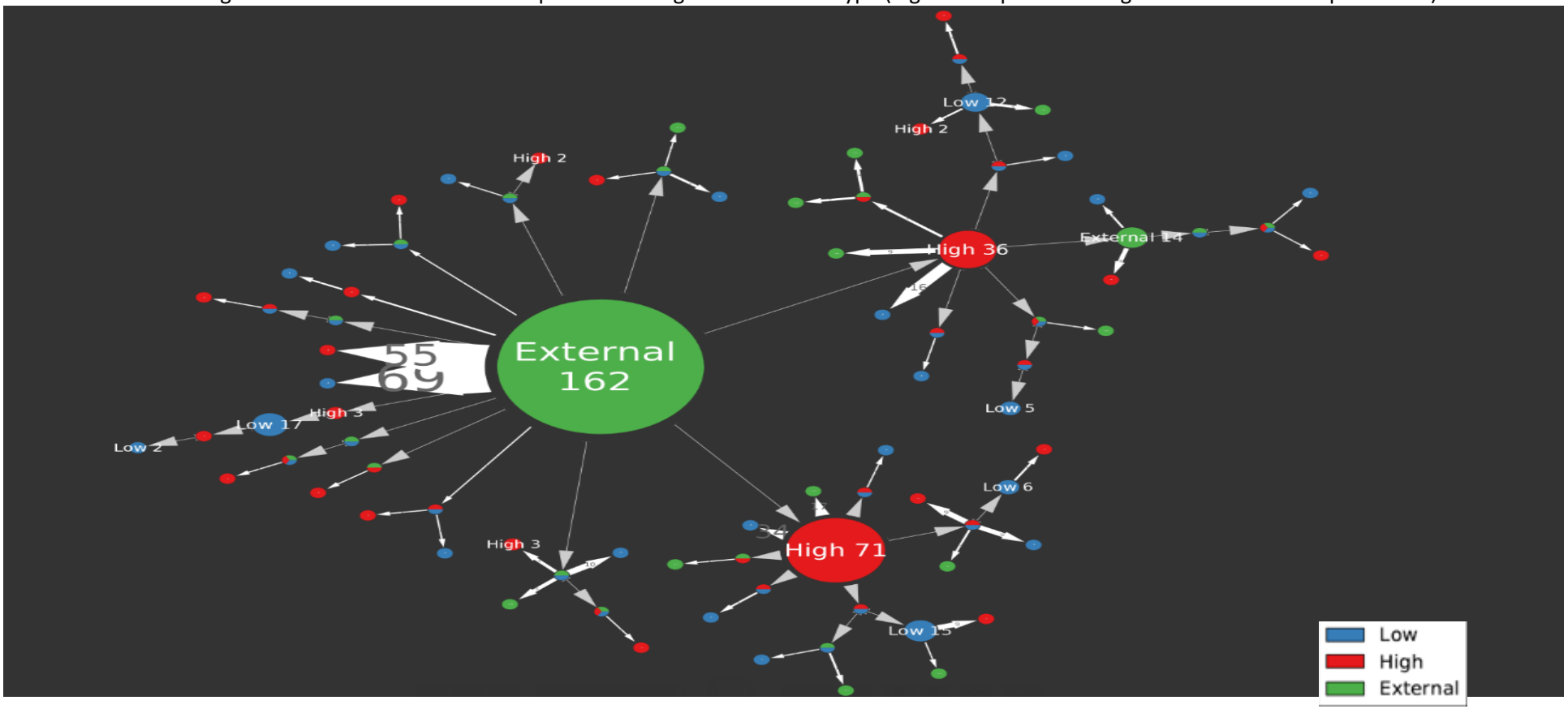
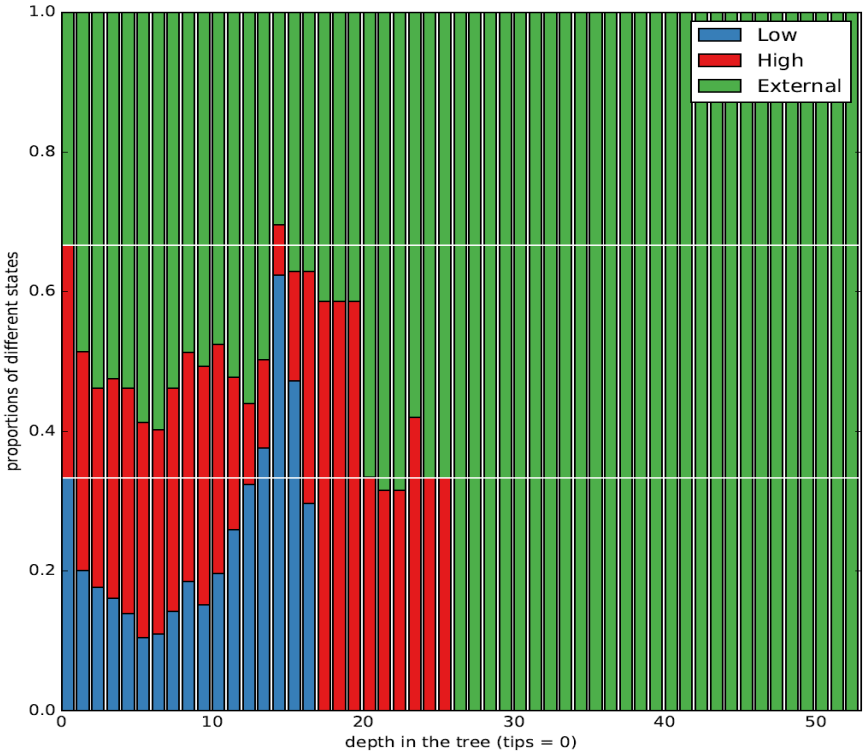
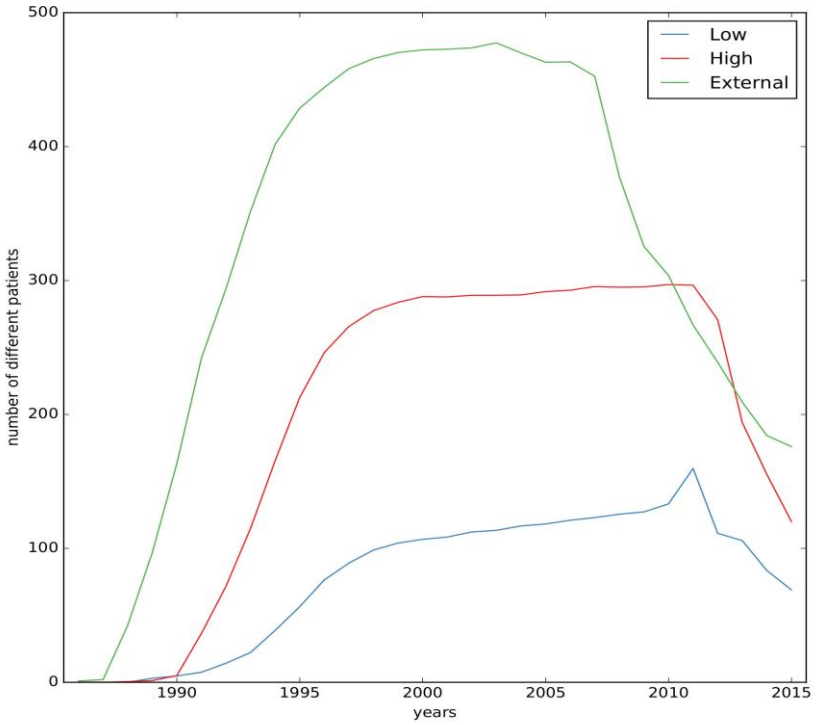


Figure 8.6: The distributions of states at different levels in the tree (6a) and across time (prevalence curve) (6b)
8.6a: At level 0, which corresponds to the tips, we have 1/3 High, 1/3 External, 1/3 Low (the random sample of 250 of each state for each tree), but moving towards the root, branches express the genetic distance (substitutions per site), firstly the Low disappears, and then the High as well, leaving us with the External nodes as the source of infection. 8.6b: Shows a prevalence curve for the numbers of individuals in each state over time. This is based on a dated phylogenetic tree, where the branches are expressed as time (instead of substitutions per site). In the case of ambiguous nodes, a proportion is counted to each state e.g. 0.5 External, 0.5 Low.





8.6 Discussion

This study confirms the findings from existing literature that migration is associated with a higher risk of HIV infection[125,128,134,137,359]. The level of risk appears to be affected by recent individual has migrated. While the effect of migration (all types) appears to be important, the large crude association observed is largely explained by the confounding variable of migration, the type of migration, and the HIV prevalence of the location to which an age. However, the location of residence impacts risk, regardless of other covariates.

8.6.1 Key findings

1. External introductions into the AC DSA are common and account for a significant proportion of new HIV infections

External sources are shown to be a substantial source of new infections overall, with viral imports, as opposed to internal circulation, playing a large role in driving and sustaining the local epidemic in the DSA. This is not only due to the founder effect of the epidemic in this region, but also because the trend is on-going, with constant re-introductions from outside the area. This indicates that migrants are important in introducing infection to the area, continually seeding new bursts of infection. In consequence, the DSA can be thought of as a 'sink' for, rather than a 'source' of, HIV infections. This aligns with evidence from Uganda, showing that a significant proportion of HIV infections within a rural community are attributed to external viral introductions[92]. This finding, coupled with a young population and many of the social drivers for HIV transmission, leads to continuously high levels of infection.

While this work suggests that the outbreak with the AC DSA appears to be fuelled by external migrants introducing infections into the area, it is important to note that the vast majority of new infections still occur in those who do not migrate (1,587 seroconversions during 112,542 person-years, compared to 385 seroconversions in 16,002 person-years for all migratory groups combined). From a public health perspective, it is important to recognise this, in order to ensure the implementation of population level interventions along with strategies targeting migrant risk groups. Although both strategies are important, the recent 'Treatment as prevention' trial run by AC showed no decrease in incidence following population treatment campaigns. This suggests that the continuous reintroduction of new infections, which is not captured by traditional public health

strategies, remains an important driver in sustaining the ongoing high levels of infection[133,355].

2. Recent migration/mobility is associated with HIV acquisition, but may be confounded.

Recent migration leads to a higher HR for HIV acquisition, with the risk being highest for those who externally migrate out of the AC DSA. This supports the hypothesis of temporary labour migration patterns contributing to the continuing high incidence of HIV in the area, with migrants being infected outside the area subsequently seeding infection back into the DSA on their return. Some of the increased risk of HIV acquisition can be explained by the age profile of migrants - young people leave home and move to urban areas to look for work. In models 1&2, migration (regardless of the type) is significantly associated with HIV acquisition. However, once we account for age, the effect of recent migration is reduced and becomes non-significant in internal migrants and in those who have in-migrated from external locations. This latter group has the biggest change in HR and shows no increase in risk compared to no recent migration, once adjusted for age. This might be explained by the fact that those who return from external migration and remain negative are a particular group of people who may have lower risk practices than the general group of people who externally migrate, given that those who externally migrate are most at risk of HIV acquisition. The change in HR across models including age is likely due to age being a significant confounder of the association between migration and HIV acquisition, where younger people are more likely both to migrate and to acquire HIV.

Additionally, age may act as a proxy for riskier sexual behaviour, e.g. a higher number of sexual partners, which may also contribute to the association between migration and HIV acquisition risk. Unfortunately the available data contains too much 'missingness' to allow credible assessment of the role played by sexual behaviour in this study. However, our tentative analysis including sexual behaviour (and thus a much-reduced sample size) does not lead to any change in the associations between age, migration and HIV, suggesting it is unlikely that age is acting only as a proxy for riskier sexual behaviour in this setting. However, given the limited quality of the sexual behaviour data, this area requires more robust data to be properly explained. It is also possible that additional factors that we do not yet know about will also confound the migration-HIV association.

Although this study has confirmed the association between migration and HIV acquisition risk, it has not shown any direct causative link. We do not know whether the increased risk is due to the mobility associated with migration, the location to which the migrant arrives, or whether migration is a marker for sexual and social behaviours which increase the risk of infection. This study appears to suggest that an element of all three factors may play a role, particularly given that riskier behaviours are associated with being young, away from home, often alone, and with being separated from family, spouse, and support networks. Further, migration is often undertaken for economic reasons, resulting in increased wealth which in turn gives access to new behaviours, for example paying sex workers.

Furthermore, risks associated with migration and HIV are traditionally attributed to young men migrating. However, this study has shown that a significant proportion of women are also migrating. Little is known about this group, about why they migrate and about the specific risk behaviours associated with these patterns. There is a need for further scientific evidence to substantiate our work and to understand this vulnerable population.

3. Background HIV prevalence influences both the risk of HIV acquisition, and where the HIV is likely to have originated from.

Transit routes have consistently been recognised as facilitating the spread of both population and disease since early in the HIV epidemic, with the notional Pan-African highway from Cairo to Cape Town being termed the "AIDS highway"[373,374]. In this study, high-prevalence areas in the DSA are directly linked to external imports given their location immediately adjacent to the N2 highway, which connects Durban in the south to Mozambique in the north (Figure 3.1). The higher rates of infection in these urban and peri-urban locations subsequently appear to spill over into low-incidence areas, suggesting that the high-prevalence areas act as bridges for HIV infections in the DSA. Therefore, those living in higher-prevalence areas are most at risk of seroconversion events as they have the greatest exposure to external imports of infection. This is true independently of one's own migration. These findings support my initial hypothesis (number 3 in section 8.2) that background prevalence of HIV influences risk, and that external introductions help to drive and sustain the AC epidemic.

Furthermore, it would be useful to understand the net flow of people within and between each HIV prevalence area within the AC DSA, and the proportion of external migrants

within each area. Unfortunately, these data were not available. If, for example, there were a net flow of people from low to high areas, and the population densities across the areas were very different and changing, this may drive the increased risk in HIV infection observed, in addition to the higher risk of a chance encounter with an HIV positive person in a high prevalence area. Furthermore, if external migrants predominantly move to high risk areas in the DSA, but have not been captured by our sampling strategy, this may also drive the effect seen. Thus, even though the effect of migration appears to be reduced in the adjusted model, migrants may still be a risk at an individual level and this needs to be considered when developing interventions. Unfortunately, the data were not available to be able to differentiate the populations between the areas, but it is important to do this to fully understand the effects observed in our study.

Reports in the literature suggest that this increased risk of HIV acquisition in high prevalence areas might be due to 'social disorganization and disequilibrium'[375]. Complex social characteristics, associated with HIV risk behaviours, are disrupted by high levels of migration, thus leading to social disequilibrium and thereby, potentially increased risk of HIV transmission within these areas. This suggests that public health interventions should not only focus on migrants as a high-risk group, but also on understanding high-migrancy settings and how these 'spaces' can be protected.

4. Rates of HIV transmission between different groups.

It is difficult to determine transmission rates between different groups using traditional epidemiological data as it is possible only rarely to infer the direction of transmission. Furthermore, the information required to track migration events is detailed, and is both time consuming to collect and complex to analyse. However, sequence data are increasingly used to attempt to address such questions - the sequences are used to infer historical events via determination of ancestral states. The associated metadata required to allow meaningful analysis are usually limited to time-specific data from the time of sampling (which is often available), rather than the detailed longitudinal data required in traditional epidemiological analysis. This study has shown that phylogenetic analysis has enhanced traditional epidemiological analysis in quantifying the extent to which migration contributes to HIV acquisition risk. Specifically, my phylogenetic analysis has allowed more precise estimation of rates of transmission, and inference of the origin and drivers of infection, both in terms of geographical patterns and in terms of the location of those

seeding infections into the DSA. By combining phylogenetic trees with extrinsic characteristics, e.g. risk groups or geographical location, it is possible to infer rates of transmission that are not easily inferred by epidemiological data alone. Therefore, by combining the two approaches, the resultant knowledge is greater than the sum of the results from each method independently[347].

8.6.2 LIMITATIONS

This study has several limitations:

1. Generalisability: The dynamics of HIV transmission are complex and are often influenced by the cultural setting. Therefore, our results may not be generalisable. This study corroborates the limited literature that exists[92], suggesting that, while these findings may be generalisable for similar rural South African settings with a culture of temporary labour migration, further studies in other settings are required to determine whether this can be extrapolated more broadly. In particular, this study does not include very long-distance migration, including those fleeing their countries of origin due to social and economic collapse, which is associated with different behaviours and risks. Generally, the distances involved in the external migration category in this study are short to medium, for example 135 miles from Durban to AC, and 380 miles from Johannesburg[347].

2. **Inability to determine causation:** Although the results are suggestive of possible mechanisms to account for the increased risk seen between migration and HIV acquisition, the results do not prove causation, as discussed above. There are many confounding factors that may explain the association seen, and further work is needed with improved datasets to fully understand these mechanisms. While migration appears to be an important factor, there are many other factors that influence risk. It may be that migration is a convenient proxy to identify these high-risk people.

3. Limitations associated with the survival analysis

Large amounts of work went into understanding the association between migration and HIV, and into developing models to advance understanding. This area is complicated by the complex social patterns involved and by the imperfect data available with which to study this field. During this work, I tried various different models to explore the limitations and benefits associated with each. The model I have used in this analysis is a plausible candidate to address the research question. However, it has limitations, and no single model can perfectly capture underlying events. For example, I acknowledge that this

cohort analysis has assumed that external migrators lost to follow-up (LTFU) have the same risk of HIV seroconversion as those remaining within the study (the baseline assumption of the model is that the risk is the same and those LTFU are lost at random). This is contrary to the hypothesis being investigated, as it is likely that those who migrate externally and are lost from the cohort have a higher risk of HIV acquisition. This is a limitation of the type of analysis undertaken.

4. Difficultly determining date of infection: One of the main limitations is the imperfect knowledge of timing of infection. This may result in misclassification of seroconversion events between the time-dependent variables, which could alter the results seen. This is likely to be particularly true between the external out-migration and in-migration categories as people cannot be tested while living externally. Therefore, the higher risk observed in external out-migrators is likely to be an underestimate. In the absence of regular testing with short inter-test periods, there will always be difficulty in attributing a specific time of infection. However, I did a sensitivity analysis with shorter window periods (between last negative and first positive test) using random imputation to estimate the date of seroconversion to ensure the results were as robust as possible. This showed on material change to the general conclusions drawn.

5. Limitations inherent in longitudinal demographic data collection: The sampling strategies and differential participation in the data collection, as described previously (in Chapters 3 and 7), often leads to high levels of missing data, particularly sexual behaviour data. When the proportion of 'missingness' is high or data are not missing at random, this limits both the reliability of the results and the range of possible analysis. Furthermore, the quality of some sexual behaviour data are likely to be poor due to the face-to-face interviews undertaken to collect the data, as these may inhibit accurate reporting[376-379].

6. Ascertainment bias: This results from sequence data only being available in those with viral loads >10,000, which over-represents new infections and those not on treatment (as discussed in Chapters 2 and 3).

7. Sampling strategy: Specific to this study, the classification of high or low prevalence areas could introduce a bias, particularly if there is differential participation between those who live in high and low prevalence areas. Furthermore, younger people are likely to live in the urban high prevalence areas where there is more work. Given our work suggests that age is a confounder for the association between migration and HIV, this may over-represent the increased risk of HIV infection in high prevalence areas.

The definition of high and low prevalence areas was based on the yearly inter-quartile range and these figures are not standardised across the areas for population density, age, or sex. High prevalence is, in part, dictated by population density. Had the data been available, I would have controlled for this.

It is also important to consider the population we have not been able to sample and to explore whether lack of data concerning this population may bias the results. It is possible that those who have not had an HIV test during the study period are systematically different to those included in the study. For example, they might have been less likely to interact with the DSS and health systems more generally. This differential health belief may mean this group has a higher rate of HIV acquisition. Furthermore, they may also represent a population with more migrants who are mobile and harder to find.

Fewer data are available on those who migrate externally for the time periods that they are outside the DSA. Therefore, HIV incidence and other data are more likely to be misclassified than in those who remain in the DSA. For example, external seroconversion events may be 'diluted' across the time when an individual is outside the DSA, as they are unlikely to test during this time. Thus, the estimates obtained for external migrants are likely to underestimate the true risks.

Of note, this study does not capture migrants who originate outside the DSA, but have spent time within the DSA (i.e. reverse temporary migration into the DSA), unless they form formal ties and are considered resident in the area. In particular, HIV positive migrants visiting the area temporarily are not captured, and this may underestimate the role of this group in introducing and driving the epidemic.

8. Lack of bioinformatics tools to handle large datasets: This is a significant limitation in undertaking studies such as this. One approach to address this limitation is to undertake a random sampling to produce a smaller dataset on which to carry out the analysis, but which allows inferences to be made regarding larger datasets.

8.6.3 Conclusion

There is a shortage of studies on migration and health and, in particular, on the relationship of migration to HIV. This study confirms that recent migration is associated with increased HIV acquisition risk[125,128,134,137,359], although the estimated association was greatly reduced once adjusted for age. However, the role of migrants in the current HIV epidemic must be further assessed and quantified. Current treatment policies fail to take

HIV transmission dynamics into account, and failure to provide testing of and treatment for mobile and migrant populations is likely to jeopardize the efficacy of interventions among key populations that share sexual networks with migrants. Although migration as a risk factor for HIV acquisition is confounded by multiple factors, it may still be a good proxy marker to identify those at increased risk of infection for the purposes of targeting prevention strategies. Without an effective prevention strategy including coverage of high-risk populations, it will be difficult to control the ongoing HIV epidemic.

However, targeting prevention strategies towards the migrant population creates a number of challenges. Firstly, this study highlights the need to consider combination prevention strategies due to the multiple factors contributing to increased risk of HIV acquisition. For example, a successful strategy might involve a combined approach including: education on both HIV prevention and the risks associated with migration, access to condoms, pre-exposure prophylaxis, and interventions that are accessible and acceptable to a young target population. Secondly, it will require the development of strategies and facilities that are flexible enough to be able to deal with migrants, rather than the fixed clinic models currently in place.

The work presented in this chapter provides evidence to substantiate and support refinements to national policies to benefit the migrant populations. I have generated evidence to show that by triangulating the information obtained from both epidem - iological and phylogenetic approaches, the resultant knowledge is greater than the sum of the results from each method independently. The addition of molecular data provides a refined understanding which furthers knowledge on which to optimise public health strategies. However, while this work has the potential to change policy on treatment access for migrant groups, it could also have a negative impact and lead to greater stigmatization. Therefore, high ethical standards must guide this research and care needs to be taken to ensure that privacy is maintained and any consequent risks are mitigated in studying these vulnerable groups. The ethical issues are discussed in Chapter 9.

Acknowledgements: The bioinformatics work presented in this chapter was undertaken in collaboration with Anna Zhukova, Olivier Gascuel from Institut Pasteur.

CHAPTER 9

ETHICAL CONSIDERATIONS IN HIV PHYLOGENETIC RESEARCH

The application of phylogenetic approaches to public health programmes has increased rapidly in the last five years. The use of these methods presents unique ethical, legal and social challenges which are not well addressed by the existing bioethics literature. Therefore, there is a need to develop an effective and sustainable model of good ethical practice in phylogenetic research, which will help minimise the risks to individuals and communities, as well as optimise the scientific and public health benefits. In this chapter, I explore the issues arising from the design, conduct and use of results from phylogenetic studies. In addition, I propose recommendations to minimise the associated risks both to individuals and to groups. Although this chapter is focused on phylogenetic studies of HIV in Africa, the conclusions are considered in a broad context and should be applicable to phylogenetic studies in general.

9.1 INTRODUCTION

Despite advances in HIV prevention, HIV incidence remains high, notably in sub-Saharan Africa which accounts for 75% of all new HIV infections worldwide[380]. To improve global HIV outcomes, new methods are needed to develop HIV prevention, treatment and care services which are evidence-based, innovative, and targeted. Such methods will need to improve our understanding of the paths and patterns of HIV transmission. As discussed in previous chapters, one novel technology that can provide such insight into transmission patterns is phylogenetics. Moreover, the combination of molecular tools and traditional epidemiology has demonstrated the potential to answer critical questions that are not easily addressed by traditional or molecular approaches alone[4,91,92]. For example, in the context of the HIV epidemic, understanding the characteristics of transmitters (determined by understanding who infects whom) remains an important challenge, which, if addressed, would enable the development of targeted prevention strategies[381].

Phylogenetic studies of HIV epidemics require sizeable datasets. One such dataset has been generated by the Phylogenetics and Networks for Generalized HIV Epidemics in

Africa (PANGEA-HIV) Consortium[65], which has used next generation sequencing (NGS) methods to generate approximately 10,000 near full length sequences, linked to clinical, demographic, and epidemiological data to assess the transmission of HIV in sub-Saharan Africa. NGS increases the potential power of phylogenetic approaches compared to the shorter length HIV sequences routinely generated by Sanger sequencing for drug resistance testing[382](outlined in Chapter 2).

The sharing of phylogenetic data on HIV has the potential to make a significant contribution to a more sophisticated and timely understanding of transmission patterns. Funding bodies such as the Wellcome Trust, the Bill & Melinda Gates Foundation and the National Institutes of Health are committed to sharing data to maximise the benefit of generating such data. Additionally, HIV sequence data used for publication in scientific journals are required to be submitted to GenBank. Furthermore, global health and HIV institutions, such as WHO and UNAIDS, have called for novel approaches to benefit HIV prevention and treatment, which may include granular approaches through phylogenetics[383]. However, the collection, storage, sharing and research use of such data raises important ethical, legal and social challenges. These issues need to be identified and addressed to facilitate the development and implementation of an effective and sustainable model of data sharing, capable of commanding well-founded public trust and confidence. This framework should be adopted into models of good ethical practice.

International ethical guidelines for research with human participants such as the Helsinki Declaration and the Council for International Organizations of Medical Sciences (CIOMS) guidelines address a number of these issues, including the need for informed consent, community engagement, risk minimization, and for considering the risks and benefits of research for groups and communities[384,385]. In addition, since the Human Genome Project launched in 1990, a large and diverse academic and policy literature has been generated on the ethical, legal and social implications (ELSI) of the research and clinical uses of genomics, historically focused on human genomic issues in high-resource settings[386]. In recent years, this has begun to be accompanied by a growing bioethics and social science literature on the ethical implications of genomic research in low- and middle-income countries (LMIC)[387-389]; the key documents, position statements and initiatives are summarized in Table 9.1. Much of this literature has emerged from the ethics programmes of large genomic research initiatives such as the Malaria Genomic Epidemiology Network

(MalariaGEN) and H3Africa[390-393] . Such literature has often drawn upon frameworks for good practice in research in low-income countries (LIC)[394] and formal international bioethics guideline documents [384,385]. However, until recently, very little had been written on the ethics of data sharing in genomics research in LMIC; some literature has now begun to emerge, again largely focusing on the experiences of particular research networks[395]. Perhaps most usefully, some recent work has begun to explore the attitudes to good practice in data sharing among key stakeholders in LMIC[396,397].

Whilst many of the ethical issues identified and discussed in this literature are likely to be of relevance to research in phylogenetics of other pathogens, HIV phylogenetic research presents many new and complex ethical issues not previously encountered. In recognition of this, and of the need for the development of models of good ethical practice in this area, I co-organised a two-day expert multidisciplinary (scientists, bioethicists, lawyers, human rights advocates, HIV activists and community engagement members from Africa) workshop, held in London in May 2017[398].

The meeting focused on identifying the critical issues arising from designing, conducting, or using results from HIV phylogenetic studies, and on making recommendations with regard to publicly releasing and publishing data obtained from HIV phylogenetic studies in an ethical manner.

This chapter summarises the findings and recommendations made at the meeting and provides a framework for both researchers and funding bodies undertaking viral genetic studies. While these recommendations may have broader applicability, the focus is on HIV phylogenetic studies in Africa. I have included it in this thesis as it is relevant to this research and addresses many of the issues raised by this work.

Table 9.1: Summary of key documents, position statements and initiatives relevant to ethical issues of HIV phylogenetics and referred to within the document

Document/ Position statement/ Initiative	Description
The Declaration of Helsinki, 1964 and updated in 2013[384]	The World Medical Association developed the Helsinki declaration as a statement of ethical principles for medical research involving human subjects, including research on identifiable human material and data. This declaration states that the interest and well-being of the individual takes precedence over the science and well-being of communities and populations. Many of the principles are relevant to performing HIV phylogenetic studies.
The Council for International Organizations of Medical Sciences (CIOMS) International Ethical Guidelines[385]	CIOMS is an international nongovernmental organization in official relationship with WHO, founded in 1949. The guidelines aim to provide internationally vetted ethical principles and detailed commentary on how universal ethical principles should be applied, with particular attention to conducting research in LIC. There have been four revisions of the guidelines since they were first published (1982), to take into account scientific developments and bring the guidelines into line with current thinking on ethics and human rights.
Wellcome Trust report on Ethical sharing of health research data in LMIC: views of stakeholders[397]	This report aims to provide evidence to inform the development, implementation and evaluation of data-sharing models and identify further research priorities. It is based on a multi-site collaborative study of stakeholder experiences and views in LIC of best practices in sharing individual-level data from clinical and public health research.
The Human Heredity and Health in Africa (H3Africa) Initiative[391,392]	H3Africa Initiative aims to facilitate a contemporary research approach to the study of genomics and environmental determinants of common diseases, with the goal of improving the health of African populations. To accomplish this, the H3Africa Initiative aims to contribute to the development of the necessary expertise among African scientists, and to establish networks of African investigators.
The ELSI (Ethical, Legal, and Social Implications) Program[386]	The ELSI Program is a multi-disciplinary program funded by the National Human Genome Research Institute at NIH. It focuses on exploring ELSI of human genomics, and developing policy options to address these implications, although the scope has broadened over years in response to rapidly evolving genomic technologies, legal and commercial developments, and translation to clinical applications. Many of the issues from human genomics also apply to viral genomics: Biobank governance is a particular focus of the "ELSI 2.0" initiative[399].

9.1.1 Phylogenetics and its role in HIV research

A background to phylogenetic analysis (the science, methods and associated assumptions) is given in Chapter 2[44-46,50,51,92,400,401]. Box 9.1 reinforces some relevant methodological considerations. The biology of the HIV virus lends itself to phylogenetic analysis, as it is a highly genetically variable virus that frequently changes to escape the host immune response[402]. The sexual transmission of HIV involves two individuals (couple) and the

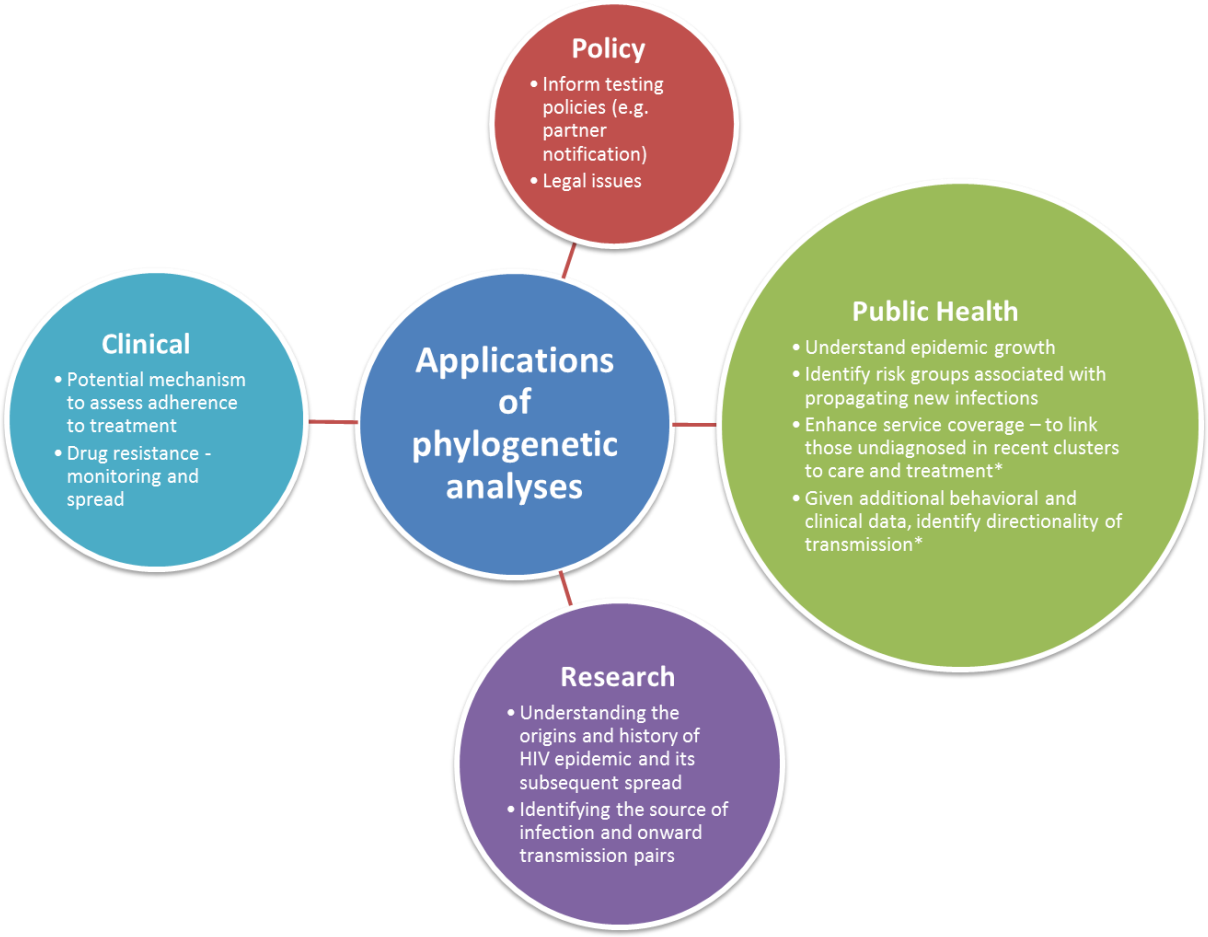
variability of the HIV virus is used to infer linkages forming “phylogenetic clusters” between couples and groups of people[52]. Caution is required in interpreting phylogenetic transmission chains, as unsampled cases may act as either a common source of infection, or as an intermediary in a transmission chain for hosts infected with genetically similar pathogens. Also, whilst the probability of common source or intermediate transmission events can be quantified, it is currently not possible to either definitively confirm or deny putative transmission pairs with genetic data alone. These limitations are relevant to understanding the ethical, legal and social implications of the technology.

Box 9.1: Key methodological considerations for constructing HIV molecular transmission clusters

- Phylogenetic support methods use bootstrap or posterior probability to identify groups more closely related to each other than to the rest of the population being analysed:
    - Bootstrapping: a statistical resampling method of random sampling of nucleotide sites with replacement. This process is repeated multiple times and the frequency of identical branch reproduction gives a bootstrap value indicating the robustness of the tree.
    - Posterior probability (PP): combines the prior probability of a tree with the likelihood of the given data to indicate the probability of the tree to be correct. The highest PP will represent the best phylogeny.
- Phylogenetic distance methods identify groups whose mean/median/maximum genetic distance suggests a common ancestor in recent time.
- Pairwise genetic distance methods identify individuals whose viral genetic distance implies a direct or indirect transmission event and combines these individuals into clusters.
- Molecular clock methods indicate the timing of the most recent common ancestor, which can contribute to understanding the timing of infection.

Phylogenetic analyses can be used widely in HIV epidemiology (see Figure 9.1 for details). For example, they can be used to study viral linkage and risk factors for epidemic spread (molecular epidemiology)[91], the growth/decline of the HIV epidemic (phylodynamic tool)[92-95], or the impact of migration on HIV spread and to identify hubs of transmission (phylogeography)[96]. Phylogenetic analysis can be extremely powerful if combined with traditional epidemiological methods, and the level of detail provided for each of these applications is expected to increase further with the development of more sophisticated mathematical models and computer simulations[351].

Figure 9.1: Applications of phylogenetic analyses



**Denotes a potential future use of phylogenetic analyses*

HIV phylogenetics has been most developed in high-income countries (HIC) with greater scientific infrastructure and where HIV is characterised by small epidemics, focused on specific risk groups, (often referred to as "concentrated epidemics"). Since sequencing of the HIV pol gene is recommended to monitor both transmitted drug resistance and emerging drug resistance on antiretroviral therapy, data are available on a large proportion of diagnosed individuals[64]. This has led to the growth of national HIV genetic databases, such as in the UK[403] and Switzerland[404], which are amenable to molecular epidemiological analysis through reconstruction of phylogenetic trees. If such datasets are linked to epidemiological surveillance and clinical cohort data, better inferences can be made with regard to both patterns of exposure and risk factors in infected and non-infected individuals. Unlike standard epidemiological data, molecular data can also allow inferences to be made from the time of transmission relative to the time of sample collection. Furthermore, the data obtained through phylogenetic analyses can be used to

validate self-reported epidemiological data in relation to sexual and other behaviours. Combining both traditional epidemiological tools and phylogenetic research is, therefore, a powerful tool for better understanding epidemic characteristics, enabling more effective and better targeted programmes for prevention, treatment, and other public health policies.

In contrast to the US and Europe, many African epidemics are much larger, and currently the sequencing of virus is not routinely undertaken. Furthermore, compliance with and the effectiveness of reporting and surveillance programmes remain limited. At best, initiatives such as PANGEA-HIV, analyse a small proportion of infected individuals in any given population. In addition, community and patient mobilisation around HIV takes very different forms compared to that in the US and Europe, and the social, political and economic context is significantly different and varies among African countries. Therefore, ethical analysis of phylogenetic work will need to take into account characteristics of the epidemic in different countries, as well as the local context.

9.2 Key ethical issues arising in phylogenetic studies of HIV transmission

Some of the ethical issues raised by HIV phylogenetic research are similar to those found in traditional epidemiologic studies. These include potential for stigmatization and risk of social harm to individuals or groups, and concerns about privacy, confidentiality and security of data. However, there are risks of harm that are particularly salient in phylogenetic research, such as the potential to provide information about people or groups of people, who are not research participants, and their behaviours within a linked network.

Of particular concern is that with only biological samples and minimal clinical and demographic information, complex social and sexual relationships may be deduced from sequence analysis. In contrast, traditional epidemiological studies would require far more information in order to draw inferences about transmission of HIV between individuals, particularly with respect to directionality of infection. In addition, with traditional epidemiologic data, there is often uncertainty about accuracy of self-reports of sexual or other contacts.

Another salient issue is the requirement from funders and publishers to share data. While this can maximize the scientific research undertaken on a dataset, it raises concerns about how the data are used, appropriate consent for such use, confidentiality of participants, and stigma relating to identified individuals and population subgroups. Hence, it is paramount to address ethical tensions (both risks and benefits) in this research and develop a working consensus on handling these challenges as science moves forward. In this section, we address the main ethical problems arising in phylogenetic research.

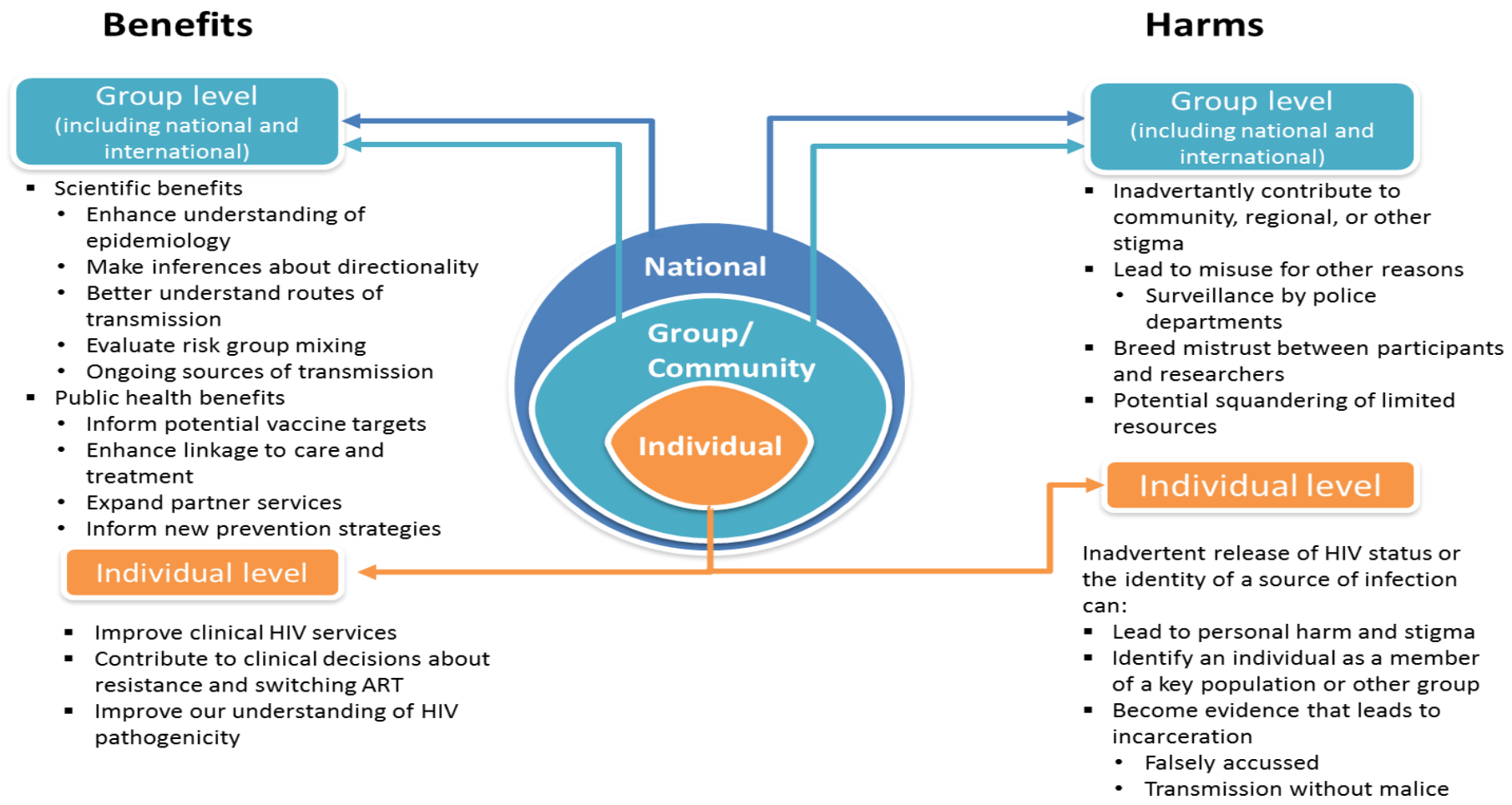
i) Risk and benefit assessments

As with any research, information obtained through phylogenetic analysis should be used to advance important, socially valuable goals, such as reducing the spread of HIV, whilst at the same time minimising the risks to individuals, groups, and populations. Harms and benefits will vary depending on whether they are assessed at the individual, group or societal level (Figure 9.2).

Risks to individuals would principally arise from either inadvertent or intentional disclosure of HIV status or transmission events, or demands for these data for judicial or extra-judicial targeting of individuals or groups. In a number of countries, phylogenetic evidence is indeed being used in criminal cases of alleged HIV transmission[21,405,406]. Breaches of confidentiality could occur through inadequate anonymization or deductive disclosure, or through misinterpretation, miscommunication, or misuse of the analytic results, or through legal action. These risks will increase if more data are generated and made easily available without adequate oversight, as sequencing technology becomes more mobile, faster and cheaper.

The key benefit of HIV sequencing to individuals arises from improvements to antiretroviral drug regimens, based on drug resistance mutation testing. Indeed, most HIV phylogenetic studies to date have been from data obtained routinely for clinical drug resistance testing, from surveillance programmes, or as part of broader research studies. At the same time, the scientific power of phylogenetic analysis allows for critical inferences potentially linking individuals' data to others in a network, enabling inference about the characteristics of networks and identification of risk groups. Additionally, NGS data might allow for assigning the directionality of transmission within a cluster[66]. This

Figure 9.2: Potential benefits (left) and harms (right) associated with HIV phylogenetic analysis



information could be used to focus public health interventions towards specific groups at high risk of both acquiring and transmitting the infection. Thus, the same data can be used for both personal and public health benefits. However, identification of these risk groups can also lead to stigma and persecution of such groups and directionality of transmission makes individuals liable for prosecution. Box 9.2 outlines a case study of migration in Botswana, as an example of the potential social harms associated with risk group identification.

Box 9.2: Migration in Botswana

Studying migrants is often fraught with both logistical and ethical problems. Migration has been identified as a key risk factor for the spread of HIV, possibly because of the lack of access to culturally and linguistically appropriate prevention information and clinical care, disruption of established social relationships and the potential for increased risky sexual practices when people are away from home[125,134,347]. However, a more nuanced approach to migration and the link to HIV is needed. Grouping mobile people together as "at risk", and particular places as "hot spots" overlooks the differences in migration flows and different risk environments. It is also important to differentiate between the characteristics of the areas between which migrants move.

Migratory populations in Botswana (documented and undocumented, skilled and unskilled) face challenges in accessing health care services, and are prohibited from receiving government provided free antiretroviral therapy (ART) drugs[361]. According to the national census[408], migrants accounted for approximately 14% of employed and 9% of unemployed populations in Francistown, Botswana. A significant proportion of HIV-infected migrants in Botswana are unaware of their positive HIV status, are not on ART, and may disproportionally contribute to new HIV transmissions. It is likely that HIV-infected migrants in Botswana have disproportionately high levels of HIV-1 RNA, and have the greatest risk of HIV transmission. Yet, the level of integration of migrants in the generalized HIV epidemic in Botswana is unknown due to exclusion from research studies. Phylogenetic studies are able to address this question. However, there are significant ethical issues and risks related to the migrants' participation in research, as migrants represent a vulnerable population, many of whom may be stigmatised, marginalized or disenfranchised, as well as subject to deportation, imprisonment or extortion, including for migrant sex workers, sexual exploitation in exchange for short-term visas[409]. Obtaining consent can inhibit participation due to privacy and identification issues due to their undocumented status. Therefore, it is ethically preferable to enrol migrants anonymously to avoid these social harms.

HIV does not have borders. However, current treatment policies in Botswana fail to take HIV transmission dynamics into account. Failure to provide ART treatment to migrants is a major ethical issue and violation of human rights[374], but is also likely to reduce the impact of Treatment-as-Prevention programmes and curtail the efficacy of interventions among key populations that share sexual network with migrants. The role of migrants in the current HIV epidemic must be assessed and quantified. High ethical standards must guide this research, as there is a risk that the results may lead to further stigmatisation e.g. driving sub-epidemics. Therefore, caution is needed as including migrants in a study increases the stakes; while it may convince governments to provide treatment, it may also risk individual expulsion from countries or lead to an adoption of policies that are discriminatory, such as routine screening on entry into the country for migrants. Health programmes for migrants need to be cross-border initiatives. While ethical barriers and challenges are still precluding the generation of evidence to inform health policy, there is a collective responsibility to change the narrative on migrants.

The decision to anonymize data used for phylogenetic analysis may present further ethical questions. Anonymization protects against individual disclosure risk, and resulting stigma or harmful social or legal consequences. Even with anonymization, there is the theoretical possibility to deductively disclose someone's identity via both HIV sequence data and non-phylogenetic data. For example, the HLA of the infected individual is imprinted on the virus due to immune selection and has the potential to allow individual identification in the future[407]. However, the likelihood of such deduction when data are fully anonymized is low, but should be acknowledged. Notwithstanding this, maintaining a link to individuals' identities may allow for direct benefits to individuals, for example, when sequence information is not currently available and would provide clinical guidance, or when information about risk groups or transmission patterns may provide a benefit to the individual in their relationships with others.

Furthermore, phylogenetic studies present at least two types of risk to groups/communities. Firstly, these studies may reveal information about the unsampled population. Secondly, if presented without care, communities may be labelled as hot spots for HIV[361]. The existence of high risk areas or groups may lead to discrimination, violence or other adverse consequences. There is also the potential for risks to be exaggerated or misrepresented through media reporting. Detailed scientific nuances or uncertainties may also be lost due to miscommunication, misunderstanding, or oversimplification in reporting of research findings.

The choice of variables used in phylogenetic analysis is an important ethical decision, by focusing on different levels and types of drivers of the epidemic. Phylogenetic analyses are often based upon individual-level demographic, behavioural or clinical variables, ignoring structural and environmental factors. This creates the misperception that certain groups (e.g., key populations (men who have sex with men (MSM) or people who inject drugs (PWID)) or others) are responsible for infecting others and sustaining the HIV epidemic. In contrast, other structural factors, which can also be understood as human rights violations, such as sexual violence, lack of access to prevention and treatment, and having experienced discrimination, may play a significant role in HIV transmission[374]. Studying these factors and their effect on HIV transmission risk can decrease a blaming mentality and create alternative understanding of how to reduce HIV transmission, and which individuals or groups are most at risk and why.

Plans for addressing risks to individuals and to groups should be developed in the planning stages of research projects. For protection of individuals, particularly the risk of criminal prosecution or other targeting based on either HIV status or HIV transmission events, anonymization of data provides considerable protection. While it is theoretically possible that individuals could be identified through re-analyzing and re-linking anonymized data from different sources, it would be difficult and require specialized expertise. In contrast, datasets with individual identifiers still linked could be subpoenaed or obtained through unauthorized means, putting individuals at risk.

Researchers, therefore, need to assess carefully the potential of identifying specific groups of people from their data, whether this identification could provide benefit in informing targeting interventions, and whether the benefits of the intervention outweigh the risks to individuals in groups by being identified. If there are other approaches that achieve the same research objective, but involve less risk, they should be preferred. In any case, an ongoing monitoring of anticipated and unanticipated risks should be built into HIV phylogenetic research, and mitigation strategies identified as early as possible.

ii) Protection of the rights and interests of study participants, while in pursuit of scientific progress and improvements to public health

Effective phylogenetic work often takes place at the interface between research and public health practice: the same data can be used for both purposes. Researchers are typically viewed as obliged, so far as possible, to protect individuals who enrol in a study from risk of harm while pursuing valuable knowledge. In contrast, public health agencies have the mission of protecting the health of the public, which sometimes involve overruling individuals' privacy interests to use data for public health decision making. In cases in which research also has implications for specific groups, further considerations relating to group harms are important. There are also research studies and public health activities that are designed to protect or enhance population health, but which may not be designed to deliver maximal benefit to individuals involved in the research[410-412]. In phylogenetic analyses, balancing commitments to scientific progress and protection of rights/interests of study participants may also present similar challenges. Box 9.3 outlines some research and clinical scenarios where special considerations of these challenges are required.

Box 9.3: Research and clinical scenarios highlight examples where undertaking HIV phylogenetics require special considerations

Serodiscordant couples: HIV transmission often occurs within discordant couples in which one person is HIV-positive and the other is not. Phylogenetic analysis allows identification of linked transmission between the members of the pair. In some HIV studies enrolling discordant couples, such confirmation is critical to the interpretation of effectiveness of interventions, for example, when interventions to reduce transmission are focused on the HIV positive partner. In such cases, it is important to ascertain whether HIV acquisition events within a discordant couple represent infection from an outside sexual partner[381,413,414]. This information is scientifically valuable, yet has the potential to cause social harm and distress to individual couples. Often such information is not made available to research participants for this reason.

Detection of acute or early HIV infection: In most countries, antiretroviral therapy (ART) is now initiated immediately on diagnosis, rather than determined by the CD4 cell count. Early treatment prevents complications of infection and stops onward transmission. But people with acute and early HIV, which can be recognized in molecular analysis as a lack of within individual viral diversity (which increases over time following infection), are still establishing an HIV reservoir, and are maximally contagious. Accordingly, molecular analysis that recognizes acute and early infection has immediate individual and public health actionable implications[415]. Researchers may need to consider what obligations they would have to notify individuals of their status as acutely infected so that appropriate counselling and care including ART can be provided. Partner tracing should also be considered.

Detection of transmission events in non-treatment-compliant patients: If antiretroviral therapy does not suppress viral replication, some new HIV transmissions from those receiving treatment could occur. The most likely reason for lack of viral suppression is poor adherence to ARV therapy, although transmitted resistant strains can also be the cause. Therefore, phylogenetics has the potential, when linked to clinical data, to identify such cases. This conflation between a patient receiving therapy and being strongly encouraged to fully adhere by their clinic, with the finding that they have infected someone else due to non-compliance, raises further ethical issues. How does a health worker respond to these data? What is the relative value given to supporting the individual in clinical care, compared to the transmission potential associated with sub-optimal antiretroviral compliance?

Furthermore, evaluation is needed to define the obligation of researchers to the study participants with respect to communicating results. In some clinical studies, there is an obligation to provide results that require clinical action. In general, this goal is currently theoretical because phylogenetic results are produced with a significant delay from sampling, therefore, any result would no longer be timely in informing clinical care. However, with the evolution of real time phylogenetic data, the situation can be expected to change. Obligations might arise to report drug resistance data to study participants as *pol* gene sequencing is commonly used for phylogenetic analysis, if this information was not previously available in the HIV care setting. A second potential issue is the source of HIV acquisition in discordant couples. As outlined in Box 9.3, one study of discordant couples found that 30% of the HIV acquisition events were not linked to the known infected partner[381]. While this information was critical for interpreting the efficacy of the prevention strategy tested in that particular trial, these results were not provided to study participants for fear of adverse consequences for the individuals involved[416]. For example,

domestic violence, loss of trust in relationships, and relationship break ups might all result from disclosure of partnerships. This example highlights the complexity of revealing transmission events between individuals in the course of research; in practice there might be no way to use this information in further counselling or follow-up with participants without increasing risks.

iii) Local social and legal context, including human rights violations

An important part of understanding risks and benefits associated with phylogenetic research is understanding the local social and legal context. These factors are vital in determining whether phylogenetic results may contribute to stigma, discrimination or the violation of human rights. For example, knowledge of local legal proceedings can help by understanding how to make the data unavailable to subpoena, and therefore, dramatically reduce the level of individual risk. The risks and benefits should be well understood by the researchers and all actors involved in projects involving the use of phylogenetics in the context of public health. It is also critical to keep in mind that social and legal context can change over the course of a research project, and periodic or on-going monitoring of the legal environment may be important to ensure that new risks are not introduced during the study. Finally, researchers need to be clear with policy-makers how proposed laws or policies could negatively impact HIV research efforts and interventions.

Persecution of key populations in Africa remains widespread, although the size and scope of the problem is poorly understood[28,417,418]. A 2014 survey of HIV-related stigma in South Africa found that one in three people living with HIV (PLWHIV) experienced discrimination, and close to half reported internalized stigma[419]. At a national level, in February 2017 senior Tanzanian ministry of health staff promoted a campaign targeting the lesbian, gay, bisexual, and transgender (LGBT) community, threatening to publish the names of individuals and directed the police to arrest and subject at least one individual to a forcible anal exam[420]. Indeed, HIV is often singled out as the exemplar infection violating human rights, often due to association with political agendas. Research methods have been used to violate individuals' rights, including the use of key population mapping by police to arrest and harass sex workers and MSM in Nigeria in 2014[421], impose travel bans on foreigners, enforce restrictions on access to housing, schooling and employment, and trigger violent attacks, including murder. There is a global human right to health[384,422], including prevention and treatment, and a right to privacy, consent, freedom from

discrimination and violence. Human rights violations against PLWHIV continue today and those affected have a right to due process and redress for these violations. Box 9.4 highlights key legal and human rights considerations that should inform the use of phylogenetic tools for public health. These considerations are based on norms provided under international human rights treaties, as well as national constitutions and legislation.

Box 9.4: Key legal and human rights considerations that should inform the use of phylogenetic tools for public health research

- Informed consent for collection and dissemination of phylogenetic data and information
- Confidentiality, safety and prevention of un-authorised use of phylogenetic data and information
- Non-stigmatisation and non-discrimination in collection and publication of phylogenetic data.
- Attention to criminalisation and other potential negative consequences relating to collection and dissemination of phylogenetic data
- Specific gender consideration and attention to the particular risks and concerns faced by women and key populations due to coercive social and legal environments
- In some countries, collection and publication of phylogenetic data may require legislative or policy change
- Community participation and accountability for collection and use of data
- Legal redress in case of misuse of phylogenetic data.

Legal associated risks of misuse of phylogenetic data:

- The risk of self-identification – you need minimal information to be able to identify yourself even with anonymization. Once you have determined this you can find out information about those around you in the network, leading to attribution of blame for infections which may increase prosecution episodes.
- Protective laws on accessing public health data:
  - Guidance to mitigate risks of potential misuse by governments and police to target vulnerable populations. Currently, these protection of privacy is very limited, but successful examples should be highlighted for good practice.
  - What data are subject to subpoena, and how this risk can be reduced?
- Phylogenetic experts need to be consistent in their statements that source attribution cannot be definitively determined from phylogenetics alone (at present, but likely to change in the future)

Globally, 72 countries (a third of them in Africa) have laws specifically allowing for HIV criminalisation[423]. Box 9.5 reviews phylogenetic analysis used in criminal convictions. Government officials or other actors may misinterpret, or wilfully misconstrue, the results of phylogenetic research in support of political agendas or criminal convictions, putting individuals at risk of criminal prosecution for HIV transmission or broader human rights abuses. A realisation from these communities of the possible consequences- for privacy and prosecution – as well as a desire to know who infected whom to assign guilt or blame for passing on or acquiring HIV – may lead to a reluctance to test, failure to disclose contacts and/or refusal of resistance testing. There is evidence that these effects have already occurred[424]. The likelihood of misuse and abuse of these data is high, particularly for stigmatised populations. However, researchers are often unprepared for the misuse of

their findings, and may be reluctant to alert ethical review committees and suspend research when risks to study participants increase, or to engage in forceful public advocacy.

Researchers may also be targeted for working with criminalized populations, as in Senegal where outreach workers targeting MSM communities have been detained, subject to laws such as proposed in Uganda which criminalized the failure to report individuals suspected of engaging in homosexual behaviours, or denied the opportunity to even apply for local research ethics committee (REC) approval because of the view that homosexuality was inimical to local "values"[425]. These are ethical challenges facing the global research community, not only those working on HIV phylogenetic studies. Box 9.6 outlines some key social and legal considerations to mitigate risks in phylogenetic studies.

Box 9.5: Use of Phylogenetic Analysis in Criminal Convictions

- Since the infamous "Florida dentist case" in the beginning of the 90s, phylogenetic analyses started to be used in court cases as a forensic tool in HIV transmission investigations, e.g., cases where one or more complainants allege that a defendant has unlawfully infected them with HIV[21].
- To date phylogenetics has been used in HIV-related court cases in approximately 10 countries[406,426,427].
- Cases can be criminal (in countries where transmission of HIV infection is specifically criminalized) or civil (in the context of general civil laws e.g., by applying physical or sexual assault laws to HIV-related cases).
- Uses in court cases include[405]:
  - Indicating the timing of the most recent common ancestor, which can contribute to understanding the timing of infection.
  - Direction of transmission
- Most HIV-specific laws are broad and vague, and as such do not require proof of transmission and prosecution can be based on potential exposure with non-disclosure.
- Phylogenetic evidence cannot stand alone in court – it should be used in the context of other evidence, such as full epidemiological investigation and contact tracing[406,426,427].
- Experts have worked with the Crown Prosecution Service in the UK to produce guidance and highlight the limitations of phylogenetics in prosecution cases including:
  - Phylogenetic information alone cannot prove transmission beyond reasonable doubt - an indirect link can never be ruled out. In contrast, significantly separated clustering can be used as evidence against direct transmission, provided the samples have been drawn close enough to the timing of transmission and do not get phylogenetically separated by onward transmission events.
  - Challenges in communicating results to non- experts – particularly the lack of certainty.
  - Identification of a source of a transmission would require two major assumptions: all strains of all patients ever infected with HIV are available as "controls", and a phylogenetic tree can flawlessly reconstruct a true epidemic history. Both assumptions are unrealistic.
  - The use of phylogenetics in public health settings is of growing concern, since it provides a powerful tool to track sources of infection and target treatment strategies, but can also be used to prosecute source cases in court. This conflict between individual and public health required careful consideration with a balanced legal view.

Box 9.6: Important social and legal considerations

- Are there particular approaches to handling reporting of results that will reduce risk?
- What role can local and external ethics committees play in addressing risks and handling the potential for political issues that arise in the local context?
- What kinds of discussions with policy makers, government officials or other stakeholders might be helpful in planning the research and communicating findings?
- What on-going monitoring will be conducted to ensure respect for study participants and impact on people living with HIV or key populations? What resources are available for advocacy and redress if concerns arise?
- Have groups of people living with HIV and key populations been meaningfully consulted? Are their views and concerns taken into account?

Misuse of phylogenetic data including seizing and subpoena of such data by police and in criminal proceeding or perceptions by people living with HIV (PLWHIV) and members of key populations that phylogenetic data might be misused against them can undermine trust in research project and in health care systems, thus risking to undermine HIV prevention and treatment programmes. Research conducted in countries where privileged information between medical practitioners and their patients could be seized in HIV-related criminal trials show that people living with HIV were more reluctant to speak openly with their practitioners about their sexual partners and practices[424].

iv) Risk mitigation strategies

Many of the risks relating to identification of individuals from phylogenetic information in environments with oppressive laws and policies can be reduced through use of anonymization. Therefore, one default presumption is that if scientific objectives can well be accomplished with anonymized data, this is preferable. This default also presumes there is no overriding interest in individuals receiving research results at the individual level. If the data are not relevant for clinical care, given, for example, significant time delay between sample collection and generation of sequence information, then there is little rationale for needing to provide data back to clinicians for treatment. There may also be cases in which sequence analysis is timely and useful, but could be provided back to clinics before phylogenetic analysis is complete, thereby allowing time to anonymize resulting sequence data before phylogenetic results are in.

If anonymization is significantly detrimental to the scientific objectives, further ethical analysis must be undertaken and specific steps to protect the data from use in harmful proceedings. These steps might be technical (storage linkage to identifiers in coded,

separate databases with controlled access) and/or legal, such as legal agreements that data will remain protected from disclosure for the duration of the study.
Furthermore, while anonymization will help address individual risks, it will not address risks to groups. Groups can be placed at risk through characterization in the research as high risk or likely to transmit virus, and these can include geographically defined groups, as well as sexual or gender minorities, those defined by ethnicity, nationality, or migration status. Mitigation plans to address these risks need to include consultation with representatives of these communities, consideration of the public health value of the findings and developing plans to communicate in formats and venues that are the least damaging to the parties that may be vulnerable. In some cases, more detailed findings might need to be communicated confidentially rather than publicly; and some group descriptors may need to be masked in publications and press releases about the research. Risk mitigation strategies must also provide for redress mechanisms in cases of abuse or misuse of phylogenetic data. These may require the establishment of ties with local legal services organizations working to protect people living with HIV and criminalized or stigmatized populations to ensure that they have access to the means to protect their rights.

A further risk mitigation strategy relates to training personnel and actors involved in phylogenetic research on the potential of harm to communities and individuals. Such training should be aimed at ensuring that research staff are sensitive to the risk of harm and understand key issues of anonymity, confidentiality, informed consent and protection of research participants and communities.

v) Valid informed consent and other safeguards

The formal requirements for the achievement of valid consent are well-established in the literature with defined guidelines[428]. Such consent should be informed, voluntary and competently given. However, whilst the formal statement of such requirements is relatively straightforward, its achievement in practice is one of the most complex aspects of good medical research practice. This is true across all settings, but particularly in LMIC. For example, it is not uncommon for participants to consent based on trust rather than a thorough understanding of the research proposed.

The issues arising in relation to consent for phylogenetic studies are likely to be multifaceted and challenging. Obtaining community assent (via community leaders) and individual informed consent is particularly challenging for complex scientific studies such as phylogenetic research, which involve concepts that are extremely hard to both explain and understand, as well as multiple possible risks and benefits. Furthermore, in almost all phylogenetic studies there will be unsampled individuals acting as either a common source of infection or as an intermediary in a transmission chain who have not consented to the participation in the study.

Due to the complex concepts involved in phylogenetic research, it may raise fears about the aims of the work and the implications of participation among research participants, frontline research staff, healthcare professionals and ethics committee members. Models that increase the understanding of phylogenetic studies, therefore, need to be designed and shared. These must emphasise the potential harms, thoughtful mitigation of harms to risk groups, processes for monitoring risk, and clear protection procedures to minimise risks. Nevertheless, with ever-advancing technologies, a comprehensive consent model will be hard to design.

Study participants and patients whose samples are being used for phylogenetic analysis should ideally have consented to such use. However, sequence data generated from drug resistance testing and other surveillance data typically does not include explicit consent to participate in large scale phylogenetics analyses and data from previous research studies often entails broad consent for HIV-related research, but rarely involves specific consent for phylogenetics. In such situations, a waiver of specific consent may be obtainable from an ethics committee. Waivers of specific consent are allowable when samples are no longer linked to identifiers, or where consent was given for sample collection for research and storage in future studies, without specific consent for the current research.

Independent review of protocols for phylogenetic studies are also essential for the protection of research participants. Social and legal considerations highlighted in Box 9.6 underline the importance of promoting responsible conduct in research while providing assurances to participants and the wider public that their welfare is fully considered as data generated from their participation contributes to knowledge and development. The role of local ethics committees is essential for providing local, independent, representation

for research participants and others affected by the research, as well as ensuring that the local context in which researchers and participants are situated is taken into account. This is particularly important when research involves people from vulnerable groups, who could be exploited or coerced to take part in research. It is, therefore, essential that ethics committees understand, define and demonstrate their role as protection of human welfare, and not as an institution that serves to reinforce a political position of a state[425]. Members of local ethics committees can also provide oversight in safe-guarding the interests of people who, due to poverty or the expectation of medical research providing access to health care, may agree to participate because of the perceived benefits, without being fully cognisant of the nature of the research and the use to be made of biological samples and other data they may provide[429].

vi) Community engagement

Researchers and health care professionals must understand the values, perspectives and concerns of prospective research participants, and those most likely to be affected by research discoveries. Researchers should engage with community members to understand the social and cultural factors that define communities and make them vulnerable, as well as the nature and evolution of their concerns. Community engagement should, therefore, occur early on in the research design process, ensuring that the research is relevant to participating communities and local perspectives are included in the design and overall conduct of the research studies[430,431]. Meaningful community engagement is particularly challenging in research naïve and low -income communities. These communities often lack authentic representative structures, have low literacy levels and experience high poverty[432,433]. Such conditions place these communities at risk of being exploited, especially when research with highly technical elements, such as viral genomics, is introduced. Therefore, effective community engagement for these studies is essential to understand the perceptions of the study, allow social and cultural contexts to be incorporated, and to maximise trust in the research team.

Community engagement with criminalized or socially marginalized populations in particular can be a challenge, as such groups may be “hidden” or fractured with divided or contested representation. As a result, research with migrants, prisoners, drug users, persons with disabilities and criminalized populations is often conducted without a

representative advocacy group. Community engagement can also be time-consuming and resource intensive.

The phylogenetics study team of the PopART study[434] in Zambia has performed extensive community engagement in those communities in which the study takes place. The process involved obtaining community input in the design stages, as well as ongoing consultation throughout the study, with the development of a feedback protocol. Box 9.7 highlights five key questions to consider in the development of responsible and ethical community engagement in phylogenetic research based on a case study of successful community engagement[435].

Box 9.7: Five Key questions for responsible and ethical community engagement in phylogenetic research

1. What is the best community engagement strategy for phylogenetic studies and how sustainable is it?
2. How to provide feedback to communities? Does informing entire communities add value and/or pose risks?
3. How valid is informed consent when a phylogenetic study is nested within an existing study or healthcare setting? Does it become a question of trust?
4. How can we avoid stigmatization when public health interventions are tailored towards specific communities?
5. How can researchers best share results at community level? Is it disrespectful not to do so?

vii) Communication and equitable data sharing

Phylogenetic inferences are based on probabilities and the results can be ambiguous. Understanding and communicating this uncertainty is key to understanding the phylogenetic results. Researchers performing phylogenetic analysis must ensure that these caveats are clearly highlighted in any dissemination, including interviews, publications, oral presentations and posters.

Mass media campaigns, as well as reporting on social media, television, radio, in newspapers and in the news has been shown to be a powerful tool to raise awareness about the disease, treatments, prevention, and public health campaigns aiming to change attitudes and behaviours. However, the way the media frames the illness and reports study outcomes can affect both the long-term and short-term success of any campaigns and may have unintentional consequences, including a lack of trust in healthcare services[436,437]. Any ambiguous or misreporting of the outcome of phylogenetic studies may result in a reduction in HIV testing rates, increased scepticism about participating in

studies, and risk groups being less likely to access healthcare. Therefore, it is essential to educate the media, the community, and local health personnel about these studies.

Care must be taken especially in reporting findings relevant to specific population subgroups, including identifiable geographic areas, population groups that may be stigmatized or targeted by government, police, others in the community, or subject to criminal charges. Researchers will need to consider the potential social harms and political impact of findings before deciding exactly what information should be publicly shared or published.

There is growing awareness that while making data more widely available may lead to the potential to unlock new avenues of research and maximise the health benefit from research investments, conversely, it also raises ethical and governance challenges. These challenges can be more complex in LMIC where inequities in resources and power imbalances may lead to the local researchers who collect the data and participants' communities being disadvantaged. There may also be heightened concerns about risks to vulnerable individuals and groups. There has been limited research in this field to address these concerns. Therefore, to bridge this gap the Wellcome Trust undertook a project to gain the views and experiences of research stakeholders in LMIC about ethical issues raised when sharing public health and epidemiology research data[397]. The aim was to develop effective and fair data-sharing policies that are locally appropriate and to provide a platform for pooling resources on knowledge about the ethics of data-sharing. The resultant report[397] highlighted four key elements of best practice in equitable data sharing: assessing value, minimising harm, promoting fairness and reciprocity, and embedding trust. Furthermore, several priorities for future research were identified: to improve consent, governance, data policy development and capacity building.

Largely as a result of the funders' requirements, many anonymised HIV sequences are being made publicy available via GenBank and LosAlamos. This is advantageous for some research studies, such as vaccine development. However, there is a real risk of a lack of awareness that every sequence is associated with a patient or study participant. Therefore, we recommend that only limited information (such as the year of sampling and the country of where the sample are collected) are routinely published with each sequence, and that any other anonymised information should be provided via a controlled

access protocol which ensures that the research use proposed is scientifically valid, does not pose any risks to study participants, and is in line with the informed consent obtained.

Phylogenetic research also often requires data sharing across multiple sites, including international research collaboration. Different patient information sheets and consent forms may allow for different levels of data sharing, and laws may differ as to how data may be re-used. Any phylogenetic researcher must abide to the levels of sharing outlined in the forms, even if this impacts on the quality of the research conducted.

9.3 Conclusions and recommendations:

In this chapter, I have presented seven key ethical issues likely to arise in phylogenetic research, as identified at an expert meeting. Given that a one-model approach to address these issues is impractical due to the broad range of variations in studies and contexts, this chapter highlights a collation of themes which are essential to consider in order to undertake phylogenetic studies in an ethically responsible manner. The critical themes are clustered into eight domains which are set out below, and these provide a framework through which to consider the issues. In addition, along with each issue I set out key features of good ethical practice in phylogenetic studies, which are the result of in-depth discussion with my co-workers.

Eight steps for ethically responsible implementation of phylogenetic analyses:

i) **A careful risk-benefit assessment** should be conducted prior to and during the design, conduct and reporting of phylogenetic analysis. Risk assessment should address risks to individuals and to groups that may be identified in the research.

ii) **Protection of the rights and interests of study participants:** Individuals who participate in studies as well as the social and geographic groups that may be identified in phylogenetic networks need to be protected. Clinically relevant results should be returned to the patient and/or care provider.

iii) **Social and legal context**: An awareness of the social environment, legal environment, human rights violations and other potential negative consequences is essential. This includes knowledge both of precedent criminal cases, and of when and how data are subject to subpoena. These challenges are context-specific and the legal, political and social environments are subject to change. Furthermore, considerations of gender-

specific risks, and concerns faced by women and key populations should be evaluated due to coercive social and legal environments.

iv) **Risk mitigation strategies** should address risks to individuals and groups, and should take into account the potential for anonymization and masking of individual and group identifiers as protective strategies, as well as accounting for the scientific needs of the project and its value in informing public health strategies. Training should be conducted for research staff regarding the risks as well as the importance of anonymity, confidentiality, informed consent and protection of research participants and communities.

v) **Monitoring and redress mechanisms** should be established to accompany and respond to misuse of phylogenetic data, including collaboration with organisations providing legal assistance to people living with HIV and key populations.

vi) **Informed consent and other safeguards**: Study participants and patients whose samples are being used for phylogenetic analysis should, in most cases, have consented to such use. In the absence of such consent, waivers of consent must have been obtained from the appropriate ethics committees. Researchers must ensure that specific populations are protected against non-stigmatisation and non-discrimination.

vii) **Community engagement:** The engagement process should be started during the research design process, thereby ensuring that the research is relevant to participating communities, and that local perspectives are included in the design and overall conduct of the research studies, including risk assessment, risk mitigation, informed consent, and communication.

viii) **Communication**: Phylogenetic results require expertise to produce and to interpret. The results are often ambiguous, meaning that it is crucial the uncertainty associated with these methods is communicated appropriately during dissemination to the wider scientific community, government bodies, media, and participating communities. Specific efforts are needed to educate public health officials, the police and communities regarding the use of phylogenetic analysis in the context of public health, including its benefits and limitations. In addition, mechanisms must be established to ensure effective communication to participants of relevant information and risks to ensure their participation is consistent with ethical approvals. This should include women and key populations.

ix) **Data governance: oversight must ensure responsible and ethical collection and use of data.** The collection and publication of data should be non-stigmatising and non-

discriminatory, and accountability for use of the data should be ensured. A governance plan should be created and must address: confidentiality and safety of phylogenetic data along with any related information; the need to prevent any un-authorised use of data/information; and protocols for data/information sharing.

Phylogenetic analysis, either alone or in combination with linked epidemiological data, is a powerful tool with the potential to help reduce the spread of the HIV epidemic. However, an effective and sustainable model of good ethical practice in phylogenetic research is required to help minimise the risks to individuals/groups participating in studies, while optimising the scientific benefits.

Acute outbreaks and epidemics, such as the West African Ebola epidemic of 2013-16, are more time-sensitive and may occur in the context of less well organised response structures. This presents challenges to those implementing ethical frameworks, as there may be additional pressure to reach a decision, and it is easier for appropriate actions to be overlooked. Notwithstanding these additional challenges, ethics committees and researchers should take into account the specific ethical considerations presented by phylogenetic studies, to the best of their ability. It is hoped that, by creating the structure outlined above, these considerations will become an established part of ethics approvals, meaning that they should automatically be enacted in future acute epidemics.

Any researcher conducting phylogenetic analysis should be aware of the risks this type of analysis poses, and should take steps to mitigate them. This is particularly pertinent in LMIC, which often has weak governance and legal structures to protect vulnerable populations. Furthermore, these issues are likely to become increasingly problematic as sequence costs decrease and data becomes more routinely available in LMIC settings. Although these approaches are unlikely to become the norm in such settings in the immediate future, looking to the future, real-time phylogenetics may be used to direct public health responses and form the basis for surveillance programmes (as it is in more developed settings). Whatever the scenario, the fundamental principle of protecting participating individuals and groups must be central to study design and implementation, and to reporting of results.

Acknowledgements

I wrote this chapter based on the discussions for the Ethics of HIV Phylogenetics meeting, which I co-organised. The work presented in this chapter has been submitted for publication. It includes comments received from co-authors (below) prior to submission of the manuscript. This chapter is an amended version of the submitted manuscript. I would like to thank Shufang Wei for producing the infographics for this chapter (Figure 9.2).

CHAPTER 10

DISCUSSION

In this final chapter, I summarise the background and rationale for this thesis, in addition to the major conclusions. My discussion draws on themes encountered throughout the thesis and I combine my findings with evidence from other related research. I set out the strengths, weakness and limitations of combining molecular studies with epidemiological analysis to enhance understanding of transmission dynamics. I make recommendations regarding how these approaches should be incorporated into the analysis of future outbreaks, and discuss the implications of this work for public health policy. Finally, I set out areas for future research which have been suggested by unanswered questions arising from my work.

10.1 BACKGROUND AND RATIONALE

Understanding the characteristics of infectious disease outbreaks is critical in developing effective public health interventions. Traditionally, this has been done using well-established empirical epidemiological methods, such as surveillance, case finding, and contact tracing. These methods document the spread of infections, their clinical manifestations, and exposure patterns in 'time, place and person'. However, they are often time-consuming, costly and logistically complex, particularly in an outbreak setting. In addition, they may be limited by case ascertainment and recall bias, resulting in a variable ability to yield a complete or accurate representation of the underlying transmission patterns or extent of the outbreak.

With recent technological advances, pathogen genome sequence data are increasingly available and offer the potential to add a new and powerful tool to supplement traditional outbreak investigation. Phylogenetic analysis uses viral sequence data to track the evolution of pathogen mutations over time, allowing reconstruction of the historical ancestry between sequences and inference of transmission events. However, phylogenetic analysis is often undertaken in isolation, without the integration of epidemiological data. Although in isolation it is able to track the origin and evolution of an epidemic,

phylogenetic analysis requires metadata to enable an understanding of outbreak characteristics, such as transmission routes or associated clinical outcomes, which are essential for control efforts.

My primary goal, as set out in Chapter 1, was to assess the utility of combining phylogenetic analysis with traditional epidemiological analysis[15] in understanding the transmission dynamics of infectious disease outbreaks. I hypothesised that the combination of these two techniques would yield incremental insight over and above what each could provide in isolation. In order to do this, I explored two contrasting infectious diseases: HIV as an example of a chronic, generalised infection; and EVD as an example of an acute, localised infection. I used these examples to draw general conclusions about the use of a combined approach in studying infectious disease outbreaks.

10.2 Summary of findings

In summary, my work provides evidence that the integration of molecular and epidemiological methods can provide a powerful tool to enhance the understanding of transmission dynamics in infectious disease outbreaks. This work has generated new knowledge about the methods and utility of such an approach. My analysis has generated two novel insights:

1. The identification of a new outbreak of HIV, previously undetected through epidemiological and surveillance techniques (Chapter 7), which, to my knowledge, is the first example of molecular methods achieving this; and
2. The first example of integrated phylogenetic and epidemiological analysis identifying a novel mechanism of transmission, namely the infection with *Ebolavirus* of children in Village X, as set out in Chapter 6.

The combined methodology is likely to be of benefit to public health by providing additional information on which to base decisions about how, when, and to whom interventions should be targeted. My main findings are discussed in detail below.

[15] Definitions used: In the context of this thesis traditional epidemiology refers to empirical 'muddy boots' field epidemiology for the Ebola work and classical demographic surveillance data for the HIV work. Molecular epidemiology refers exclusively to viral sequence data and resulting phylogenetic tree analysis (Chapter 2).

The key lessons learnt regarding the future integration of these combined methods include:

- Phylogenetic analysis is an important complementary tool to use alongside epidemiological analysis, but it is currently under-used;
- Where possible, epidemiological and molecular data should be integrated to identify cases that may be linked;
- Clusters identified by molecular analysis should be prioritised for further investigation, with related epidemiological data being used to enable true clusters to be defined; and
- The maximal benefit is likely to be seen in the real-time application of these combined methods.

My work on a long-standing HIV epidemic in KwaZulu Natal, South Africa, has been used to demonstrate how these methods can enhance:

- Understanding of the origins of outbreaks and drivers of ongoing outbreaks/infections;
- Understanding of the relationship of the current outbreak to previous/ other concomitant outbreaks;
- Understanding of the patterns of viral mixing within and between outbreaks;
- Identification of unknown outbreaks/clusters of infection – to my knowledge, this is the first time molecular methods have detected a previously unknown outbreak;
- Knowledge of fine-scale transmission events between individuals, including patterns of mixing within and between outbreaks, and correlates of infection.

My work on the 2013-2016 Ebola outbreak in Sierra Leone has demonstrated how these methods can be used to improve:

- Understanding of the origin of new cases;
- Determining the number of introductions into a population and the trajectory of the outbreak, e.g. number of generations and hubs of infection (super-spreaders);
- Identifying a previously unidentified novel mechanism of transmission.

Although this combined approach offers incremental insight over and above the sum of the two individual methods, it is important to note that the prevailing role of traditional epidemiology remains crucial. Molecular epidemiology should be seen as complementary to traditional epidemiological analysis i.e. as an additional tool to supplement the core approach.

In terms of the practical application of these tools, 'complete' datasets incorporating good quality epidemiological and sequence data rarely, if ever, exist in an outbreak setting. Therefore, in practice, public health interventions are developed and implemented on the basis of incomplete data. Likewise, the analysis in my thesis has been carried out using routinely collected data from disparate studies, including data with many inconsistencies and limitations. Despite this, the analysis has yielded valuable results, and one of the strengths of this work has been to demonstrate that novel methods can improve understanding despite these real world complexities.

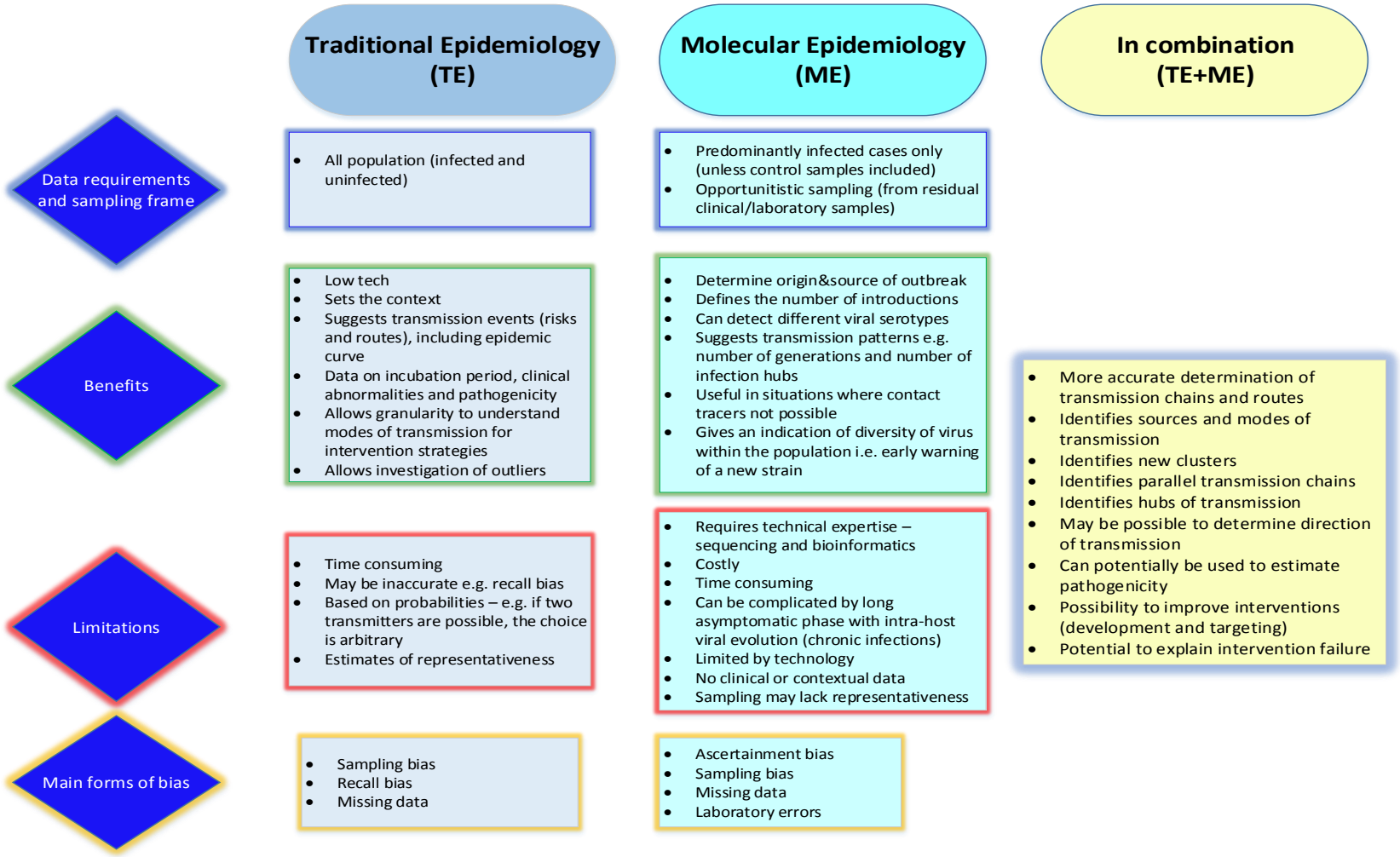
The strengths and limitations of this work are summarised below (section 10.4), together with detailed specific examples of novel insights provided by this research (section 10.3). Figure 10.1 summarises the utility of individual and combined approaches to transmission event reconstruction.

10.3 Key findings: Specific examples of the benefits of a combined approach to investigating transmission dynamics, together with the public health implications of these key findings

1. **Improved ability to determine the origin(s) of an outbreak and the relationship either to previous/other outbreaks, or to other clusters within the current outbreak**

Ebola: My analysis in Chapter 6 showed that phylogenetic techniques can help to confirm the origin(s) of outbreaks. Specifically, it showed that the Village X outbreak originated from the Aberdeen fishing village in Freetown. Other phylogenetic studies concerning the West African outbreak have confirmed the benefit of these techniques. For example, phylogenetics was used to demonstrate that the concomitant outbreak of EVD in the Democratic Republic of Congo in 2014 was not related to the West African Ebola outbreak (Chapter 4). From a public health perspective, understanding the origin of introduction events could influence control measures. In the example concerning DRC, had the introduction been from a neighbouring country, it would have supported the implementation of border control measures. Given that it was confirmed as a random emergence event, no such measures were necessary, meaning that human and financial resources that otherwise might have been deployed were conserved.

Figure 10.1 Summary of the utility of traditional and molecular epidemiological methods both individually and combined to determining transmission event reconstruction



2. Improved ability to determine factors that drive outbreaks, by understanding patterns of mixing within and between outbreaks

HIV: Chapter 8 set out my analysis which demonstrated the complementary and synergistic value of epidemiological and novel phylogenetic studies. Both methods elucidated and quantified the role of migration in increasing the risk of HIV acquisition. However, the molecular analysis threw new light on to the role and patterns of external migration, which led to repeated introductions of the infection into the Africa Centre Demographic Surveillance Area (AC DSA). Traditional risk factor epidemiological work had not been able definitively to confirm this. My work showed that infections are introduced from outside the DSA into high-prevalence areas, which subsequently seed infections in low-prevalence areas. These patterns are likely driven by circular labour migration and they help to drive and sustain the high incidence of HIV in the AC DSA. My phylogenetic analysis complemented the traditional epidemiological analysis, which had shown that both migration and prevalence of HIV in the area of residence were independent risks associated with the transmission of HIV (in the AC DSA). External migrants had the highest risk of HIV seroconversion, but this was confounded by age.

From a public health perspective, my work demonstrates that the AC DSA does not have a closed population. Therefore, interventions targeted solely at the AC population, in isolation, are unlikely to be successful. Instead, interventions (and infrastructures) need to be integrated with regional /national programmes, and they need to be flexible enough to provide education and treatment to mobile migrant populations. Furthermore, although the outbreak within the AC DSA appears to be fuelled by external migrants introducing infections into the area, the vast majority of new infections still occur in those who don't migrate (Chapter 8). From a public health perspective, it is important to recognise this, to ensure implementation of both population level interventions, as well as the targeting of risk groups.

3. Earlier identification of new emergent clusters

HIV: In Chapter 7, I detailed a new emergent cluster in the AC DSA that had not been identified by routine surveillance or epidemiological analysis, and which is likely still to be propagating new infections. Molecular epidemiology initially identified the cluster and the likely timeframes for cluster development (via R_0 calculations), and epidemiological data subsequently allowed me to identify the mechanisms driving the development of the

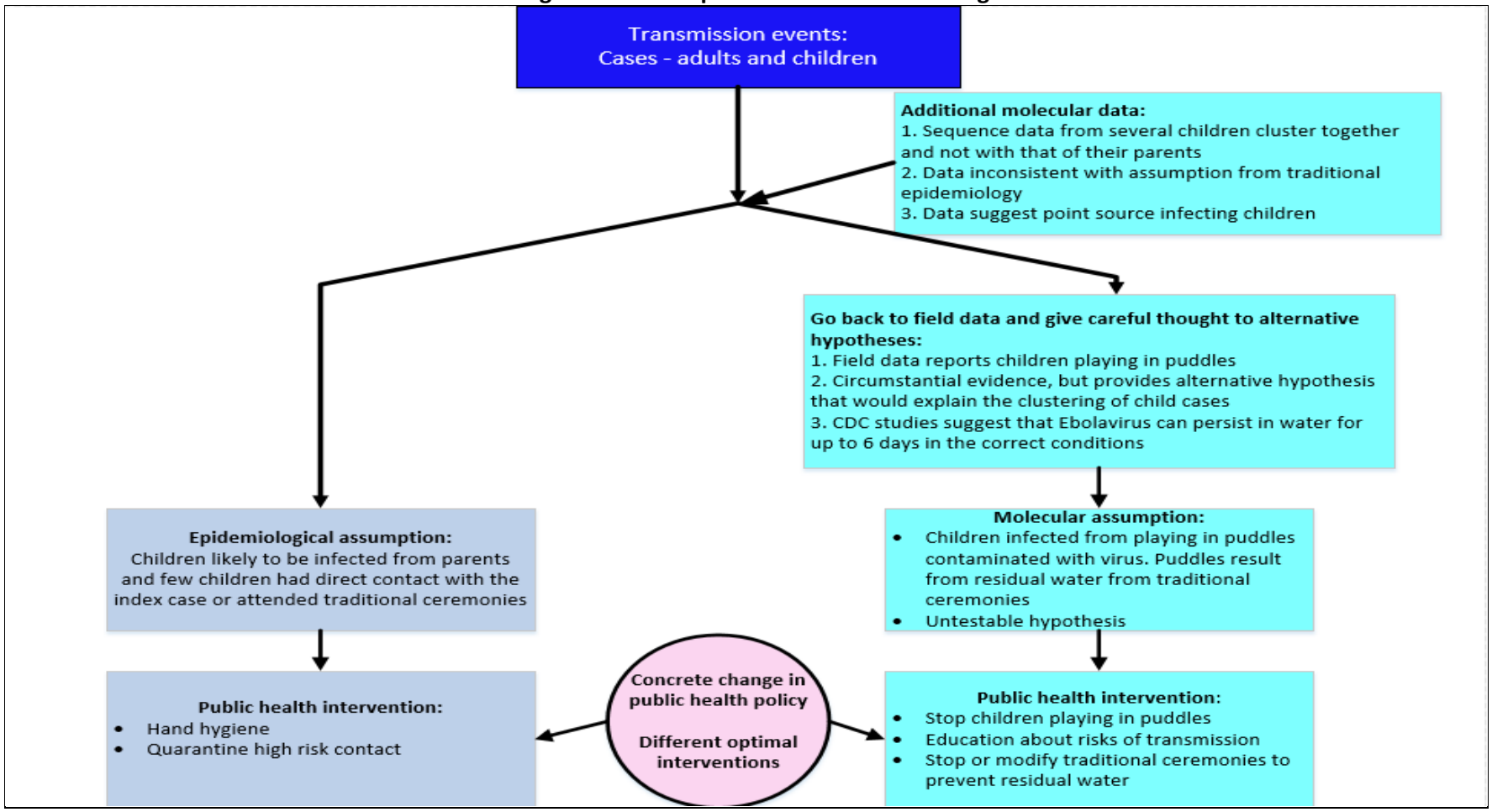
cluster. To my knowledge, this is one of the first examples of molecular methods alone identifying a new cluster in an existing outbreak.

Public health implications: Prompt identification of new clusters is crucial in allowing early interventions to be deployed in order to avert as many cases as possible. The combined approach yields a more detailed understanding of the fine scale dynamics of transmission, which in turn facilitates the formulation of interventions, for example targeted and combination interventions. Further efforts are needed to control outbreaks of HIV around new industrial developments, such as mines. Although this was already a known risk factor, my work adds weight to the need for new industrial/mining developments to be required to implement health and safety policies that address the risk of HIV acquisition, as well as improving education and rolling out awareness campaigns.

4. Identifying novel transmission mechanisms

Ebola: In Chapter 6, I demonstrated that combining the two methods has the potential either to help determine transmission events where there is uncertainty, or to confirm epidemic parameters. In addition, I explored the discrepancies in results generated by the two methods. For example, Figure 10.2 shows the process of exploring a discrepancy in results between epidemiological assumptions and molecular data in the Village X analysis. The molecular data were inconsistent with the epidemiological assumption of parent-to-child transmission due to the fact that the genetic sequences of samples taken from the children appeared to cluster and were similar to one another. This is highly unlikely to have been the case had the epidemiological assumption of parent-to-child transmission been correct. As a consequence of this realisation, I reviewed the epidemiological data and identified a potential novel transmission pathway: children playing in puddles created by the traditional cleansing/burial ceremonies. Although this hypothesis is untestable, it provides a plausible mechanism of transmission which best explains both the molecular and epidemiological data. Epidemiological data in isolation would have led to the wrong conclusion. It is interesting to note that it was the combination of molecular analysis and epidemiological data that allowed me to hypothesise the transmission route. Additional data from CDC, as set out in Chapter 6, was required to confirm the lifecycle of the *Ebolavirus* in water and thus to confirm the potential transmissibility via this route.

Figure 10.2: Diagram highlighting the different assumptions drawn from epidemiological and molecular data from Village X, leading to different optimal intervention strategies



In addition to my work, recent publications have also highlighted the utility of these combined methods in determining transmission routes. Research from the West African Ebola outbreak identified super-spreader events using combined methods, including one traditional burial ceremony which was linked to over 300 transmissions[244].

There are many public health benefits from identifying and clarifying transmission routes in outbreaks, including the following:

- Novel data allows confirmation of routes of transmission, thereby allowing appropriate implementation of interventions to prevent further transmission. Strength of evidence is important because it can influence both how and when interventions are deployed, particularly in resource limited settings. It also helps to inform approaches to future outbreak control.
- In the Village X case, providing education and public health guidance to prevent the build-up of standing water and/or children playing in puddles could have prevented cases. Due to the fact that interventions involving a behavioural change are difficult to implement, having stronger evidence to confirm the need for interventions gives conviction both to those determining intervention strategies, and to those affected. This should increase the probability of interventions being implemented successfully.
- Furthermore, identifying super-spreader events is crucial in targeting focused interventions. The prevention of a snowballing cascade of infections, particularly in early outbreak settings, can have a substantial impact on the number of cases. This was demonstrated in my work in Chapter 5, where I discussed the potential for publicly-available information to inform policy with respect to vaccination of healthcare workers in Ebola outbreaks. In many epidemics HCWs have been found to be the transmission link to both other HCWs and the general population, particularly in early outbreak settings. My work showed that prophylactic vaccination of HCWs could have a substantial impact on preventing early chains of transmission, before the risk of epidemic spread is recognised, and as a result, reduce the risk of future large epidemics.

5. Determining the number of introductions and parallel chains of transmission

Ebola: During a large and complex outbreak, particularly if multiple introductions or transmissions are involved, parallel transmission chains may develop and may overlap

significantly, both in time and space. Epidemiological analysis alone cannot distinguish between multiple strains co-circulating in a population. For example, transmission from different viral lineages can occur on the same day, but might later incorrectly be considered by epidemiological analysis to be directly linked. Molecular approaches can more accurately define these complex transmission dynamics and, therefore, are of benefit in these circumstances. In addition, molecular methods can unravel uncertainties in transmission that have already been identified by epidemiological analysis. An example of this is the application of these methods in the analysis of outbreaks of Dengue in which there were multiple strains circulating at the same time[330].

In addition, phylogenetic analysis of the Village X outbreak (Chapter 6) confirmed a single point source introduction with only one infecting strain (the Index Case). Although epidemiological data had led to this being hypothesised, the use of molecular methods was necessary for confirmation. This was important because the local area had recently experienced other cases of EVD and it was possible that the introduction to Village X originated from one of these locations. Had the source of the infection been different, it is possible that other public health measures might have been necessary. This type of analysis was also used during the West African Ebola outbreak to determine if new cases originated from unrecognised local transmissions, undetected intra-country transmission, imported cases from neighbouring countries, or new emergence events [106,160,247,285] (Chapter 4).

Public health implications: Understanding transmission dynamics helps to predict the trajectory of an outbreak; for example, timelines of infections may vary between a single point source outbreak and outbreaks involving multiple introductions. This can have a material effect on the choice of intervention strategy and on the evaluation of current interventions. A good example of this is that unrecognised local transmissions might suggest that contact tracing has been inadequate, leading to efforts to improve this intervention.

6. The need for and development of a consensus concerning the key features of good ethical practice in phylogenetic studies

My research has highlighted a number of key ethical issues arising from phylogenetic analysis, either on its own or in combination with linked epidemiological data. For

example, molecular data creates the possibility of identifying infected individuals and linked cases despite anonymised data. This could lead to stigma, discrimination, harassment and criminal convictions for individuals or vulnerable groups. This is particularly pertinent in low and middle-income countries (LMIC), which may have weak governance and legal systems and may fail to protect vulnerable populations. This problem is likely to be exacerbated by the decreasing costs of genetic analysis and the increasing availability of data. The principle of protecting participating individuals and groups must be central to future study design and implementation, and to the reporting of results. Therefore, a consensus regarding the key features of good ethical practice in phylogenetic studies, specifically for HIV, was developed and was presented in Chapter 9.

10.4 Strengths and weaknesses: Is either discipline sufficient on its own?

Notwithstanding the promising potential of molecular analysis, traditional epidemiological methods remain the cornerstone of outbreak investigation. Epidemiological data are required to determine key epidemic parameters such as the biological characteristics of a pathogen, the infectious period, the incubation period, the mode of transmission, and clinical manifestations associated with an outbreak. These parameters cannot be determined using molecular data alone, and are critical to optimising public health interventions.

However, epidemiological data have limitations: to allow the construction of robust transmission trees, data must be good quality, detailed and complete. In practice, this is rarely achieved, particularly in outbreak settings, given how challenging and time-consuming data acquisition is. Therefore, scenarios in which epidemiological data alone may be sufficient to determine key outbreak transmission dynamics are restricted to certain specific scenarios which are relatively uncommon; for example, simple, point source, isolated outbreaks that are easy to sample accurately.

Molecular studies are also of use on their own, particularly in detailing the origin and distribution of infections across a sampled population via genotyping of the pathogen and the construction of phylogenetic trees. There are also instances in which molecular data reveal information not available from epidemiological methods alone, e.g. co-circulating

strains and parallel transmission chains. However, molecular epidemiology is arguably more limited when used in isolation. Although it can confirm a transmission event, for example if sequences A and B have a common ancestor, unless the sequences are linked to epidemiological metadata, meaningful interpretations regarding the transmission route or the risks of transmission events cannot be made.

Moreover, traditional epidemiology is often required to identify the cases from which samples can be taken for molecular analysis. In this way molecular analysis can be entirely dependent upon samples from epidemiological studies, while the reverse is not the case. Thus, although the methods are somewhat interdependent, traditional epidemiology serves as a framework through which to interpret molecular analysis. This framework allows sufficient granularity to interpret sequence data and to explain any outliers. Without such a framework, the impact of molecular analysis on public health outcomes is limited.

Both methods have limitations, particularly with regard to the effect of unsampled populations and missing data. These are common problems and can result in both methods generating incorrect inferences, as outlined in Chapter 2. Due to the fact that molecular analysis is based on the observation and detection of tiny differences between samples (i.e. genetic mutations), any significant misclassification can alter the results substantially. This was evident in my analysis of the Village X Ebola outbreak (Chapter 6), in which the date of sample collection lacked consistency: some samples were taken at the time of diagnosis, while other were taken two weeks after diagnosis. This inconsistency altered the resultant transmission tree, as set out in my discussion in Chapter 6. Furthermore, pathogen evolution itself can be a limitation in these studies. For example, pathogen evolution can be a source of both type I errors (false positives when unrelated sequences are mistakenly clustered because they acquire the same, rare mutations independently e.g. drug resistant mutations in HIV) and type II errors (false negatives occur when related sequences evolve so rapidly that their relatedness is masked by background noise). Fully understanding these limitations and how to mitigate them requires further work.

Thus, while both methods are of value when used in isolation, my work has shown that, when combined, they offer additional insight over and above that offered by each

separately. Notwithstanding this, epidemiological analysis remains the cornerstone of public health responses to outbreaks, and the epidemiological data gathered offers a framework with which to use and interpret novel molecular techniques. Without such a framework, molecular analysis offers limited incremental insight.

10.5 The benefit of supplementing traditional epidemiology with molecular techniques in acute and chronic outbreaks

a) Acute outbreaks (epidemics)

Many viral infections cause relatively acute and localised outbreaks leading to recovery or death within weeks or months. An analytical approach combining traditional and molecular epidemiological techniques has shown the potential to benefit investigation of the transmission dynamics of acute viral outbreaks, such as Ebola. As shown by the example of the 554 *Ebolavirus* sequences from Sierra Leone discussed in Chapter 6, this combined approach is particularly useful for determining the parallel evolution of multiple viral lineages (although clinically not relevant), an application for which traditional epidemiology alone has limited value. This finding has been reinforced by recent research into Dengue outbreaks as mentioned above[330].

It is highly likely that a combined approach would be of benefit in other situations to understand acute infectious outbreaks. Probable examples include: the investigation of influenza to distinguish between seasonal strains and novel emerging strains, all of which cause similar symptoms; and settings in which contact tracers are not trained, or where information-gathering is challenging (e.g. language barriers or personal hazard due to contagious diseases), as it may be easier to obtain a clinical sample for molecular analysis. Table 10.1 summarises the potential application of molecular methods to acute outbreak investigation.

b) Chronic outbreaks (endemic)

Chronic viral infections can cause both localised and generalised outbreaks. The primary benefits shown by my work of a combined approach in the context of chronic outbreak settings, are in identifying emerging clusters of recent infections against the background of an endemic setting, and in highlighting vulnerable groups and transmission routes for the targeting of interventions. The long-term surveillance work at AC highlights the challenges

Table 10.1 Application of molecular methods to steps undertaken in acute outbreak investigations.

Steps in outbreak investigation	Application of molecular methods
1. Establish the existence of an outbreak	- Genotype pathogen - Visualise new cluster on phylogenetic tree - Identify and confirm clustering
2. Verify the diagnosis	- Molecular studies provide confirmation of diagnosis by definition as they use viral sequence data
3. Define case definition and identify cases	- Set criteria to define a 'new outbreak' - Use historical data to contextualise outbreak compared to previous cases - Use molecular studies to identify additional undiagnosed cases e.g. asymptomatic transmission - Set criteria for defining new cluster and expanding clusters
4. Describe the outbreak and develop hypotheses	- Identify origin of outbreak - Identify number of introductions – point source or multiple introductions - Identify parallel transmission chains - Identify modes of transmission (with metadata) - Describe cases in relation to known risk factors or potential sources - Identify high-risk groups, vulnerable groups, super-spreaders
5. Evaluate hypotheses and refine	- Both epidemiology and molecular data allow inference of various epidemic parameters. Molecular data often requires metadata to allow meaningful conclusion to be draw. The inferences should be compared to confirm hypotheses or allow exploration of alternative hypotheses where there are discrepancies
6. Implement control and prevention measures	- Earlier action - Development of optimal strategies e.g. targeting of interventions towards risk groups, identification of need for flexible or combination intervention, Identification of novel intervention strategies - Determining viral strains and new strains can impact vaccination development and strategies
7. Evaluate interventions and initiate surveillance	- To identify where interventions have been successful at preventing transmissions and where they have failed i.e. where are ongoing transmissions arising - Develop methods to undertake real-time surveillance to rapidly identify new clusters/outbreaks

in identifying emergent clusters by epidemiological methods alone. Although such clusters might eventually be identified in this way, my work has shown that phylogenetic techniques are faster and more effective at doing so with likely benefits for preventative interventions. Therefore, there is a clear case to supplement traditional epidemiological surveillance programmes with novel molecular analysis.

However, it is worth noting that phylogenetic analysis of chronic infections with a life-long phase (e.g. HIV, hepatitis, chronic Epstein-Barr Virus, Cytomegalovirus and Herpes Simplex Virus) can be more complicated. In these instances, many additional factors must be considered, including: the high proportion of asymptomatic, undiagnosed, and unsampled infections; lack of knowledge of the time of infection; viral evolution over time, with the development of multiple intra-host quasi-species; long latency periods; low viral load in

some cases as a result of treatment or elite control; and phylogenetic uncertainty, as a result of these variations. For example, the HIV viruses residing within a host often descend from a single transmitted/founder virus. The high mutation rate of HIV, coupled with long delays between infection and diagnosis, make isolating and characterizing the 'infecting strain' a challenge. The single consensus sequence at the time of sampling will be different to the sequence of the virus at the time of infection, limiting our ability to identify the transmission event accurately. Furthermore, the longer time interval between infection and sampling results in a more diverse intra-host virus, which in turn increases the inferred range of dates on which infection might have occurred, makes it harder to determine the most likely common ancestor. However, Next Generation Sequencing (discussed later in this chapter) may change this, as it allows sequencing of all quasi-species within an individual. Therefore, the methods used to reconstruct transmission trees must be adapted to different pathogens, outbreak settings, and sequencing technology.

10.6 Development of new models combining epidemiological and sequence data

Since I began this research in 2015, bioinformatics models have been developed which use statistical methods to combine one epidemiological parameter (e.g. time of sampling) with genetic sequence data (either raw sequence data or phylogenetic trees) to reconstruct networks of transmission events, such as Outbreaker[330], Scotti[438] and BEASTlier[439]. These models remain under development, and as the methods used become more sophisticated, they will require further validation. At present, published works using these models are limited to simulated research settings with the capability to analyse only small numbers. Current versions are relatively simple, and because they only include one epidemiological parameter, they do not incorporate the complexities of real data (e.g. non-sampled patients) or the uncertainties related to phylogenetic analyses, nor do they account for circumstances specific to different disease outbreaks. Furthermore, Outbreaker, for example, uses raw nucleotide sequence data exclusively, which only describes genetic distances between isolates, compared to ancestral relationships determined by phylogenetic trees. This misses the information provided by phylogenetic analyses, including tree topology, the incorporation of evolutionary models into the tree (outlined in Chapter 2), and the acknowledgement of potentially unsampled intermediates within the

tree[440]. As a result, the utility of these sequence-based models is restricted to densely sampled outbreaks with a low proportion of unsampled cases.

Despite the inevitable issues arising from the use of new models, my work suggests that valuable information can be obtained from them. Not only did my use of Outbreaker in analysing the Village X outbreak allow confirmation of the epidemic structure, number of generations, and transmission hubs which had been suggested by traditional epidemiological work, but it also implied novel transmission events that were inconsistent with the assumptions made by traditional epidemiological analysis, specifically that the sequences obtained from children differed from what one would have expected had transmission been parent-to-child. This led to careful consideration of other possible modes of transmission and, ultimately, to the insight that a novel mode of transmission was responsible. In their current forms, these models remain prone to generating erroneous conclusions, and further work is needed to refine and validate them. In the meantime, results need to be interpreted carefully.

During the course of my research, bioinformatics models have evolved rapidly. Indeed, one of the models I used, the outbreak module of BEAST2 (Chapter 7), was only released for public use earlier this year (2017). While the development and refinement of these models are not the focus of this thesis, they are worthy of mention and should be further explored once the techniques are more established. These models have the potential to advance the application of integrated methods for benefit of public health.

10.7 Limitations

The primary limitation of this study is that the 'true' underlying transmission chain is rarely known, meaning it is impossible to confirm which is the optimal method. I encountered several other types of limitations while undertaking this work, including scientific limitations, data related limitations, practical limitations and personal challenges, which I outline in this section.

10.7.1 Scientific limitations

The varied scientific limitations are outlined in the relevant sections throughout this thesis, and an overarching theme was the bias arising from the sampling strategies used. As molecular samples are often obtained opportunistically, for example from other studies,

from outbreak surveillance, or from those attending clinics/health centres, they incur the same sampling bias as the underlying (epidemiological) study.

As discussed in Chapter 2, molecular studies on their own have several significant limitations over and above those of epidemiological analysis. The specific limitations I encountered during the course of my work are set out below:

a) Ascertainment bias: Although both traditional and molecular epidemiological studies can be limited by ascertainment bias (discussed in Chapter 2), this may have a greater impact on molecular analysis. Viral molecular studies using field data suffer due to the fact that they may only be able successfully to sequence samples from individuals with high viral loads. This biases the results either towards new infections, or towards those not on treatment. The type of sample taken is important in this regard: dried blood spots permit only a small sample volume and so require a higher concentration of the virus in the sample; other samples such as blood or saliva contain more total RNA or DNA and so allow a sequence to be obtained with a much lower viral load.

b) Inability to include uninfected cases: Molecular analysis of an infectious agent is, by definition, not possible from non-infected individuals. This means that it is not possible to make comparisons with those who are not infected with respect to risks associated with disease transmission. However, control groups can be obtained, for example, from samples of historical outbreaks, or from those with a different strain or with another infection. This may provide important information to determine the likely origins and subsequent divergence of outbreaks. However, classical epidemiology is necessary to facilitate the comparison of infected and uninfected populations within an outbreak in order to determine transmission routes and risks of transmission.

c) Imperfect inference of timing of infection and inability to determine direction of transmission: Although molecular studies allow inference of epidemiological linkage between individuals, it is difficult to infer who infects whom. However, next generation sequencing does enable directionality to be inferred (see section 10.9 below).

d) Cost and timing: Sequencing is not cheap and requires expertise, expensive technology and bioinformatics platforms. Therefore, it is hard to implement large-

scale sequencing programmes in resource-limited countries. However, with advances in technology, the costs are decreasing and there are already instances in which sequencing studies may be more cost-effective than large, time-consuming epidemiological studies. This is particularly true when real-time sequencing is available, which will enable generation of sequences when they are most needed i.e. at the beginning of an outbreak.

e) Technical expertise is required: With improving technology, the expertise required to undertake sample processing, sequencing, and phylogenetic tree generation may decrease over time. However, the interpretation and analysis of phylogenetic data, in combination with other metadata to contextualise them, still requires a highly skilled technical workforce, and this is unlikely to change. Phylogenetic analysis requires many technical assumptions to be made at each step of the analysis, for example which reference genome to use, and understanding how to interpret and communicate the uncertainty from maximum likelihood results also requires expertise. This is essential to ensure accurate representation of the conclusions drawn, as discussed in Chapter 9.

f) Lack of consistency in standard reporting of analytical methods for phylogenetic studies: The field of HIV phylogenetics is advanced compared to other pathogens, and there are standard methods and criteria for defining clusters. Unfortunately, this is not the case in less established fields such as the phylogenetic analysis of *Ebolavirus*. Thorough and transparent reporting is the norm in the field of HIV phylogenetics, meaning parameters, such as the thresholds used to define clusters, are readily available. Unfortunately, due to the fact that *Ebolavirus* is better conserved than HIV, it is not possible to extrapolate from the parameters used in the latter case to the benefit of the former. This presents a challenge to *Ebolavirus* phylogenetics, as detailed methodological reporting in published papers is vague, and many studies do not define their criteria for reporting clusters. There is a need for future research to provide clear guidance on how to reproduce the methods used therein.

10.7.2 Data limitations

This thesis uses routinely collected data from real life outbreak settings (both emergency and chronic) in low-income countries. Such data are inherently challenging to work with,

regardless of the origin, and have multiple intrinsic limitations. I have undertaken a series of investigations using a wide range of disparate datasets that are individually incomplete and do not provide the power accurately to assess complete transmission dynamics. However, in combination, and in reviewing the totality of evidence, the conclusions drawn have both moved this field forwards, and also provided benefits to advance public health measures.

Routinely collected data provide a wealth of information that can be used to answer highly relevant questions at low cost. However, there is an inefficiency in accessing and processing routinely collected data across all settings; for example, appropriately archiving data (labelling and storing them) so that they can be reused, while maintaining governance, security and confidentiality. Overcoming these limitations, to allow timely access to pre-existing data during an outbreak, could benefit public health responses in a cost-effective manner. For example, during the 2013-2016 West African Ebola outbreak, expediting the identification of the magnitude and modes of Ebola transmission could have led to a faster and more effective public health response.

The type and depth of analyses carried out are highly dependent on the datasets available: on the information they contain, the quality of these data, and the amount of 'missingness'. There were several data-related problems during my work, which limited the analyses I was able to undertake. These included difficulties in gaining access to data despite the appropriate approvals; data loss by collaborators; misunderstandings with respect to data ownership at an individual versus institutional level; and technical issues with mislabelled data. Discussion of these issues is presented below, along with a summary of the key challenges and lessons learnt. These issues merit discussion in terms of the impact they had on the results obtained.

a) Data access and sharing

The recent West African Ebola outbreak showed the potential for data to be used to help shape and improve the public health response in international health emergencies[441]. Previous emerging infection outbreaks, such as the 2003 Severe Acute Respiratory Syndrome (SARS) epidemic, have also demonstrated the power of an international response based on information and evidence obtained in real-time[442]. These data can prove vital, particularly in the early stages of a fast-moving emergency, in which effective

public health planning can have a substantial impact on the overall trajectory of the outbreak[168,441]. During the recent Ebola outbreak, in many instances, the data and insights from the academic community were crucial in planning the short and medium-term strategies. Examples include: the scale-up of Ebola treatment centres[119], anthropological insights which promoted safe burial procedures[443], and transmission data informing vaccine policies (as discussed in Chapters 4 and 5)[444].

However, there is no best-practice standard for data sharing during outbreaks. During the Ebola outbreak there were many instances when individuals and organisations were unwilling to share data in real-time, with potentially serious effects on the success of the response, particularly in the early stages[441]. Reports suggest that the main reasons given for not sharing data were the perceived disincentives to sharing data (e.g. it would jeopardise subsequent publication and allow others to publish the data) as well as the confidentiality issues associated with public release[441]. However, in an unfolding international health emergency, the unwillingness to share data can impact the success of the response and threaten lives[445]. Box 10.1 summarises some specific challenges that I encountered during this work with respect to data access and sharing.

During the Ebola and subsequent Zika virus outbreaks, there have been numerous calls from leading scientific organisations and funding bodies (including the Wellcome Trust[397,446] and the Bill and Melinda Gates Foundation[447]), health agencies (WHO[448], UK DOH[441]) and individuals, to encourage data release and sharing, particularly during health emergencies. Indeed the Wellcome Trust convened a meeting with leading international stakeholders, which "*affirmed that timely and transparent pre-publication sharing of data and results during public health emergencies must become the global norm*"[446]. In order to facilitate this, a standardised framework, together with an easily accessible and readily useable standardised platform for data sharing, is required. This must be user-friendly, and should require minimal computing capacity, in order to be widely accessible, including to the rural and remote environments in which many of these emerging outbreaks occur. Platforms should not only collect data for scientific sharing, but should also be optimised for public health use in early epidemics. Furthermore, calls have been made for journals to help address the disincentives for sharing data, i.e. to provide incentives by only publishing data-driven research from public health emergencies if the data have been shared with authorities and legitimate bodies responding to the emergency at the earliest possible

opportunity[441]. Many journals heeded the call and signed a joint statement to this effect[449].

Henceforth, the public health and scientific communities as a whole, need to address this issue, promoted by leadership from major journals and international collaborating bodies, who often do not share their data widely and limit access to a select few. With good practice frameworks, the aim would be to create data sharing practices that enabled the development of evidence-based control measures in future outbreaks.

Box 10.1: Examples of data access and sharing issues experienced during this work with potential solutions

1. In the case of Ebola, data sources were fragmented - data were collected at the district level without any single centralised data structure. Much of the information was recorded on paper-based forms with no time, resources or funding to convert this to electronic data. While a significant amount of time and effort had been put into training data collectors and into collecting these data, the information was unusable in the form in which it was collected. Such data are not included in my work.
   Potential Solution: Prioritisation in future outbreaks should be to streamline data collection and invest in further training both to avoid the collection of redundant data and to maximise the development of datasets that can benefit public health response.

2. All data were nationally owned and required both national and ethical approvals in order to gain access. We obtained approvals from the Sierra Leone Chief Medical Officer and the Sierra Leone ethics board (in addition to LSHTM ethics board) as part of the collaborative National Ebola Data Archive project. Due to the decentralisation of all the data, accessing datasets required approval from different people at different locations with varying protocols and policies across sites. The data were either very easily accessed with no real confirmation of approval, or stringently protected, preventing even authorised access. Understanding and adhering to these local protocols was extremely time-consuming and, in some circumstances obtaining the extra authorisation required was unachievable(e.g. from district tier staff).
   Potential Solution: Internationally agreed, widely disseminated recommendations regarding data access protocols and ownership/stewardship for outbreak datasets during emergencies are needed to minimise this barrier. Furthermore, a national data guardian with responsibility for data stewardship should be identified in all future emergencies. Rapid approval for data access to legitimate collaborators who could produce valuable evidence on which to base public health responses is key in early outbreak settings.

3. There are several datasets that I know to be in existence which could have complemented the data presented in this work by enhancing the linkage coverage. However, the approval process for large international organisations takes from months to years. Due to the time limitations, I was unable to access some data sources for inclusion in this study. Furthermore, data were unable to be released until they were 'cleaned'. For both of these reasons, as of September 2017, WHO and CDC Ebola data sets were not available for use (beyond what was obtained as part of the National Ebola Archive project in-country). This is a considerable delay from the end of the epidemic (in June 2016 all three countries were declared 'Ebola-free').
   Potential Solution: WHO and CDC are both large organisations with capacity and resources to lead the way in forming standardised data collection platforms (as discussed above). WHO is the global leader in international emergency outbreak response, and it should be leading the way with respect to data sharing.

b) Data completeness

Dataset completeness was compromised in two main ways:

i) Incomplete data collection: This is present in all studies to varying degrees, but it is particularly striking in some datasets used in this work. This could result in missing/not completed variables biasing results (discussed further in Chapter 2). Where there were substantial missing data, this has been highlighted in the analysis. It is possible to adjust for missing data by undertaking multiple imputation, but this is only reliable where missing data are missing for no more than 20% of instances for a given variable. There are many variables where the proportion of missing data are ≥50%. Therefore, it was not appropriate to use multiple imputation methods.

ii) Lack of access to datasets I had initially anticipated including: This resulted in poor sample coverage and exclusion of data due to poor linkage of sequence data and metadata. The 'missingness' of data limited the power of my analyses and, in specific circumstances, prevented analysis being undertaken.

c) Data integration

A large component of this project was cleaning, formatting and merging different datasets. During this process, I identified several problems with the data:

i) Lack of consistency with patient identifier: In the Ebola datasets the unique patient identifier, used between and across datasets, were not consistent in identifying the same patient, and there were numerous examples of the same patient being identified by different numbers. To address this, in collaboration with the data manager, I created an algorithm to match patients using other identifiers. This was challenging due to the unclean nature of the data (e.g. geographical spellings and discrepancies in dates), which led to numerous 'fuzzy matches' being disregarded (as described in Chapter 3). I only included those whom I had confidence had been correctly identified. I was conservative in matching patients across datasets in order to maintain robustness. To improve the quality of data and data linkage in future outbreak settings, a unique patient identifier is a key requirement.

ii) Concerns regarding the reliability of the data: Given the type of data used, reliability is a known and accepted risk. The results obtained are only as good as the data used. Examples of unreliability include unrealistic ages and instances of a single patient being recorded as both male and female in different entries. These

data were excluded from my analysis. To improve reliability, data should be subject to regular quality checks to ensure that datasets accurately reflect the population they represent.

iii) The use of free text: There are many instances in which data were not coded. This is particularly true of geographical data. Inconsistent recording and spelling made it difficult to use this data fully, as the incremental gain is not always justified by time required to clean the data.

iv) Lack of consistency of definitions: Ebola datasets were often cross-sectional and, therefore, the evolution of symptoms or final outcomes was difficult to track over time. For example, 'alert' and 'laboratory' data record the 'outcome' at a specific point in time i.e. the time of symptom onset and sampling, respectively. Therefore, if the patient is dead, the outcome is definitive, but if the patient is alive, this tells us nothing about the final outcome, which would be the key indicator for any risk factor analysis. 'Outcome' was used interchangeably with 'final outcome' or 'outcome at a specified point in time' across many datasets, which makes interpretation impossible. There were only two datasets that reportedly collated and updated results, so we used these as our first linkage steps and then attempted to fill the gaps where possible with other data sources.

Understanding how the data are recorded, as well as fully understanding the variable recorded, is essential to allow both accurate data interpretation and prioritisation of multiple data sources to enable the most comprehensive and accurate data linkage possible.

d) Limitations associated with specific data types

Different types of data have inherent advantages and disadvantages, and analysis needs to be interpreted in the light of these limitations, as outlined below.

i) Demographic surveillance data

One of the main limitations of this type of data is the low participation rate. The inherent limitations associated with this type of data in the context of the AC are discussed in Chapter 3. However, my work identified several additional limitations:

- High rates of missing data – it has been suggested that this is associated with participation 'fatigue'.

- Reporter bias: Field reports suggest that surveillance participants 'learn' how to answer the questions, either to give the answers they think questioners want to hear, or to minimise the time taken to complete the survey. This is particularly true with respect to sexual behaviour data.
- The data were gathered prior to the formulation of the research question, and, thus, are not focused on that topic. Therefore, they do not always provide the information needed. The data are collected to answer a range of questions, which evolve over time and, as a result, variables addressing specific questions may not be included, or may be missing in some years.

ii) Ebola outbreak response data

During an emergency response, efforts are, rightly, not focused on data collection or research. Clinical data were particularly challenging to collect - given the infection control measures necessary, nothing could be taken out of the 'hot zone' meaning that records had to be transcribed verbally over a physical barrier, while still in full protective outfit. Despite these obstacles, a wealth of useful data were collected. The challenge was how best to utilise the data available in a cost and time-efficient manner, to maximise the potential to be able to look back and improve both immediate and future responses.

e) Data security

As highlighted by my work, data security is a crucial aspect of data management, particularly with respect to phylogenetic data. There are many examples of national frameworks for data security with respect to personally-identifiable information, such as the Caldicott principles in the UK[450,451]. However, international recommendations are required, and should include the following concerns: data with personal identifiers should be anonymised before sharing and should only be shared on a need to know basis with approved collaborators; physical storage of data should be secure and it should be encrypted and backed up; and all research projects should have a data security protocol. Many of these important principles were not in place for the Ebola datasets.

10.7.3 Practical limitations

a) Cultural and communication challenges in obtaining information

International outbreak response work involves interactions and negotiations with a wide range of international actors. During this, I encountered organisational barriers and

bureaucracy, navigated relationships with varying personalities, dealt with sensitivities arising from cultural differences, and had to consider different political agendas. Requests were made for payments in return for the provision of data, an issue which required a careful response. Furthermore, I experienced different cultural views on gender equality, which made the work more challenging. My ability to undertake the project successfully varied depending on the state of emergency. It was far easier to garner cooperation during the worst periods of the outbreak, while people's focus was on response to the outbreak. However, once the focus shifted away from the immediate response to longer term issues, it became much harder to access data.

b) Personal risks of work in these settings

Furthermore, I experienced first-hand the very real dangers of working in outbreak settings, not just related to the epidemic. Africa has the highest rate of fatalities from road traffic injuries worldwide[452,453], and road traffic accidents (RTAs) are widely acknowledged to be the greatest risk to those working there. Unfortunately, I was involved in a serious RTA in Sierra Leone. This experience highlighted that the risks in these settings are real, and often arise from things we take for granted ourselves, such as safe transport systems.

10.8 Implications and recommendations for public health and policy

When I started work on this thesis, I hoped to find ways that my work could help to improve public health policy, and I have considered this issue throughout my research. Overall, combining traditional and molecular epidemiological approaches allows a more detailed understanding of outbreaks and, therefore, of steps that should be taken to limit their impact. Examples of possible benefits to public health policy arising from my work were discussed in section 10.3 above. They include the following themes: earlier implementation of interventions, development of optimised intervention strategies, targeting of interventions towards risk groups, development of appropriate combination interventions, and implementation of focused strategies to supplement population-level strategies in hyper-endemic settings.

In addition, phylogenetic research points towards some broader conclusions regarding the use of combined techniques to benefit public health and policy in general. These are discussed below.

a) Potential use in evaluating and refining public health interventions

Phylogenetics can be used to evaluate various public health interventions and, if appropriate, to identify why an intervention might be failing. For example, the first time this was used (in the HTPN 052 trial, discussed in Chapter 9)[413], serodiscordant HIV couples were randomly assigned either to standard clinical care (i.e. according to CD4+ count, or AIDS defining illness), or to immediate treatment (of the infected partner) regardless of CD4+ count. The hypothesis being tested was that this treatment should reduce, if not eliminate, transmission within couples receiving treatment. However, the trial data showed that infections continued to occur despite treatment. Phylogenetics was used to investigate how and why treatment was failing. The analysis showed that many of the transmissions had arisen from a source other than the index partner participant (20% of serconversion). Therefore, to successfully prevent these transmissions, an alternative intervention would be needed. Phylogenetics has significant potential to improve public health interventions in related scenarios, although substantive examples are still awaited given the technology is not yet widely adopted.

b) Potential to inform Global Health responses

An additional prospective use of phylogenetic analysis is its incorporation into global infectious disease preparedness and surveillance systems. These systems are led by WHO, aimed at the timely identification of increased incidence of endemic or emerging infections. In the case of the 2013-2016 West African Ebola outbreak, these systems performed poorly, and the four independent panels convened to review the response all implicated the slow recognition and response from WHO as a significant factor in the failure to contain the outbreak. Furthermore, another common recommendation of the panels was to develop stronger national core capacities in public health across the globe in order better to detect and respond to hazards, including infectious disease outbreaks, as agreed under the International Health Regulations (IHRs). This includes a global need for stronger infectious disease surveillance which echoes recommendations arising from the Kikwit Ebola outbreak in 1995, some 20 years earlier.

Integrating phylogenetic data into the standard epidemiological surveillance tools would provide an additional level of information which may assist in key areas. For example, in the case of the recent Ebola outbreak, earlier identification of the infectious agent and scale of the outbreak may have reduced the time taken to respond. An additional example

is provided by outbreaks of influenza – whereby novel strains and more pathogenic strains may require differential responses to seasonal influenza strains. Phylogenetics is able to determine strain type, parallel chains of transmission, and identifying the emergence of zoonotic strains in humans - allowing a reduced response time. Clearly, the incorporation of phylogenetics into surveillance strategies requires an effective infrastructure to collect and analyse patient samples; an infrastructure which is not currently present. However, it is hoped that this will develop over time, allowing the incorporation of phylogenetic analysis into global public health planning

c) Requirement for a global framework to address ethical implications

In the past, responses to ethical challenges have often been reactive rather than forward-looking. In addition, in outbreak settings it is, understandably, the case that patient care is prioritised. However, the probable adoption of phylogenetic techniques into public health responses provides the opportunity to tackle the related ethical concerns proactively, not just after the next public health emergency. I have set out at length in Chapter 9 the issues specific to scientific research, and although these remain relevant, in the global public health domain there are additional considerations that warrant mentioning. For example, although existing policy must balance individual freedoms with the public good, such as in setting out quarantine requirements, phylogenetic datasets may have implications that are more specific to certain individuals (source attribution), with potentially more serious outcomes for those individuals (e.g. in contexts of political or ethnic tension). Although it may be impossible to avoid any such problems, it is important to ensure that due consideration is given to such matters and that outcomes are not simply the by-products of policy decisions. Moreover, it is also important to ensure that these decisions are made at the appropriate level so that impartiality and accountability can be assured. To this end, the incorporation of phylogenetic considerations into existing global ethics frameworks would be of significant benefit and, I hope, is something that my work might help facilitate.

10.9 Considerations to advance the integration of phylogenetic analysis and traditional epidemiology for transmission dynamic studies in outbreak investigations

Below I set out six areas of future work that would benefit the integration of these two disciplines.

1. **Develop a framework to integrate epidemiological and molecular analysis**

This is a key requirement. As a minimum, sequence data, if available, should be used to confirm the origin of an outbreak as well as the source(s) (point source or multiple introductions) in complex outbreaks. In the simplest form, all epidemiological datasets could incorporate a single 'sequence' variable, such as cluster designation. This would be particularly useful in outbreaks that might include multiple viral lineages or parallel chains of transmission. In such cases, the cluster designation would ensure that the resulting transmission events are accurately classified based on genetic relatedness, preventing the incorrect inference of linkage between different viral lineages as can occur with traditional epidemiological analysis.

Of course, my hope is that these analytical tools can be integrated to a far greater extent than the bare minimum. To gain maximum benefit, new models are required which allow the combination of genetic data and multi-parameter epidemiological data within larger datasets. It seems probable that the greatest degree of insight will be provided by overlaying a phylogenetic tree with multiple epidemiological parameters, an undertaking that will require extensive testing and, thus, collaboration with mathematical modellers and statisticians, as well as with epidemiologists and geneticists.

2. **Data: improved quality and availability**

The quality of analysis is dependent upon the quality of the underlying data. This is particularly true of public health interventions based on real-time data. Therefore, there is a need to develop a sustainable process of data collection and management; one that is both time and cost-efficient. For example, implementing a standardised electronic data platform with an intuitive user interface (to minimise user expertise) could significantly improve the quality of data gathered during outbreaks and could facilitate real-time uploading of these data. In addition, a standardised cleaning process, with inbuilt quality

assurance checks could optimise data completeness. A standardised platform could be developed in such a way that it could rapidly be adjusted for disease-specific parameters, while still providing a low-tech and effective data collection tool.

Pre-determined structures for data governance would also be needed. Software packages already exist that have been designed for this purpose, for example EpiInfo[454], although their implementation in resource-poor settings which may have limited public health and IT infrastructures, is likely to remain a challenge.

Finally, as discussed in section 10.7.2 above, a best-practice standard for data sharing during health emergencies is needed. Real-time, evidence-based public health interventions are likely to be affected a great deal by the quality and availability of data, so the development of a template to facilitate the sharing of (deanonymised) data is essential.

3. **Real-time analysis for infectious disease surveillance and outbreaks**

Related to point number 2 above, the possibility of integrated analysis being used in real-time holds enormous promise. The possibility for prompt and effective interventions to be targeted at the right populations could save many lives. While such real-time use is dependent upon high-quality data, it is also necessary that these techniques are developed and incorporated into surveillance and response toolkits. This process requires the collaboration of numerous different disciplines. Without a cross-discipline effort to develop and apply advanced integrated analysis, real-time phylogenetic and epidemiological analysis will remain no more than a promising future opportunity.

4. **Next Generation Sequencing (NGS) as the mainstay of sequencing technology**

NGS can generates datasets on an unprecedented scale and is currently revolutionizing genome analysis. Over the last three years, NGS has become the first-choice method of genome sequencing in research, due to the lower cost and quicker turn-around time (both increasing the quantity and rate of data collection by several orders of magnitude) compared to traditional Sanger sequencing[455,456]. Given the rapid developments in the NGS field, it is worth considering the impact of NGS will have on this field.

NGS produces multiple short genetic reads and is capable of identifying multiple strains or different species within one individual. This has the potential to benefit phylogenetic analysis, particularly in resolving difficult phylogenetic relationships by utilising the multiple sequences within one host.

For example, it enables resolution of the direction of infection and identification of date of infection [351]. This could revolutionize the development of targeted interventions towards high-risk and vulnerable groups, but will come at the cost of magnified ethical considerations to mitigate the risks of identifying those who transmit an infection.

Phylogenetic methods will require modification and validation for use with NGS data. The challenges inherent in this have led to a time lag in applying NGS to phylogenetic studies. These include: the uncertainty about which analyses and evolutionary models are appropriate for these data, the challenges associated with assembling the short reads to genomes for use in conventional phylogenetic analysis, and the adaptation of models to account for the higher sequencing error rate known to occur with NGS (an error rate of 0.01% is currently considered good in NGS, but if this is extrapolated to the genome level, this would equate to thousands of errors, which could substantially affect the phylogenetic results obtained). Therefore, while NGS has the potential to benefit phylogenetic analysis, its application will require refinement and may be problematic in many cases. Furthermore, NGS analysis will present computational and statistical challenges, requiring significant development of bioinformatics.

5. Reinforcing a framework for the reporting of molecular epidemiological studies

Field *et al*. highlight the discrepancies in the reporting of 'molecular epidemiological' studies and set out recommendations for good scientific reporting (outlined in Chapter 2) to ensure quality and transparency[1]. The importance of such a framework being more widely applied as a standard publication requirement has become apparent during the course of this work (as described in the limitations above). There is a need for more rigorous reporting of molecular studies and journals should encourage transparent reporting in line with STROME-ID as a publication requirement. This should specifically include clear guidance on how to reproduce the methods use and how to interpret the results (including any uncertainty).

Furthermore, this work has shown the power of a combined approach, incorporating both molecular and epidemiological datasets. STROME-ID was an extension of the initial STROBE guidance for reporting epidemiological studies. Thus, it already includes variables which cover reporting of epidemiological studies[457]. However, further extensions to STROME-ID would benefit this evolving field, for example recommendations for methods of reporting data linkage to combine sequence and epidemiological data, and an extension for NGS.

6. An ethical framework

As discussed in Chapter 9, there is a need to develop an effective and sustainable model of good ethical practice in phylogenetic research to help minimise the risks to individuals/groups participating in studies while optimising the scientific benefits.

10.10 Recommendations for future research

Table 10.2 summarises this thesis in the context of historical work, the key developments and goals for future work. In summary, I propose the following recommendations for ongoing research in this area:

- The incorporation of these integrated methods into real-time surveillance mechanisms. One potential option would be to set up an algorithm for automated generations of phylogenetic trees from sequence data at AC, which could then be combined with epidemiological parameters. For example, a first step could be to further evaluate specifically the 'New Cluster' (and other emergent clusters) to understand if it is still expanding and what is driving this cluster of infections. This would provide a good case study to secure funding for more extensive research
- Extend this work to other pathogens. Different approaches are required for different pathogens and research questions
- Explore using NGS techniques and the added benefit this may bring to understanding transmission dynamics
- Develop statistical techniques to infer the proportion of the unsampled population. This will allow more accurate conclusions to be drawn about the representativeness and generalisability of the results, which may impact on intervention development
- Formal analyses of new bioinformatics models comparing transmission trees developed through epidemiology versus phylogenetic versus a combined model to

evaluate which programmes provide the most accurate representation of the true underlying transmission dynamics.

10.11 Conclusions

In this thesis, I have demonstrated how integrating molecular and epidemiological analysis offers an enhanced understanding of the transmission dynamics of infectious disease outbreaks and how this knowledge can be used to significantly improve public health interventions. Therefore, I believe should these combined methods should be incorporated into the standard epidemiological analysis of transmission dynamics in future outbreaks. This is critical at a time when infectious disease outbreaks have led to some of the most significant global health threats of the recent past.

Table 10.2 This thesis in context: Summary of key findings

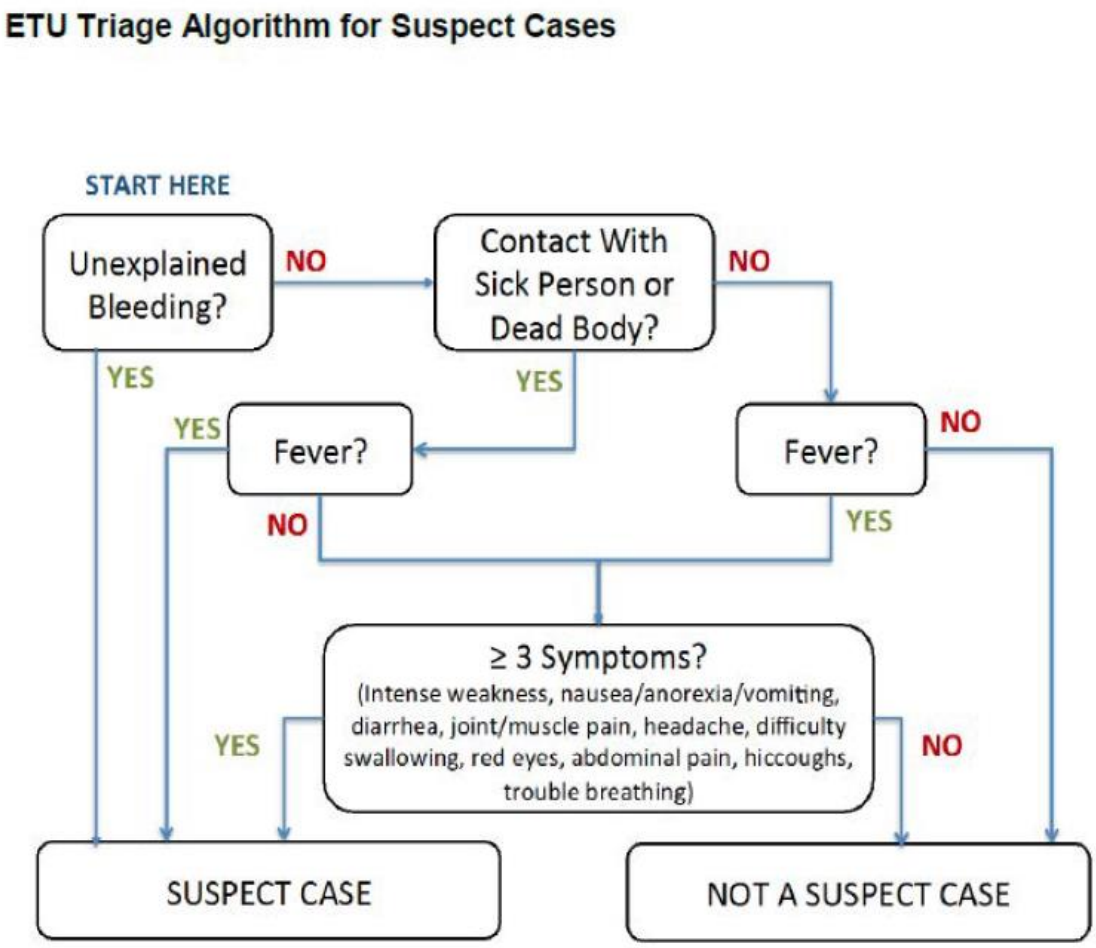
Evidence before this study	• Contact tracing and traditional epidemiology was used as the mainstay for investigating transmission dynamics • Phylogenetics analysis emerged as a novel research tool to define transmission events, but the approaches were used in isolation from epidemiological and clinical methods. The utility of phylogenetics as a tool for outbreak investigation remained unclear and it had not been previously used in real-time in outbreaks of emerging infections or for population level investigation of generalised outbreaks • Few examples existed using a combined approach – these were limited to concentrated outbreaks in high resource research settings with good sample coverage. This initial work suggested the potential to enhance knowledge by using such combined approaches • No work had investigated these approaches in real life settings i.e. at a population level, in acute outbreak settings or for generalised outbreaks.
Added value of using a combined approach – scientific contribution and implications for public health prevention	Both disciplines are complementary and most powerful when used in combination Novel insights into transmission dynamics of outbreaks from this work include: • The HIV outbreak at AC is seeded by external introductions • Provided further evidence to support the association of both migration and HIV acquisition risk, and high HIV incidence in areas with new industrial developments • Earlier identification of new emerging clusters • The Village X outbreak originated from the fishing village in Freetown and was the only introduction into the village (point-source outbreak) • Identified novel mechanisms of transmission allowing targeting of interventions e.g. the Village X hypothesis of children infected from contaminated water, which appears to be more consistent with the data, rather than the parent to child transmission • Identification of super-spreading events/groups This combined approach holds promise for: • Understanding the sources and modes of transmission with outbreaks • Earlier identification of emerging clusters • More detailed information on clusters • Understanding complex transmission chains with parallel transmission events • Informing targeted intervention strategies All of the above require finer scale transmission understanding than is possible through either discipline alone
Moving forwards/ Next steps	• The maximal benefit is likely to be seen in the real-time application of these combined methods • Incorporation of combined approach into the framework for standard outbreak investigation analyses (Table 10.1) • Increased used with increasingly available sequencing and automated analysis tools • Next generation sequencing adds another dimension to potential analyses by enabling easier identification of time of infection and directionality of infection by using depth of sequencing and intra-host variation • Increased technological capacity to support such methods

APPENDICES

LIST OF APPENDICES:

Numbered by chapter for ease of referencing

APPENDIX 6.1: MSF Clinical Triage protocol[325]



Full clinical guidance found at: https://www.medbox.org/ebola-guidelines/filovirus-haemorrhagic-feverguideline/previous.

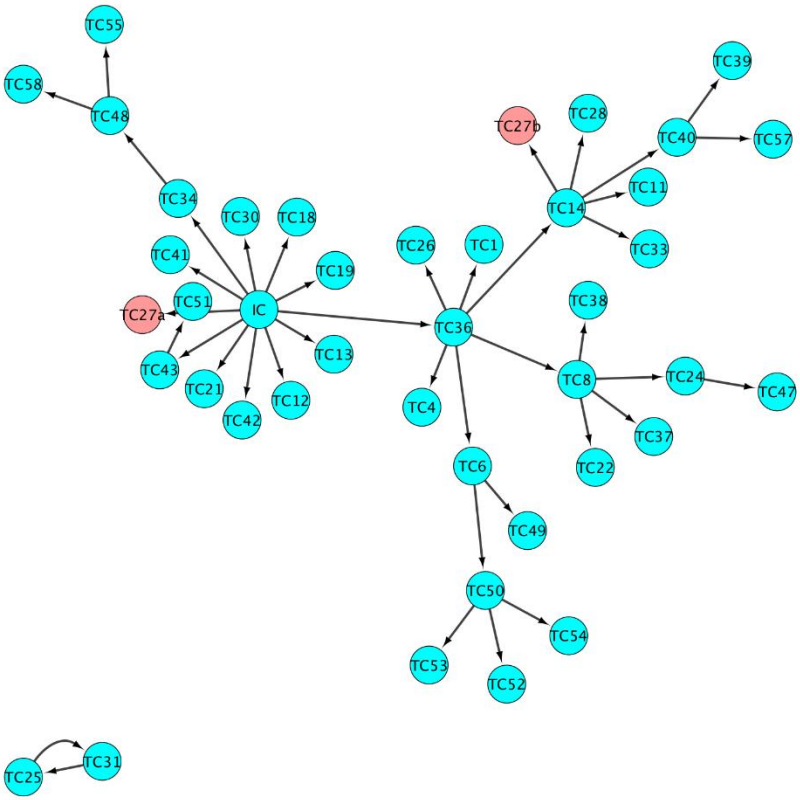
APPENDIX 6.2: Probability of transmission events as determined by Outbreaker

Ancestor sequence	Transmitted to seqence	Probability of transmission event
TC36	TC1	0.34
TC36	TC4	0.71
TC36	TC6	0.89
TC36	TC8	0.63
TC14	TC11	0.94
IC	TC12	0.99
IC	TC13	0.58
TC36	TC14	0.58
IC	TC18	0.37
IC	TC19	0.36
IC	TC21	0.66
TC8	TC22	0.38
TC8	TC24	0.37
TC31	TC25	0.5
TC36	TC26	0.84
IC	TC27a	0.3
TC14	TC27b	0.96
TC14	TC28	0.93
IC	TC30	0.29
TC25	TC31	0.5
TC14	TC33	0.87
IC	TC34	0.66
IC	TC36	0.54
TC8	TC37	0.42
TC8	TC38	0.33
TC40	TC39	1
TC14	TC40	0.94
IC	TC41	0.37
IC	TC42	0.36
IC	TC43	0.35
TC24	TC47	0.21
TC34	TC48	0.21
TC6	TC49	0.49
TC6	TC50	0.41
TC43	TC51	0.23
TC50	TC52	0.38
TC50	TC53	0.3
TC50	TC54	0.38
TC48	TC55	0.14
TC40	TC57	0.06
TC48	TC58	0.07

APPENDIX 6.3: Number of Single Nucleotide Polymorphisms in pairwise comparison between each of the 42 sequences

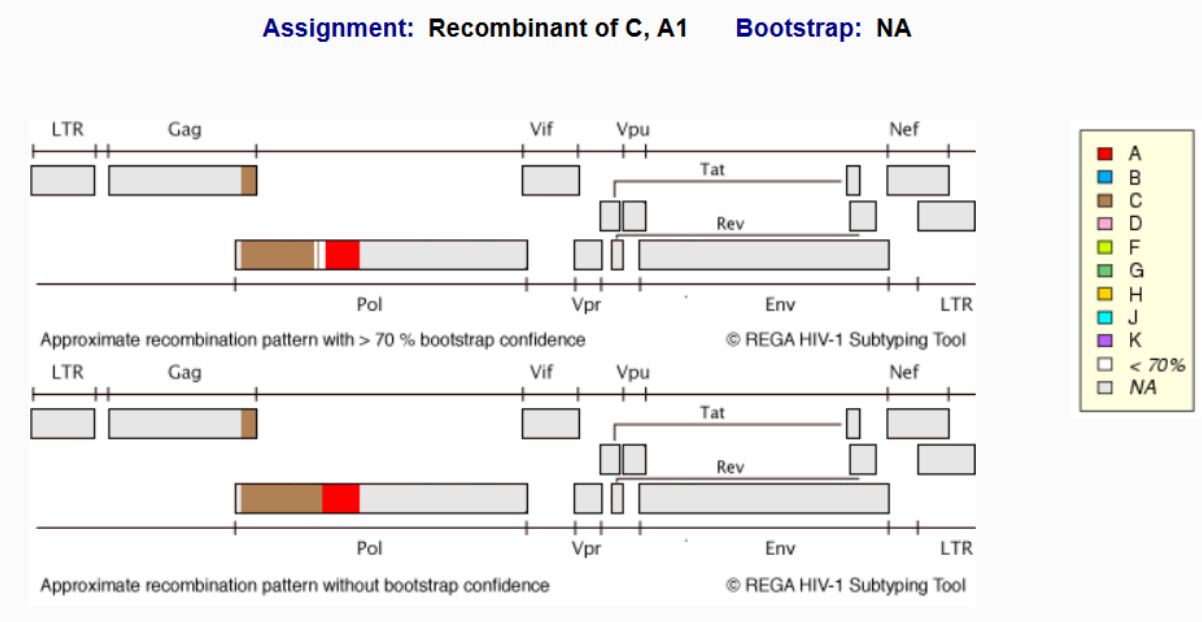
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	NN	OO	PP
A	NA	0	1	1	0	2	2	2	1	0	0	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	0	3	3	4	2	5	2	2	1	2	1	2	2	3	2
B	0	NA	1	1	0	2	2	2	1	0	0	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	0	3	3	4	2	5	2	2	1	2	2	2	2	3	2
C	1	1	NA	0	1	1	1	1	2	1	1	1	3	6	1	0	2	0	3	3	4	1	2	1	2	5	1	1	4	2	3	3	4	1	1	0	3	1	1	1	2	1
D	1	1	0	NA	1	1	1	1	2	1	1	1	3	6	1	0	2	0	3	3	4	1	2	1	2	5	1	1	4	2	3	3	4	1	1	0	3	1	1	1	2	1
E	0	0	1	1	NA	2	2	2	1	0	0	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	0	3	3	4	2	5	2	2	1	2	2	2	2	3	2
F	2	2	1	1	2	NA	2	2	3	2	2	2	2	5	2	1	1	1	2	2	5	2	3	2	1	4	2	2	5	3	2	2	3	2	2	1	2	1	0	2	3	2
G	2	2	1	1	2	2	NA	2	3	2	2	2	4	7	2	1	3	1	4	4	5	2	3	2	3	6	2	2	5	3	4	4	5	2	2	1	4	1	2	2	3	2
H	2	2	1	1	2	2	2	NA	3	2	2	2	4	7	2	1	3	1	4	4	5	2	3	2	3	6	2	2	5	3	4	4	5	2	2	1	4	1	2	2	3	2
I	1	1	2	2	1	3	3	3	NA	1	1	1	3	6	3	2	4	2	3	3	3	3	4	3	4	6	1	1	4	4	5	3	6	3	3	2	3	2	3	3	4	3
J	0	0	1	1	0	2	2	2	1	NA	0	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	0	3	3	4	2	5	2	2	1	2	2	2	2	3	2
K	0	0	1	1	0	2	2	2	1	0	NA	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	0	3	3	4	2	5	2	2	1	2	2	2	2	3	2
L	0	0	1	1	0	2	2	2	1	0	0	NA	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	0	3	3	4	2	5	2	2	1	2	1	2	2	3	2
M	2	2	3	3	2	2	4	4	3	2	2	2	NA	5	13	3	3	3	2	2	11	4	5	4	3	6	2	2	7	5	4	2	5	4	5	3	2	4	2	5	6	4
N	5	5	6	6	5	5	7	7	6	5	5	5	5	NA	7	6	6	6	5	5	7	7	8	7	6	8	5	5	8	6	7	5	8	7	7	6	5	7	5	7	8	7
O	2	2	1	1	2	2	2	2	3	2	2	2	13	7	NA	1	3	1	4	4	7	2	3	2	3	6	2	2	5	3	4	4	5	2	2	1	4	2	2	2	3	2
P	1	1	0	0	1	1	1	1	2	1	1	1	3	6	1	NA	2	0	3	3	4	1	2	1	2	5	1	1	4	2	3	3	4	1	1	0	3	1	1	1	2	1
Q	3	3	2	2	3	1	3	3	4	3	3	3	3	6	3	2	NA	2	3	3	6	3	4	3	2	5	3	3	6	4	3	3	4	3	3	2	3	3	1	3	4	3
R	1	1	0	0	1	1	1	1	2	1	1	1	3	6	1	0	2	NA	3	3	4	1	2	1	2	5	1	1	4	2	3	3	4	1	1	0	3	1	1	1	2	1
S	2	2	3	3	2	2	4	4	3	2	2	2	2	5	4	3	3	3	NA	2	5	4	5	4	3	6	2	2	5	5	4	2	5	4	4	3	2	4	2	4	5	4
T	2	2	3	3	2	2	4	4	3	2	2	2	2	5	4	3	3	3	2	NA	5	4	5	4	3	6	2	2	5	5	4	2	5	4	4	3	2	4	2	4	5	4
U	3	3	4	4	3	5	5	5	3	3	3	3	11	7	7	4	6	4	5	5	NA	5	6	5	5	9	3	3	8	6	7	4	8	5	7	4	5	5	5	7	7	5
V	2	2	1	1	2	2	2	2	3	2	2	2	4	7	2	1	3	1	4	4	5	NA	3	2	3	6	2	2	5	3	4	4	5	2	2	1	4	1	2	2	3	2
W	3	3	2	2	3	3	3	3	4	3	3	3	5	8	3	2	4	2	5	5	6	3	NA	3	4	7	3	3	6	4	5	5	6	3	3	2	5	2	3	3	4	3
X	2	2	1	1	2	2	2	2	3	2	2	2	4	7	2	1	3	1	4	4	5	2	3	NA	3	6	2	2	5	3	4	4	5	2	0	1	4	2	2	2	3	2
Y	3	3	2	2	3	1	3	3	4	3	3	3	3	6	3	2	2	2	3	3	5	3	4	3	NA	4	3	3	6	4	3	3	4	3	3	2	3	3	1	3	4	3
Z	6	6	5	5	6	4	6	6	6	6	6	6	6	8	6	5	5	5	6	6	9	6	7	6	4	NA	6	6	9	7	6	5	1	5	6	5	6	5	4	6	7	5
AA	0	0	1	1	0	2	2	2	1	0	0	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	NA	0	3	3	4	2	5	2	2	1	2	2	2	2	3	2
BB	0	0	1	1	0	2	2	2	1	0	0	0	2	5	2	1	3	1	2	2	3	2	3	2	3	6	0	NA	3	3	4	2	5	2	2	1	2	2	2	2	3	2
CC	3	3	4	4	3	5	5	5	4	3	3	3	7	8	5	4	6	4	5	5	8	5	6	5	6	9	3	3	NA	6	7	5	8	3	5	4	5	5	5	5	6	5
DD	3	3	2	2	3	3	3	3	4	3	3	3	5	6	3	2	4	2	5	5	6	3	4	3	4	7	3	3	6	NA	5	5	6	3	3	2	5	3	3	3	4	3
EE	4	4	3	3	4	2	4	4	5	4	4	4	4	7	4	3	3	3	4	4	7	4	5	4	3	6	4	4	7	5	NA	4	5	4	4	3	4	4	2	4	5	4
FF	2	2	3	3	2	2	4	4	3	2	2	2	2	5	4	3	3	3	2	2	4	4	5	4	3	5	2	2	5	5	4	NA	5	4	4	3	2	4	2	4	5	4
GG	5	5	4	4	5	3	5	5	6	5	5	5	5	8	5	4	4	4	5	5	8	5	6	5	4	1	5	5	8	6	5	5	NA	5	5	4	5	5	3	5	6	5
HH	2	2	1	1	2	2	2	2	3	2	2	2	4	7	2	1	3	1	4	4	5	2	3	2	3	5	2	2	3	3	4	4	5	NA	2	1	4	1	2	2	3	2
II	2	2	1	1	2	2	2	2	3	2	2	2	5	7	2	1	3	1	4	4	7	2	3	0	3	6	2	2	5	3	4	4	5	2	NA	1	4	2	2	2	3	2
JJ	1	1	0	0	1	1	1	1	2	1	1	1	3	6	1	0	2	0	3	3	4	1	2	1	2	5	1	1	4	2	3	3	4	1	1	NA	3	0	1	1	2	1
KK	2	2	3	3	2	2	4	4	3	2	2	2	2	5	4	3	3	3	2	2	5	4	5	4	3	6	2	2	5	5	4	2	5	4	4	3	NA	3	2	4	5	4
LL	1	2	1	1	2	1	1	1	2	2	2	1	4	7	2	1	3	1	4	4	5	1	2	2	3	5	2	2	5	3	4	4	5	1	2	0	3	NA	1	2	3	2
MM	2	2	1	1	2	0	2	2	3	2	2	2	2	5	2	1	1	1	2	2	5	2	3	2	1	4	2	2	5	3	2	2	3	2	2	1	2	1	NA	2	3	2
NN	2	2	1	1	2	2	2	2	3	2	2	2	5	7	2	1	3	1	4	4	7	2	3	2	3	6	2	2	5	3	4	4	5	2	2	1	4	2	2	NA	3	2
OO	3	3	2	2	3	3	3	3	4	3	3	3	6	8	3	2	4	2	5	5	7	3	4	3	4	7	3	3	6	4	5	5	6	3	3	2	5	3	3	3	NA	3
PP	2	2	1	1	2	2	2	2	3	2	2	2	4	7	2	1	3	1	4	4	5	2	3	2	3	5	2	2	5	3	4	4	5	2	2	1	4	2	2	2	3	NA

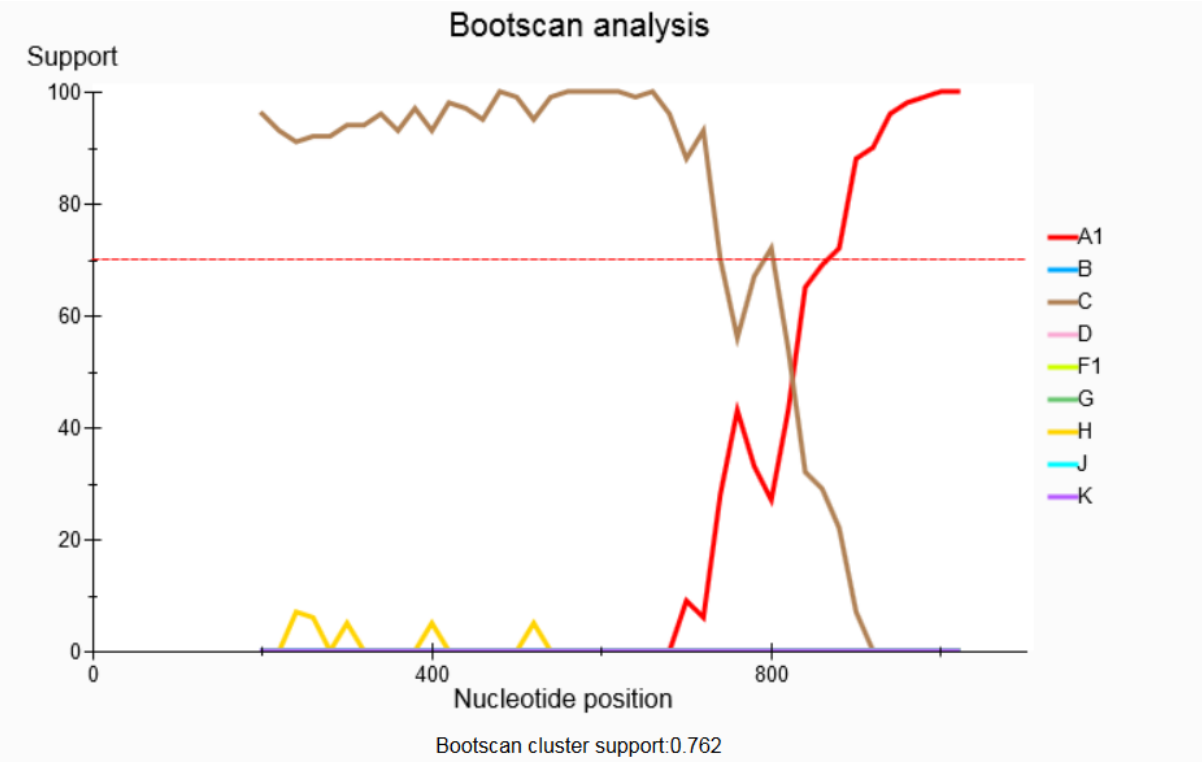
APPENDIX 6.4: Outbreaker transmission tree showing displacement across generations of duplicate sample



Key: Pink =duplicate sample

APPENDIX 7.1: Example recombinant sequence in dataset: non sub-type C sequences





APPENDIX 7.2: Comparator table including all negative population at AC (therefore, at risk of HIV acquisition)

		All Sequences	All negatives	p-value: All vs Negs
Eligible individuals		1,376	31,902	
BASIC DEMOGRAPHIC INFORMATION				
Gender				
	Male (%)	30	45.7	
	Female (%)	70	54.3	**p<0.001**
Age (years)				
	Mean (years)	33.5	35.5	
	SD	11.7	20.2	**p<0.001**
	<20 (%)	8.3	20.4	
	20-24 (%)	17.1	23.9	
	25-29 (%)	18.5	14.1	
	30-34 (%)	15.4	7.0	
	35-39 (%)	14.0	3.6	
	40-44 (%)	9.7	3.2	
	45-49 (%)	6.3	3.2	
	>50 (%)	10.7	24.6	**p<0.001**
Wealth quintiles				
	Most deprived (%)	18.6	18.5	
	2nd most deprived (	19.9	18.9	
	Middle quintile (%)	21.8	18.1	
	2nd least deprived (	20.5	18.2	
	Least deprived (%)	14.1	17.1	
	Missing (%)	5.1	9.2	**p<0.001**
Maximum Education				
	None (%)	5.3	19.3	
	Primary (year 1-7)(%	31.9	34.6	
	Secondary (8-12)(%)	43.6	20.4	
	Tertiary (>12)(%)	0.9	1.0	
	Missing(%)	18.3	24.6	**p<0.001**
In employment				
	Yes, FT (%)	17.7	18.0	
	Yes, PT (%)	7.9	3.4	
	No (%)	51.4	52.2	
	Missing (%)	23.0	26.4	**p<0.001**
Marital Status				
	Never married (%)	72.1	28.9	
	Married (%)	7.3	5.6	
	Engaged (%)	6.9	0.9	
	Divorced/Separate d/Widowed (%)	4.0	4.9	
	Missing (%)	9.7	59.7	p=0.7
Residence HIV prevalence				
	High (%)	30.3	19.0	
	Mid (%)	48.8	51.8	
	Low (%)	16.1	29.2	
	External (%)	4.7	0.0	
	Missing (%)	0.1	7.3	**p<0.001**
	HIV prev median	25.9	23.7	
	HIV prevalence IQR	22.1-30.8	18.4-27.8	**p<0.001**

		All Sequences	All negatives	p-value: All vs Negs
CLINCIAL VARIABLES				
ART treatment				
	Yes (%)	13.8	0.0	
	No (%)	36.1	1.6	
	Missing (%)	50.1	98.4	**p<0.001**
SEXUAL BEHAVIOUR				
Multiple partners ever reported				
	Yes (%)	9.5	5.5	
	No (%)	67.2	22.7	
	Missing (%)	23.3	71.8	**p<0.001**
Circumcised	Men only			
	Yes (%)	8.8	13.2	
	No (%)	79.5	44.3	
	Missing (%)	11.7	42.5	**p<0.001**
Condom use				
	Always (%)	12.9	5.6	
	Sometimes (%)	20.2	7.2	
	Never (%)	14.0	7.2	
	Missing (%)	52.9	80.0	**p<0.001**

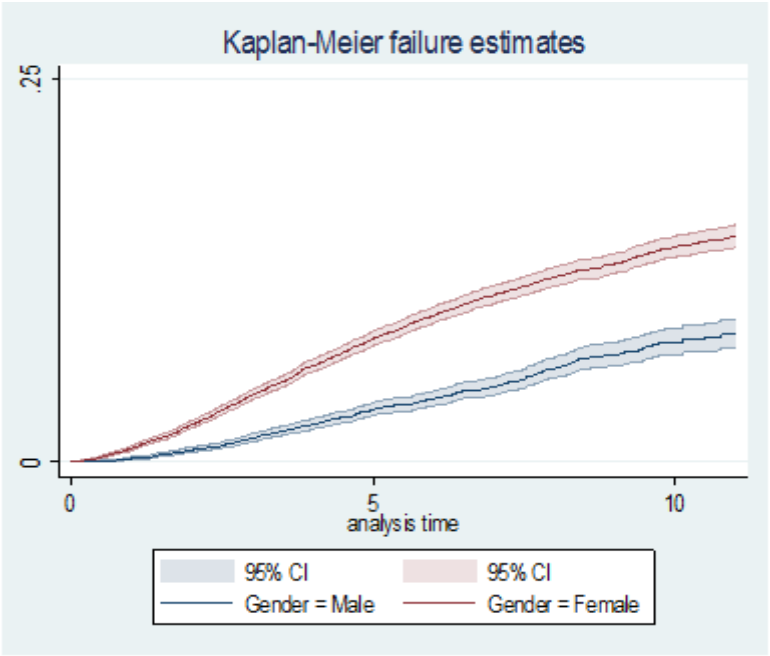
APPENDIX 7.3: Pearson correlation coefficient for main risk factors in the analysis

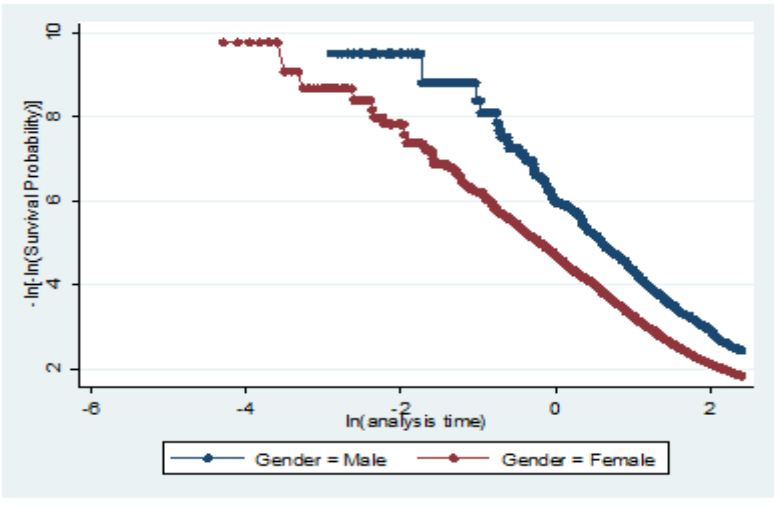
	Gender	Age at sample	Employment	Education	Wealth	HIV prevalence of dwelling	Circumcised	Ever reported mulitple partners
Gender	1							
Age at Sample	-0.1169	1						
Employed	-0.1548	0.0675	1					
Education	-0.03	0.0425	0.2222	1				
Wealth	0.0009	-0.02	0.1509	0.1964	1			
HIV prevalence of dwelling	0.0387	-0.0335	0.0874	-0.0005	0.1991	1		
Circumcised	.	-0.0236	0.0959	0.0225	0.0125	0.0179	1	
Ever reported mulitple partners	-0.3683	-0.0469	0.0367	0.1312	-0.0071	-0.0258	0.041	1

All correlates <0.8

APPENDIX 8.1: Proportional Hazard (PH) assumption violation explored by gender

Both the Kaplan-Meier curve for probability of HIV seroconversion by Gender over time and the graphical representation of the PH log curve show that the probability of seroconversion (or hazard, respectively) are more similar between genders at the start of the study and diverge over time. There are no significant deviations or crossing of the curves, therefore the PH models were retained.





Appendix 8.2: Interacted models to explore for effect modification

MEN	Uninteracted model (table 8.4 - column 4)		A: Mig#age + prev		B: Mig#age + prev#age		C: Mig + Prev#age	
	HR	ci95	HR	ci95	HR	ci95	HR	ci95
No migration	1.00						1.00	1.000,1.000
Internal migration	1.30	0.79,2.15					1.25	0.754,2.071
In migration from External	1.04	0.65,1.66					1.00	0.624,1.591
Out migration (External time)	1.81	1.27,2.57					1.71	1.185,2.473
Low Prevalence	1.00							
Mid Prevalence	1.36	1.02,1.81	1.31	0.981,1.751				
High Prevalence	1.99	1.46,2.72	1.90	1.386,2.592				
Age: <20	1.00							
20-25	3.46	2.54,4.74						
26-40	3,16	2.28,4.39						
41-60	1.02	0.67,1.55						
>60	0.46	0.24,0.88						
Interactions:								
Int_mig_L20			2.60	0.928,7.261	2.59	0.924,7.238		
Int_mig_20_25			1.24	0.577,2.653	1.23	0.574,2.639		
Int_mig_26_40			0.63	0.199,1.973	0.63	0.199,1.982		
Int_mig_41_60			2.83	0.675,11.90	2.90	0.689,12.225		
Int_mig_G60			0.00	0.000,.	0.00	0.000,0.000		
Inward_mig_L20			2.59	0.624,10.78	2.62	0.631,10.914		
Inward_mig_20_25			1.10	0.593,2.048	1.11	0.597,2.063		
Inward_mig_26_40			0.76	0.335,1.743	0.76	0.333,1.735		
Inward_mig_41_60			0.00	0.000,.	0.00	0.000,0.000		
Inward_mig_G60			0.00	0.000,.	0.00	0.000,0.000		
Out_mig_L20			3.11	1.402,6.913	2.89	1.148,7.295		
Out_mig_20_25			1.58	1.021,2.444	1.79	1.048,3.068		
Out_mig_26_40			1.01	0.566,1.797	1.02	0.526,1.994		
Out_mig_41_60			3.89	1.573,9.629	2.95	1.056,8.241		
Out_mig_G60			0.00	0.000,.	0.00	0.000,0.000		
Med_prev_L20					1.14	0.549,2.366	0.97	0.516,1.831
Med_prev_20_25					1.48	0.915,2.395	1.45	0.955,2.191
Med_prev_26_40					1.38	0.830,2.297	1.73	1.092,2.752
Med_prev_41_60					0.96	0.425,2.154	0.81	0.393,1.655
Med_prev_G60					1.29	0.323,5.177	1.33	0.331,5.318
High_prev_L20					1.84	0.834,4.051	1.57	0.780,3.172
High_prev_20_25					2.30	1.388,3.821	2.25	1.443,3.504
High_prev_26_40					1.83	1.059,3.172	2.30	1.390,3.814
High_prev_41_60					1.14	0.428,3.034	0.98	0.396,2.404
High_prev_G60					1.69	0.281,10.116	1.73	0.289,10.391
Observations	68949		68949		68949		68949	
AIC	5708		5716		5722		5718	
	BEST MODEL							

WOMEN	Uninteracted model (table 8.4 - column 5)		A: Mig#age + prev		B: Mig#age + prev#age		C: Mig + Prev#age	
	HR	ci95	HR	ci95	HR	ci95	HR	ci95
No migration	1.00						1.00	1.000,1.000
Internal migration	1.08	0.84,1.37					1.04	0.813,1.327
In migration from External	0.86	0.66,1.12					0.83	0.636,1.078
Out migration (External time)	1.12	0.92,1.35					1.08	0.883,1.311
Low Prevalence	1.00							
Mid Prevalence	1.27	1.11,1.44	1.26	1.109,1.434				
High Prevalence	1.55	1.33,181	1.54	1.326,1.797				
Age: <20	1.00							
20-25	1.49	1.30,1.71						
26-40	0.65	0.56,0.76						
41-60	0.17	0.14,0.21						
>60	0.03	0.02,0.05						
Interactions:								
Int_mig_L20			1.27	0.827,1.956	1.27	0.827,1.957		
Int_mig_20_25			0.90	0.610,1.325	0.89	0.602,1.307		
Int_mig_26_40			0.81	0.451,1.437	0.82	0.459,1.462		
Int_mig_41_60			2.26	1.055,4.846	2.26	1.056,4.850		
Int_mig_G60			0.00	0.000,0.000	0.00	0.000,0.000		
Inward_mig_L20			1.11	0.571,2.154	1.11	0.571,2.155		
Inward_mig_20_25			0.75	0.531,1.046	0.75	0.531,1.046		
Inward_mig_26_40			0.78	0.423,1.420	0.77	0.418,1.404		
Inward_mig_41_60			1.79	0.443,7.256	1.79	0.441,7.225		
Inward_mig_G60			19.82	2.621,149.95	20.88	2.724,159.97		
Out_mig_L20			0.92	0.610,1.392	0.83	0.540,1.282		
Out_mig_20_25			0.98	0.770,1.251	1.04	0.792,1.359		
Out_mig_26_40			1.41	0.978,2.042	1.37	0.914,2.049		
Out_mig_41_60			1.88	0.692,5.133	2.11	0.741,5.981		
Out_mig_G60			0.00	0.000,0.000	0.00	0.000,0.000		
Med_prev_L20					1.08	0.857,1.368	1.15	0.918,1.436
Med_prev_20_25					1.31	1.064,1.609	1.32	1.088,1.601
Med_prev_26_40					1.30	0.993,1.714	1.21	0.944,1.554
Med_prev_41_60					1.54	1.017,2.317	1.45	0.978,2.158
Med_prev_G60					1.90	0.595,6.067	1.98	0.625,6.245
High_prev_L20					1.38	1.051,1.816	1.47	1.126,1.915
High_prev_20_25					1.80	1.416,2.278	1.81	1.440,2.263
High_prev_26_40					1.21	0.856,1.716	1.12	0.807,1.559
High_prev_41_60					1.56	0.928,2.614	1.47	0.887,2.438
High_prev_G60					4.31	1.250,14.829	4.29	1.255,14.670
Observations	113312		113312		113312		113312	
AIC	23854		23857		23862		23859	
	BEST MODEL							

A difference in AIC of >2 between models suggests a significant difference between the models with the lower score representing the best fit. Therefore, the uninteracted model (as presented in the main thesis) appears to fit the data best.

NB The unexpectedly high HR for the older age ranges (>60 predominantly) is likely to be due to the small numbers and seroconversions contributing to this category.

REFERENCES

1. Field N, Cohen T, Struelens MJ, et al. Strengthening the reporting of molecular epidemiology for infectious diseases (STROME-ID): an extension of the STROBE statement. *The Lancet Infectious Diseases* 2014; **14**(4): 341-52. doi:10.1016/s1473-3099(13)70324-4.
2. Maxmen A. How the fight against ebola tested a culture's traditions. 2015. https://news.nationalgeographic.com/2015/01/150130-ebola-virus-outbreak-epidemic-sierra-leone-funerals/. (accessed 13 April 2016)
3. WHO. How to conduct safe and dignified burial of a patient who has died from suspected or confirmed Ebola virus disease. 2014. http://www.who.int/csr/resources/publications/ebola/safe-burial-protocol/en/ (accessed 16 August 2016).
4. Dennis AM, Herbeck JT, Leigh Brown A, et al. Phylogenetic studies of transmission dynamics in generalized HIV epidemics: An essential tool where the burden is greatest? *Journal of Acquired Immune Deficiency Syndrome* 2014; **67**(2): 181-95.
5. Chalmet K, Staelens D, Blot S, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. *BMC Infectious Diseases* 2010; **10**: 262. doi:10.1186/1471-2334-10-262.
6. Pillay D, Herbeck JT, Cohen MS, et al. PANGEA-HIV: Phylogenetics for generalised epidemics in Africa. *The Lancet Infectious Diseases*; **15**(3): 259-61.
7. International AIDS Society Conference. Personal Communication. Durban; 2016.
8. Moon S, Sridhar D, Pate MA, et al. Will Ebola change the game? Ten essential reforms before the next pandemic. The report of the Harvard-LSHTM Independent Panel on the Global Response to Ebola. *The Lancet* 2015; **386**(10009): 2204-21. doi:10.1016/s0140-6736(15)00946-0.
9. WHO. Report of the Ebola Interim Assessment Panel, 2015. http://www.who.int/csr/resources/publications/ebola/ebola-panel-report/en/. (Accessed 16 August 2016)
10. United Nations. Protecting humanity from future health crises. 2016. http://www.un.org/News/dh/infocus/HLP/2016-02-05_Final_Report_Global_Response_to_Health_Crises.pdf (accessed 16 August 2016).
11. National Academy of Medicine. Global Health Risk Framework. 2016. https://nam.edu/initiatives/global-health-risk-framework/ (accessed 25 August 2017).
12. Kilbourne ED. the molecular epidemiology of influenza. *Journal of Infectious Diseases* 1973; **127**: 478-87.
13. Vineis P, McMichael AJ. Bias and confounding in molecular epidemiological studies: special considerations. *Carcinogenesis* 1998; **19**(12): 2063-67.
14. Hall A. What is molecular epidemiology? *Tropical Medicine and International Health* 1996; **1**: 407-8.
15. Monis PT, Andrews RH. Molecular Epidemiology: Assumptions and limitations of commonly applied methods. *International ournal of Parasitology* 1998; **28**(6): 981-7.
16. Chia KS, Shi CY, Lee JH, Seow A, Lee HP. Molecular epidemiology: issues in study design and statistical analysis. *Annals of the Academy of Medicine Singapore* 1996; **25**: 55-63.
17. Darwin C. On the Origin of Species, by means of natural selection, or the preservation of favoured races in the struggle for life. London: Murray, J; 1859.
18. Pereira L, Cassai E, Honess TW, Roizman B, Terni M, Nahmias A. Variability in the structural polypeptides of herpes simplex virus 1 strains: potential application in molecular epidemiology. *Infection and Immunology* 1976; **13**: 211-20.

19. Summers WC. Molecular epidemiology of DNA viruses: applications of restriction endonuclease cleavage site analysis. *The Yale Journal of Biology and Medicine* 1980; **53**(1): 55-9.
20. McDade JE, Anderson BE. Molecular epidemiology: applications of nucleic acid amplification and sequence analysis. *Epidemiological Review* 1996; **18**: 90-7.
21. Ou CY, Ciesielski CA, Myers G, al. e. Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992; **256**: 1165-71.
22. Campbell P. Editorial: Benefits of sharing. *Nature* 2016; **30**: 129.
23. Estevezj. The Sanger (chain-termination) method for DNA sequencing. 2012. http://commons.wikimedia.org/wiki/File:Sanger-sequencing.svg?uselang=pt-br (accessed 5 August 2016).
24. WHO. A research and development blueprint for action to prevent epidemics. 2016. http://www.who.int/csr/research-and-development/en/ (accessed 12 July 2016).
25. Harris SR, Cartwright EJP, Török ME, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet Infectious Diseases* 2013; **13**(2): 130-6. doi:10.1016/s1473-3099(12)70268-2.
26. Goedhals D, Rossouw I, Hallbauer U, Mamabolo M, de Oliveira T. The tainted milk of human kindness.*The Lancet* 2012; **380**(9842):702.doi:10.1016/s0140-6736(12)60957-x.
27. Butler D. Lawyers call for science to clear AIDS nurses in Libya. *Nature* 2006; **443**(7109): 254. doi:10.1038/443254b.
28. de Oliveira T, Pybus OG, Rambaut A, et al. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* 2006; **444**(7121): 836-7.
29. Rosenthal E. HIV injustice in Libya--scapegoating foreign medical professionals. *New England Journal of Medicine* 2006; **355**(24): 2505-8. doi:10.1056/NEJMp068241.
30. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 2011; **364**(1): 33-42. doi:10.1056/NEJMoa1012928.
31. Hayward AC, Fragaszy EB, Bermingham A, et al. Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. *The Lancet Respiratory Medicine*; **2**(6): 445-54. doi:10.1016/S2213-2600(14)70034-7.
32. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotie sequences. *Journal of Molecular Evolution* 1980; **16**(2): 111-20.
33. Yang Z. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution* 1996; **42**(2): 294-307.
34. Holmes P. Molecular Evolution: a phylogenetic approach. Oxford: Blackwell Publishing; 1998.
35. Jukes T, Cantor C. Evolution of protein molecules. New York: Academic Press; 1969.
36. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 1981; **17**(6): 368-76.
37. Haseagwa M, Kishino H, al. e. Dating of the human-ape splitting by molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 1985; **22**(2): 160-74.
38. Lanave C, Preparata G, al. e. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 1984; **20**(1): 86-93.
39. Brown AE. Using a phylogenetic approach that combines laboratory and clinical data to enhance understanding of HIV transmission events among men who have sex with men. London: University College London; 2009.
40. Tamura K, Stecher G, Peterson D,et al. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* 2013; **30**(12): 2725-9. doi:10.1093/molbev/mst197.
41. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 2014; **30**(22): 3276-8. doi:10.1093/bioinformatics/btu531.

42. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution* 2010; **27**(2): 221-4. doi:10.1093/molbev/msp259.
43. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; **23**(21): 2947-8. doi:10.1093/bioinformatics/btm404.
44. Vandamme A. Basic concets of molecular evolution. Cambridge, UK: Cambridge University Press; 2009.
45. Efron B, Halloran E, et al. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences* 1996; **93**(14): 7085-90.
46. Baldauf SL. Phylogeny for the faint of heart: a tutorial. *Trends in Genetics* 2003; **19**(6): 345-51. doi:10.1016/s0168-9525(03)00112-4.
47. Riley LW. Molecular epidemiology of infectious disease: principles and practices. Washington, DC: ASM Press; 2004.
48. Bioinformatics SIo. Mutation rates of different types of organisms. http://viralzone.expasy.org/all_by_species/4136.html (accessed 31 July 2016).
49. Palmer S, Vuitton D, Gonzales MJ,et al.. Reverse transcriptase and protease sequence evolution in two HIV-1-infected couples. *Journal of Acquired Immune Deficiency Syndrome* 2002; **31**(3): 285-90.
50. Huè S, Clewley JP, Cane PA, Pillay D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 2004; **18**(5): 719-28.
51. Huè S, Clewley JP, Cane PA, Pillay D. Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. *AIDS* 2005; **19**(4): 449-50.
52. Grabowski MK, Redd AD. Molecular tools for studying HIV transmission in sexual networks. *Current Opinions in HIV and AIDS* 2014; **9**(2): 126-33. doi:10.1097/COH.0000000000000040.
53. Wensing AM, Calvez V, Gunthard HF, et al. 2014 Update of the drug resistance mutations in HIV-1. *Topics in Antiviral Medicine* 2014; **22**(3): 642-50.
54. Frost SD, Volz EM. Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society London B Biological Sciences* 2013; **368**: doi:20120208.
55. Robinson K, Ryson N, Cohen T, al. e. How the dynamics and structure of sexual contact networks shape pathogen phylogenies. *PLoS Computational Biology* 2013; e1003105: doi:e1003105.
56. Sackett DL. Bias in analytic research. *Journal of Chronic Diseases* 1979; **32**: 51-63.
57. Rothman KJ, Greenland S, Lash TL. Validity in epidemiologic studies. Philadelphia: Lippincott Williams & Wilkins; 2008.
58. Gill ON, Adler MW, Day NE. Monitoring the prevalence of HIV. *British Medical Journal* 1989; **299**: 1295-8.
59. UNAIDS. Global HIV/AIDS response: epidemic update and health sector progress towards universal access. Geneva, 2011.
60. Brenner B, Wainberg MA, Roger M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. *AIDS* 2013; **27**: 1045-57.
61. Brenner BG, Roger M, Routy JP, et al. High rates of forward transmission events after acute/early HIV-1 infection. *The Journal of Infectious Diseases* 2007; **195**(7): 951-9. doi:10.1086/512088.
62. Wiens JJ, Servedio MR. Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. *Systems Biology (Stevenage)* 1998; **47**(2): 228-53.
63. Fiscus SA, Pilcher CD, Miller WC, et al. Rapid, real-time detection of acute HIV infection in patients in Africa. *The Journal of infectious diseases* 2007; **195**(3): 416-24. doi:10.1086/510755.

64. Pozniak A, Gazzard B, Anderson J, et al. British HIV Association (BHIVA) guidelines for the treatment of HIV-infected adults with antiretroviral therapy. *HIV medicine* 2003; **4 Suppl 1**: 1-41.
65. Huè S, Brown AE, Ragonnet-Cronin M, et al. Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS* 2014; **28**(13): 1967-75. doi:10.1097/QAD.0000000000000383.
66. Volz EM, Frost SDW. Inferring the Source of Transmission with Phylogenetic Data. *PLOS Computational Biology* 2013; **9**(12): e1003397. doi:10.1371/journal.pcbi.1003397.
67. Tanser F, de Oliveira T, Maheu-Giroux M, T. B. Concentrated HIV subepidemics in generalized epidemic settings. *Current opinion in HIV and AIDS* 2014; **9**(2): 115-25.
68. Karim AQ, Kharsany AB, Frohlich JA, et al. Stabilizing HIV prevalence masks high HIV incidence rates amongst rural and urban women in KwaZulu-Natal, South Africa. *International Journal of Epidemiology* 2011; **40**(4): 922-30.
69. Anderson SJ, Cherutich P, Kilonzo N, et al. Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *The Lancet* 2014; **384**(9939): 249-56.
70. Boily MC, Pickles M, Lowndes CM, et al. Positive impact of a large-scale HIV prevention program among female sex workers and clients in Karnataka state, India. *AIDS* 2013; **27**(9): 1449-60.
71. UNAIDS. Joint United Nations Programme on HIV/AIDS. UNAIDS World AIDS Day Report. Geneva, 2012.
72. Shisana O, Rehle T, Simbayi LC,et al. South African national HIV prevalence, incidence and behaviour survey, 2012. Cape Town: Human Sciences Research Council; 2014.
73. UNAIDS. 90-90-90: An ambitious treatment target to help end the AIDS epidemic. Geneva, 2014. http://www.unaids.org/en/resources/documents/2017/90-90-90. (Accessed 3 February 2017)
74. Sun G, Rossi JJ. MicroRNAs and their potential involvement in HIV infection. *Trends in pharmacological sciences* 2011; **32**(11): 675-81. doi:10.1016/j.tips.2011.07.003.
75. Clavel F, Guetard D, Brun-Vezinet F, et al. Isolation of a new human retrovirus from West African patients with AIDS. *Science* 1986; **233**(4761): 343-6.
76. Hemelaar J. The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine* 2012; **18**(3): 182-92. doi:10.1016/j.molmed.2011.12.001.
77. Robertson DL, Anderson JP, Bradac JA, et al. HIV-1 nomenclature proposal. *Science* 2000; **288**(5463): 55-6.
78. Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications. *ISRN Microbiology* 2013; **2013**: 481314. doi:10.1155/2013/481314.
79. Plantier JC, Leoz M, Dickerson JE, et al. A new human immunodeficiency virus derived from gorillas. *Nature Medicine* 2009; **15**(8): 871-2. doi:10.1038/nm.2016.
80. Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet Infectious Diseases* 2011; **11**(1): 45-56. doi:https://doi.org/10.1016/S1473-3099(10)70186-9.
81. Avert. HIV strains and types. 2017. https://www.avert.org/professionals/hiv-science/types-strains (Accessed 14 October 2017).
82. US Department of Health and Human Services. HIV lifecycle. https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/73/the-hiv-life-cycle (Accessed 26 June 2016).
83. An P, Winkler CA. Host genes associated with HIV/AIDS: advances in gene discovery. *Trends in Genetics* 2010; **26**(3): 119-31. doi:10.1016/j.tig.2010.01.002.
84. WHO. WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children. Geneva, 2007.
85. Castilla J, Del Romero J, et al. Effectiveness of highly active antiretrovirral therapy in reducing heterosexual tranmission of HIV. *Journal of Acquired Immune Deficiency Syndrome* 2005; **40**(1): 96-101.

86. Simon V, Ho DD, Abdool Karim Q. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *The Lancet* 2006; **368**(9534): 489-504. doi:10.1016/s0140-6736(06)69157-5.
87. WHO. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV. Geneva, 2015. http://www.who.int/hiv/pub/guidelines/earlyrelease-arv/en/.(Accessed 3 April 2016)
88. Yerly S, Kaiser L, Race E,et al. Transmission of antiretroviral-drug-resistant HIV-1 variants. *The Lancet* 1999; **354**(9180): 729-33. doi:10.1016/s0140-6736(98)12262-6.
89. Pao D, Fisher M, Huè S, et al. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS* 2005; **19**(1): 85-90.
90. Brown AE, Gifford RJ, Clewley JP, et al. Phylogenetic Reconstruction of Transmission Events from Individuals with Acute HIV Infection: Toward More-Rigorous Epidemiological Definitions. *The Journal of Infectious Diseases* 2009; **199**(3): 427-31. doi:10.1086/596049.
91. de Oliveira T, Kharsany AB, Graf T, et al. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *The Lancet HIV* 2017; **4**(1): e41-e50. doi:10.1016/s2352-3018(16)30186-2.
92. Grabowski MK, Lessler J, Redd AD, et al. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Medicine* 2014; **11**(3): e1001610. doi:10.1371/journal.pmed.1001610.
93. Volz EM, Ionides E, Romero-Severson EO,et al. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Medicine* 2013; **10**(12): e1001568; discussion e. doi:10.1371/journal.pmed.1001568.
94. Gray RR, Tatem AJ, Lamers S, et al. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* 2009; **23**(14): F9-f17. doi:10.1097/QAD.0b013e32832faf61.
95. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* 2013; **110**(1): 228-33. doi:10.1073/pnas.1207965110.
96. Faria NR, Azevedo Rdo S, Kraemer MU, et al. Zika virus in the Americas: Early epidemiological and genetic findings. *Science* 2016; **352**(6283): 345-9. doi:10.1126/science.aaf5036.
97. Auerbach DM, Darrow WW, Jaffe HW, Curran JW. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. *The American Journal of Medicine* 1984; **76**(3): 487-92.
98. Worobey M, Watts TD, McKay RA, et al. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 2016; **539**(7627): 98-101. doi:10.1038/nature19827 http://www.nature.com/nature/journal/v539/n7627/abs/nature19827.html#supplementary-information.(Accessed 16 February 2017)
99. WHO. Ebola virus disease: Fact sheet. http://www.who.int/mediacentre/factsheets/fs103/en/ (accessed 17th December 2014).
100. WHO. Ebola haemorrhagic fever in Zaire, 1976. Report of an international commission. *Bulletin of the World Health Organization* 1978; **56**(2): 271-93.
101. Dowell SF, Mukunu R, Ksiazek TG, Khan AS, Rollin PE, Peters CJ. Transmission of Ebola hemorrhagic fever: a study of risk factors in family members, Kikwit, Democratic Republic of the Congo, 1995. *The Journal of Infectious Diseases* 1999; **179 Suppl 1**: S87-91. doi:10.1086/514284.
102. CDC. About Ebola Virus Disease. 2016. https://www.cdc.gov/vhf/ebola/about.html (Accessed 11 August 2016).

103. Kuhn JH, Andersen KG, Baize S, et al. Nomenclature- and database-compatible names for the two Ebola virus variants that emerged in Guinea and the Democratic Republic of the Congo in 2014. *Viruses* 2014; **6**(11): 4760-99. doi:10.3390/v6114760.
104. Dudas G, Rambaut A. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus Outbreak. *PLOS Current Outbreaks* 2014; doi:10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.
105. Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 2014; **345**(6202): 1369-72. doi:10.1126/science.1259657.
106. Carroll MW, Matthews DA, Hiscox JA, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature* 2015; **524**(7563): 97-101. doi:10.1038/nature14594.
107. Messaoudi I, Amarasinghe GK, Basler CF. Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nature Reviews Microbiology* 2015; **13**(11): 663-76. doi:10.1038/nrmicro3524.
108. CDC. Outbreaks Chronology: Ebola Virus Disease. 2016. https://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html (Accessed 11 August 2016).
109. WHO. Ebola response roadmap: Situation response Update (18/02/2015). http://apps.who.int/ebola/en/ebola-situation-report/situation-reports/ebola-situation-report-18-february-2015 (Accessed 19 February 2015).
110. Ebola haemorrhagic fever in Sudan, 1976. Report of a WHO/International Study Team. *Bulletin of the World Health Organization* 1978; **56**(2): 247-70.
111. Jahrling PB, Geisbert TW, Dalgard DW, et al. Preliminary report: isolation of Ebola virus from monkeys imported to USA. *The Lancet* 1990; **335**(8688): 502-5.
112. Le Guenno B, Formenty P, Wyers M, et al. Isolation and partial characterisation of a new strain of Ebola virus. The *Lancet* 1995; **345**(8960): 1271-4.
113. WHO. Democratic Republic of Congo: "classic" Ebola in a country experiencing its seventh outbreak. 2014. http://www.who.int/csr/disease/ebola/ebola-6-months/drc/en/ (Accessed 12 August 2016).
114. Olivero J, Fa JE, Real R, et al. Recent loss of closed forests is associated with Ebola virus disease outbreaks. *Scientific Reports* 2017; **7**(1): 14291. doi:10.1038/s41598-017-14727-9.
115. Leroy EM, Rouquet P, Formenty P, et al. Multiple Ebola virus transmission events and rapid decline of central African wildlife. *Science* 2004; **303**(5656): 387-90. doi:10.1126/science.1092528.
116. Swanepoel R, Leman PA, Burt FJ, et al. Experimental inoculation of plants and animals with Ebola virus. *Emerging Infectious Diseases* 1996; **2**(4): 321-5. doi:10.3201/eid0204.960407.
117. Leroy EM, Kumulungui B, Pourrut X, et al. Fruit bats as reservoirs of Ebola virus. *Nature* 2005; **438**(7068): 575-6. doi:10.1038/438575a.
118. Leendertz SA, Gogarten JF, Dux A et al.. Assessing the Evidence Supporting Fruit Bats as the Primary Reservoirs for Ebola Viruses. *EcoHealth* 2016; **13**(1): 18-25. doi:10.1007/s10393-015-1053-0.
119. WHO Ebola Response Team. Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections. *New England Journal of Medicine* 2014; **371**(16): 1481-95. doi:doi:10.1056/NEJMoa1411100.
120. WHO. Ebola R&D landscape of clinical candidates and trials. 2015. https://www.google.co.uk/search?q=WHO+Ebola+R%26D+landscape+of+clinical+trails+2013&ie=utf-8&oe=utf-8&client=firefox-b-ab&gfe_rd=cr&ei=xiL9V7SWEOLv8AfUqJGwDA (ccessed 10 October 2016).
121. CDC. Ebola Virus Disease. 2016. https://www.cdc.gov/vhf/ebola/ (Accessed 13 April 2017).

122. Georges-Courbot MC, Sanchez A, Lu CY, et al. Isolation and phylogenetic characterization of Ebola viruses causing different outbreaks in Gabon. *Emerging Infectious Diseases* 1997; **3**(1): 59-62. doi:10.3201/eid0301.970107.
123. Suzuki Y, Gojobori T. The origin and evolution of Ebola and Marburg viruses. *Molecular Biology and Evolution* 1997; **14**(8): 800-6.
124. Leroy EM, Epelboin A, Mondonge V, et al. Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector Borne Zoonotic Diseases* 2009; **9**(6): 723-8. doi:10.1089/vbz.2008.0167.
125. McGrath N, Eaton JW, Newell M-L, Hosegood V. Migration, sexual behaviour, and HIV risk: a general population cohort in rural South Africa. *The Lancet HIV*; **2**(6): e252-e9. doi:10.1016/S2352-3018(15)00045-4.
126. Tanser F, Hosegood V, Bärnighausen T, et al. Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *International Journal of Epidemiology* 2008; **37**(5): 956-62. doi:10.1093/ije/dym211.
127. Hosegood V, Benzler J, Solarsh GC. Population mobility and household dynamics in rural South Africa: implications for demographic and health research. *Southern African Journal of Demography* 2005; **10**(1/2): 43-68. http://www.jstor.org/stable/20853278.
128. Welz T, Hosegood V, Jaffar S, Batzing-Feigenbaum J, Herbst K, Newell ML. Continued very high prevalence of HIV infection in rural KwaZulu-Natal, South Africa: a population-based longitudinal study. *AIDS* 2007; **21**(11): 1467-72. doi:10.1097/QAD.0b013e3280ef6af2.
129. South African National Department of Health. National Antenatal HIV Prevalence Survey (2013). Pretoria, 2016. https://africahealthnews.com/antenatal-hiv-prevalence-survey-south-africa-published/. https://africahealthnews.com/antenatal-hiv-prevalence-survey-south-africa-published/
130. Tanser F, Barnighausen T, Cooke GS, Newell ML. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. *International Journal of Epidemiology* 2009; **38**(4): 1008-16. doi:10.1093/ije/dyp148.
131. Tanser F, Barnighausen T, Sartorius B. Identifying 'corridors of HIV transmission' in a severely affected rural South AFrican population: A case for a shift toward targeted prevention strategies. *In preparataion* 2017.
132. Houlihan CF, Bland RM, Mutevedzi PC, et al. Cohort Profile: Hlabisa HIV Treatment and Care Programme. *International Journal of Epidemiology* 2011; **40**(2): 318-26. doi:10.1093/ije/dyp402.
133. Pillay D. Personal communication. 2017.
134. Camlin CS, Hosegood V, Newell ML, McGrath N, Barnighausen T, Snow RC. Gender, migration and HIV in rural KwaZulu-Natal, South Africa. *PLoS One* 2010; **5**(7): e11539. doi:10.1371/journal.pone.0011539.
135. Muhwava W, Hosegood V, Nyirenda M, Herbst A, Newell M-L. Levels and determinants of Population Movements and Migration in rural KwaZulu Natal, South Africa; 2010.
136. Tanser F, Barnighausen T, Vandormael A, Dobra A. HIV treatment cascade in migrants and mobile populations. *Current Opinion in HIV and AIDS* 2015; **10**(6): 430-8. doi:10.1097/COH.0000000000000192.
137. Lurie MN, Williams BG. Migration and health in Southern Africa: 100 years and still circulating. *Health Psychology and Behavioral Medicine* 2014; **2**(1): 34-40. doi:10.1080/21642850.2013.866898.
138. Hosegood V, McGrath N, Moultrie T. Dispensing with marriage: Marital and partnership trends in rural KwaZulu-Natal, South Africa 2000-2006. *Demographic Research* 2009; **20**(13): 279-312.
139. Tanser F, Barnighausen T, Grapsa E, Zaidi J, Newell ML. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science* 2013; **339**(6122): 966-71. doi:10.1126/science.1228160.

140. Africa Centre. Africa Center Demographic Information System (ACDIS) Fieldwork Training Manual. Mtubatuba; 2008.
141. Filmer D, Pritchett LH. Estimating Wealth Effects without Expenditure Data-or Tears: An appplication to educational enrollments in states of India. *Demography* 2001; **38**(1): 115-32.
142. Rutstein SO, Johnson K. The DHS Wealth Index: DHS Comparative Reports No.6. Calverton, Maryland, USA: ORC Macro, 2004.
143. Bor J, Barnighausen T, Newell C, Tanser F, Newell ML. Social exposure to an antiretroviral treatment programme in rural KwaZulu-Natal. *Tropical Medicine and International Health* 2011; **16**(8): 988-94. doi:10.1111/j.1365-3156.2011.02795.x.
144. Los Alamos HIV database. http://www.hiv.lanl.gov/ (accessed 7 February 2015).
145. Manasa J, Danaviah S, Pillay S, et al. An affordable HIV-1 drug resistance monitoring method for resource limited settings.*Journal of Visualized Experiments* 2014; 10.3791/51242.: doi:10.3791/51242.
146. Manasa J, Danaviah S, Lessells R, et al. Increasing HIV-1 drug resistance between 2010 and 2012 in adults participating in population-based HIV surveillance in rural KwaZulu-Natal, South Africa *AIDS Research and Human Retroviruses* 2016; **32**(8): 763-9. doi:10.1089/AID.2015.0225.
147. Kearse M, Moir R, Wilson A, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012; **28**(12): 1647-9. doi:10.1093/bioinformatics/bts199.
148. HIV-1 Quality Assessment Tool. http://bioafrica.net/tools/pppweb.html (accessed 24 August 2017).
149. Gifford RJ, Liu TF, Rhee S-Y, et al. The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance. *Bioinformatics* 2009; **25**(9): 1197-8. doi:10.1093/bioinformatics/btp134.
150. Pineda-Pena AC, Faria NR, Imbrechts S, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infection Genetics and Evolution* 2013; **19**: 337-48. doi:10.1016/j.meegid.2013.04.032.
151. Africa Centre. Personal communication with AC. 2016.
152. Larmarange J, Mossong J, Bärnighausen T, Newell ML. Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa *PLoS One* 2015; doi:http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123345.
153. Baisley C. Linkage to care at the Africa Centre. 2016.
154. Statistics Sierra Leone. Sierra Leone's 2015 Census Provisional Results. Freetown, 2016. https://sl.one.un.org/2016/04/15/sierra-leones-2015-census-provisional-results-launched/. (Accessed 4 February 2016)
155. Central_Intelligence_Agency_US.Gov. The World Factbook - Sierra Leone. 2017. https://www.cia.gov/library/publications/the-world-factbook/geos/sl.html, (Accessed 6 Sptember 2017)
156. Arias A, Watson SJ, Asogun D, et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evolution* 2016; 10.1093/ve/vew016: doi:10.1093/ve/vew016.
157. Trombley AR, Wachter L, Garrison J, et al. Comprehensive Panel of Real-Time TaqMan™ Polymerase Chain Reaction Assays for Detection and Absolute Quantification of Filoviruses, Arenaviruses, and New World Hantaviruses. *The American Journal of Tropical Medicine and Hygiene* 2010; **82**(5): 954-60. doi:10.4269/ajtmh.2010.09-0636.
158. WHO. Ebola situation reports: archive. 2016. http://www.who.int/csr/disease/ebola/situation-reports/archive/en/ (Accessed 9 October 2016).

159. WHO. Case definition recommendations for Ebola or Marburg virus diseases. August 2014 2014. http://www.who.int/csr/resources/publications/ebola/case-definition/en/.
160. Coltart CEM, Lindsey B, Ghinai I,et al. The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2017; **372**(1721): doi:10.1098/rstb.2016.0297.
161. WHO. Ebola maps. 2016. http://www.who.int/csr/disease/ebola/maps/en/, http://www.who.int/csr/disease/ebola/maps/en/. (Accessed 6 January 2017)
162. WHO. Global Health Observatory data repository. 2015. http://apps.who.int/gho/data/node.main.A1444?lang=en&showonly=HWF. (Accessed 6 January 2017)
163. WorldBank. Indicators. 2016. http://data.worldbank.org/indicator/, (Accessed 6 January 2017)
164. MSF. Guinea: Ebola epidemic declared. 2014. http://www.msf.org.uk/article/guinea-ebola-epidemic-declared (Accessed 10 October 2016).
165. CDC. Previous Updates: 2014 West Africa Outbreak. 2014. http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/previous-updates.html (Accessed 10 October 2016).
166. WHO. Ebola virus disease in Guinea (Situation as of 25 March 2014). 2014. http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4065-ebola-virus-disease-in-guinea-25-march-2014.html (Accessed 10 October 2016).
167. Baize S, Pannetier D, Oestereich L, et al. Emergence of Zaire Ebola virus disease in Guinea. *New England Journal of Medicine* 2014; **371**(15): 1418-25. doi:10.1056/NEJMoa1404505.
168. Coltart CE, Johnson AM, Whitty CJ. Role of healthcare workers in early epidemic spread of Ebola: policy implications of prophylactic compared to reactive vaccination policy in outbreak prevention and control. *BMC Med* 2015; **13**: 271. doi:10.1186/s12916-015-0477-2.
169. Aljazeera. Guinea residents 'refusing' Ebola treatment. 2014. http://www.aljazeera.com/news/africa/2014/09/guinea-residents-refusing-ebola-treatment-201492751955453636.html (accessed 09 December 2016)
170. BBC. Ebola outbreak: Guinea health team killed. 2014. www.bbc.co.uk/news/world-africa-29256443 (Accessed 09 December 2016)
171. Phillip A. Eight dead in attack on Ebola team in Guinea. 'Killed in cold blood'. The Wasington Post. 2014. https://www.washingtonpost.com/news/to-your-health/wp/2014/09/18/missing-health-workers-in-guinea-were-educating-villagers-about-ebola-when-they-were-attacked/?utm_term=.bdfc66109340 (Accessed 19 October 2016)
172. The Advisory Board Company. Health workers killed in Guinea for distributing information about Ebola. 2014. https://www.advisory.com/daily-briefing/2014/09/19/health-workers-killed-in-guinea-for-distributing-information-about-ebola (Accessed 19 October 2016)
173. Tech Times. Ebola update: Guinea president declares virus outbreak as national health emergency. 2014. http://www.techtimes.com/articles/13018/20140814/ebola-update-guinea-president-declares-virus-outbreak-as-national-health-emergency.htm (Accessed 10 October 2016).
174. ACTUU224. Le gouverneur interdit tout les spectacles prévus pour la Tabaski, M. Thug réagit. 2014. http://www.actu224.com/le-gouverneur-interdit-tout-les-spectacles-prevus-pour-la-tabaski-m-thug-reagit/ (Accessed 10 October 2016).
175. Keïta F. Fête de Tabaski: Toutes les manifestations culturelles interdites à Conakry... Africa Guinee. 2014. http://www.africaguinee.com/articles/2014/10/03/fete-de-tabaski-toutes-les-manifestations-culturelles-interdites-conakry (Accessed 10 October 2016).

176. BBC. Ebola crisis: Guinea begins compensation payments. 2014. http://www.bbc.co.uk/news/world-africa-297450269 (Accessed 10 October 2016).
177. The Guardian. Bandits in Guinea steal blood samples believed to be infected with Ebola 2014. 2016. https://www.theguardian.com/world/2014/nov/21/bandits-guinea-steal-blood-samples-possibly-infected-with-ebola (Accessed 7 October 2017).
178. Humanitarian Data Exchange. Guinea ETCs. 2016. https://data.humdata.org/dataset/ebola-treatment-centers/resource/007c38f0-2efe-4f4a-8b47-f7eb5a4d20eb (Accessed 9 October 2016).
179. WHO. New Ebola cases confirmed in Guinea as WHO warns of more possible flare ups. 2016. http://www.who.int/csr/disease/ebola/new-ebola-cases-confirmed-guinea/en/ (Accessed 10 October 2016).
180. WHO. Criteria for declaring the end of the Ebola outbreak in Guinea, Liberia or Sierra Leone. 2015. http://www.who.int/csr/disease/ebola/declaration-ebola-end/en/, http://www.who.int/csr/disease/ebola/declaration-ebola-end/en/.(Accessed 14 October 2016).
181. WHO. Ebola virus disease in Liberia. 2014. http://www.who.int/csr/don/2014_03_30_ebola_lbr/en/ (Accessed 9 October 2016)
182. CDC. Previous Case Counts. 2016. http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/previous-case-counts.html (Accessed 9 October 2016)
183. Ladner JT, Wiley MR, Mate S, et al. Evolution and Spread of Ebola Virus in Liberia, 2014-2015. *Cell Host and Microbe* 2015; **18**(6): 659-69. doi:10.1016/j.chom.2015.11.008.
184. Reuters. WHO says Guinea Ebola outbreak mall as MSF slams international response. . 2014. http://www.reuters.com/article/us-guinea-ebola-idUSBREA301X120140401 (Accessed 9 October 2016).
185. MSF. Ebola: Pushed to the limit and beyond. 2015. http://www.msf.org/en/article/ebola-pushed-limit -and-beyond (Accessed 9 October 2016).
186. BBC. Seven die in Monrovia Ebola outbreak. 2014. http://www.bbc.co.uk/news/world-africa-278883639 (Accessed 9 October 2016).
187. AllAfrica. Liberia: Ebola kills doctor at Redemption Hospital. 2014. http://allafrica.com/stories/201407021024.html (Accessed 9 October 2016)
188. BBC. Ebola outbreak: Liberia shuts most border points. 2014. http://www.bbc.co.uk/news/world-africa-28522824 (Accessed 9 October 2016)
189. NewYorkTimes. In Liberia, Home Deaths Spread Circle of Ebola Contagion. 2014. http://www.nytimes.com/2014/09/25/world/africa/liberia-ebola-victims-treatment-center-cdc.html (Accessed 9 October 2016).
190. WHO. Liberia: Ebola clinic fills up within hours of opening. 2014. http://www.who.int/features/2014/liberia-ebola-clinic/en/ (Accessed 9 October 2016)
191. US Government. The U.S Government Response to the Ebola Outbreak. 2014. http://m.state.gov/md233996.htm (Accessed 9 October 2016)
192. NewYorkTimes. Empty Ebola Clinics in Liberia Are Seen as Misstep in U.S Relief Effort. 2015. http://www.nytimes.com/2015/04/12/world/africa/idle-ebola-clinics-in-liberia-are-seen-as-misstep-in-us-relief-effort.html?_r=0 (Accessed 9 October 2016)
193. NBCNews. Ebola Burial Teams in Sierra Leone Go On Strike Over Hazard Pay. 2014. http://www.nbcnews.com/storyline/ebola-virus-outbreak/ebola-burial-teams-sierra-leone-go-strike-over-hazard-pay-n220871, (Accessed 9 October 2016)
194. News24. Last Ebola-free region of Liberia falls to Virus. 2014. http://www.news24.com/Africa/News/Last-Ebola-free-region-of-Liberia-falls-to-virus-20140822-3 (accessed 9 October 2016).
195. Fallah M, Dahn B, Nyenswah TG, et al. Interrupting Ebola Transmission in Liberia Through Community-Based Initiatives. *Annals of Internal Medicine* 2016; **164**(5): 367-9. doi:10.7326/M15-1464.

196. WHO. Liberia and Guinea step up coordination to stem new cases of Ebola. 2016. http://www.who.int/csr/disease/ebola/liberia-guinea-flareups-update/en/ (Accessed 9 October 2016).
197. NIH. Genetics of the 2014 Ebola Outbreak. 2014. https://www.nih.gov/news-events/nih-research-matters/genetics-2014-ebola-outbreak(Accessed 9 October 2016)
198. Sack K, Fink, S., Belluck, P., Nossiter, A. How Ebola Roared Back. New York Times. 2014. http://www.nytimes.com/2014/12/30/health/how-ebola-roared-back.html (Accessed 2 October 2016).
199. WHO. Ebola in Sierra Leone: A slow start to an outbreak that eventually outpaced all others. 2015. http://www.who.int/csr/disease/ebola/one-year-report/sierra-leone/en/ (accessed 2 October 2016).
200. Schoepp RJ, Rossi CA, Khan SH, Goba A, Fair JN. Undiagnosed acute viral febrile illnesses, Sierra Leone. *Emerging Infectious Diseases* 2014; **20**(7): 1176-82. doi:10.3201/eid2007.131265.
201. Washington MMM. Effectiveness of Ebola Treatment Units and Community Care Centers - Liberia, September 23-October 31, 2014. *Morbidity and Mortality Weekly Report* 2015; **64**(3): 67-9. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6403a6.htm.
202. Nature. 365 days: Nature's 10. *Nature* 2014; **516**(7531): 311-9.
203. UK Government. UK Action Plan to Defeat Ebola in Sierra Leone. 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/357703/UK_action_plan_to_defeat_Ebola_in_Sierra_Leone_-_background_paper.pdf2016). (Accessed 9 October 2016).
204. Park DJ, Dudas, G., Wohl, S., et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* 2015; **161**(7): 1516-26.
205. Lu HJ, Qian, J., Kargbo, D.,et al.. Ebola Virus Outbreak Investigation, Sierra Leone, September 28 - November 11, 2104. *Emering Infectious Dieases* 2015; **21**(11).
206. News24. Ebola ravages health care in Freetown. 2014. http://www.news24.com/Africa/News/Ebola-ravages-health-care-in-Freetown-20140906, (Accessed 4 October 2016).
207. WHO. Ebola Response Roadmap Situation Report. 2014, http://apps.who.int/iris/bitstream/10665/144032/1/roadmapsitrep_19Nov14_eng.pdf (Accessed June 14 2016).
208. UNICEF. Sierra Leone launches three-day, door-to-door Ebola prevention campaign. 2014, http://www.unicef.org/wcaro/english/media_8582.html. (Accessed 14 June 2016).
209. AssociatedPress. Sierra Leone cancels all soccer matches over Ebola Outbreak. August 5, 2014 2014. http://www.nydailynews.com/sports/soccer/sierra-leone-cancels-soccer-matches-ebola-outbreak-article-1.1892588, (Accessed 14 June 2016).
210. O'Carroll L. Ebola Epidemic: Sierra Leone quarantines a million people. The Guardian. 2014 25/09/2014. https://www.theguardian.com/world/2014/sep/25/ebola-epidemic-sierra-leone-quarantine-un-united-nations (Accessed 12 June 2016).
211. Broadcast Message to the Nation By His Excellency Dr. Ernest Bai Koroma On the March to Zero Ebola Cases March 31, 2015 *Nationa Ebola Response Centre* 2015, http://nerc.sl/?q=president-koroma&page=1. (Accessed 10 June 2016).
212. Maxmen A. In Sierra Leone, Quarantines without food threathen Ebola response. 2015. http://america.aljazeera.com/articles/2015/2/19/in-sierra-leone-quarantined-ebola-survivors.html. (Accessed 12 June 2016).
213. CBCNews. Ebola outbreak: Why Liberia's quarantine will fail. 2014. http://www.cbc.ca/news/world/ebola-outbreak-why-liberia-s-quarantine-in-west-point-slum-will-fail-1.27442922016). (Accessed 12 June 2016).

214. Drazen JM, Kanapathipillai R, Campion EW, et al. Ebola and quarantine. *New England Journal of Medicine* 2014; **371**(21): 2029-30. doi:10.1056/NEJMe1413139.
215. AssociatedPress. Ebola cases appear in last untouched district in Sierra Leone. October 16, 2014 2014. http://www.foxnews.com/world/2014/10/16/ebola-cases-appear-in-last-untouched-district-in-sierra-leone.html. (Accessed 4 July 2016).
216. Shuchman M. Sierra Leone doctors call for better Ebola care for colleagues. *The Lancet* 2014; **384**(9961): e67. doi:10.1016/s0140-6736(14)62388-6.
217. Richards P, Amara J, Ferme MC, et al. Social pathways for Ebola virus disease in rural Sierra Leone, and some implications for containment. *PLoS Neglected Tropical Diseases* 2015; **9**(4): e0003567. doi:10.1371/journal.pntd.0003567.
218. Levine R, Ghiselli M, Conteh A, et al. Notes from the Field: Development of a Contact Tracing System for Ebola Virus Disease - Kambia District, Sierra Leone, January-February 2015. *Morbidity and mortality weekly report* 2016; **65**(15): 402. doi:10.15585/mmwr.mm6515a4.
219. Koroma E. President Koroma Warns Port Loko, Kambia to Stop Secret and Unsafe Burials. 2015. http://statehouse.gov.sl/index.php/contact/1184-president-koroma-warns-port-loko-kambia-to-stop-secret-and-unsafe-burials. (Accessed 4 July 2016).
220. Awoko. NERC launches Operation Northern Push. 2015. http://awoko.org/2015/06/18/sierra-leone-news-nerc-launches-operation-northern-push/**.** (Accessed 4 July 2016).
221. WHO. Stopping Ebola: It takes collaboration to care for a village. 2015. http://www.who.int/features/2015/stopping-ebola-in-kambia/en/. (Accessed 4 July 2016).
222. CDC. Sierra Leone Trial to Introduce a Vaccine against Ebola (STRIVE) Q&A. 2016. https://www.cdc.gov/vhf/ebola/strive/qa.html, (Accessed 6 July 2016).
223. O'Carroll L, Fofana U. WHO officially declares Sierra Leone Ebola-free. 2015. https://www.theguardian.com/world/2015/nov/07/world-health-organisation-sierra-leone-ebola-free, (Accessed 9 September 2016).
224. WHO. WHO statement on end of Ebola Flare-up in Sierra Leone. 2016.
225. Fasina FO, Shittu A, Lazarus D, et al. Transmission dynamics and control of Ebola virus disease outbreak in Nigeria, July to September 2014. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European Communicable Disease Bulletin* 2014; **19**(40): 209-20.
226. WHO. Mali: Details of the additional cases of Ebola virus disease. http://www.who.int/mediacentre/news/ebola/20-november-2014-mali/en/ (Accessed 30 January 2015).
227. Reuters. Doctor who treated source of second Mali Ebola outbreak dies. http://www.reuters.com/article/2014/11/20/health-ebola-mali-idUSL6N0TA62R20141120 (Accessed 30 January 2015).
228. Reuters. Mali says has no remaining Ebola cases as last patient recovers. http://www.reuters.com/article/2014/12/11/us-health-ebola-mali-idUSKBN0JP2HG20141211 (Accessed 30th January 2015).
229. Simon-Loriere E, Faye O, Faye O, et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* 2015; **524**(7563): 102-4. doi:10.1038/nature14612.
230. CDC. Cases of Ebola diagnosed in the United States. http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/united-states-imported-case.html (Accessed 12 January 2015).
231. BBC News. Ebola outbreak: nurse infected in Spain. http://www.bbc.co.uk/news/world-europe-29514920 (accessed 30 January 2015).
232. WHO. Ebola virus disease – United Kingdom. http://www.who.int/csr/don/30-december-2014-ebola/en/ (Accessed 30 January 2015).

233. WHO. The outbreak of Ebola virus disease in Senegal is over. http://www.who.int/mediacentre/news/ebola/17-october-2014/en/ (Accessed 30 January 2015).
234. WHO. Ebola virus disease - Italy. 2015. http://www.who.int/csr/don/13-may-2015-ebola/en/ (Accessed 10 October 2016).
235. Piot P. Ebola's perfect storm. *Science* 2014; **345**(6202): 1221. doi:10.1126/science.1260695.
236. Chowell G, Nishiura H. Characterizing the transmission dynamics and control of ebola virus disease. *PLoS Biology* 2015; **13**(1): e1002057. doi:10.1371/journal.pbio.1002057.
237. WHO. Factors that contributed to undetected spread of the Ebola virus and impeded rapid containment. 2015. http://www.who.int/csr/disease/ebola/one-year-report/factors/en/. (Accessed 10 October 2016).
238. Nations U. World Urbanization Prospects, 2014. https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Highlights.pdf. (accessed ST/ESA/SER.A/352, https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Highlights.pdf (Accessed 10 October 2016).
239. Neiderud CJ. How urbanization affects the epidemiology of emerging infectious diseases. *Infection Ecology & Epidemiology* 2015; **5**: 27060. doi:10.3402/iee.v5.27060.
240. Gomes MF, Pastore YPA, Rossi L, et al. Assessing the international spreading risk associated with the 2014 West African ebola outbreak. *PLoS currents* 2014; **6**: doi:10.1371/currents.outbreaks.cd818f63d40e24aef769dda7df9e0da5.
241. Bogoch, II, Creatore MI, Cetron MS, et al. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. *The Lancet* 2015; **385**(9962): 29-35. doi:10.1016/s0140-6736(14)61828-6.
242. Heymann DL, Chen L, Takemi K, et al. Global health security: the wider lessons from the west African Ebola virus disease epidemic. *The Lancet* 2015; **385**(9980): 1884-901. doi:10.1016/s0140-6736(15)60858-3.
243. Richards P, Amara, J., Ferme, M.,et al.. Social Pathways for Ebola Virus Disease in Rural Sierra Leone and some Implications for Containment. *PLOS Neglected Tropical Diseases* 2014. Doi.org/10.1371/journal.pntd.0003567.
244. WHO. Sierra Leone: a traditional healer and a funeral. 2014. http://www.who.int/csr/disease/ebola/ebola-6-months/sierra-leone/en/, (Accessed 10 October 2016).
245. Pandey A, Atkins KE, Medlock J, et al. Strategies for containing Ebola in West Africa. *Science* 2014; **346**(6212): 991-5. doi:10.1126/science.1260612.
246. Nielsen CF, Kidd S, Sillah AR, et al.. Improving burial practices and cemetery management during an Ebola virus disease epidemic - Sierra Leone, 2014. *Morbidity and Mortality Weekly Report* 2015; **64**(1): 20-7.
247. Fang LQ, Yang Y, Jiang JF, et al. Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. *Proceedings of the National Academy of Sciences* 2016; **113**(16): 4488-93. doi:10.1073/pnas.1518587113.
248. Frieden TR, Damon I, Bell BP,et al. Ebola 2014--new challenges, new global response and responsibility. *New England Journal of Medicine* 2014; **371**(13): 1177-80. doi:10.1056/NEJMp1409903.
249. Kucharski AJ, Camacho A, Flasche S,et al.. Measuring the impact of Ebola control measures in Sierra Leone. *Proceedings of the National Academy of Sciences* 2015; **112**(46): 14366-71. doi:10.1073/pnas.1508814112.
250. Kickbusch I, Reddy KS. Community matters - why outbreak responses need to integrate health promotion. *Global Health Promotion* 2016; **23**(1): 75-8. doi:10.1177/1757975915606833.
251. Crowe S, Hertz D, Maenner M, et al. A plan for community event-based surveillance to reduce Ebola transmission - Sierra Leone, 2014-2015. *Morbidity and mortality weekly report* 2015; **64**(3): 70-3.

252. Olu OO, Lamunu M, Nanyunja M, et al. Contact Tracing during an Outbreak of Ebola Virus Disease in the Western Area Districts of Sierra Leone: Lessons for Future Ebola Outbreak Response. *Frontiers in Public Health* 2016; **4**: 130. doi:10.3389/fpubh.2016.00130.
253. Fast SM, Mekaru S, Brownstein JS, Postlethwaite TA, Markuzon N. The Role of Social Mobilization in Controlling Ebola Virus in Lofa County, Liberia. *PLoS currents* 2015; **7**: doi:10.1371/currents.outbreaks.c3576278c66b22ab54a25e122fcdbec1.
254. Marais F, Minkler M, Gibson N, et al. A community-engaged infection prevention and control approach to Ebola. *Health Promotion International* 2016; **31**(2): 440-9. doi:10.1093/heapro/dav003.
255. Carrion Martin AI, Derrough T, Honomou P, et al. Social and cultural factors behind community resistance during an Ebola outbreak in a village of the Guinean Forest region, February 2015: a field experience. *International Health* 2016; **8**(3): 227-9. doi:10.1093/inthealth/ihw018.
256. Greiner AL, Angelo KM, McCollum AM, et al. Addressing contact tracing challenges-critical to halting Ebola virus disease transmission. *International Journal of Infectious Diseases* 2015; **41**: 53-5. doi:10.1016/j.ijid.2015.10.025.
257. Rivers CM, Lofgren ET, Marathe M, et al. Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. *PLoS Currents* 2014; **6**: doi:10.1371/currents.outbreaks.4d41fe5d6c05e9df30ddce33c66d084c.
258. Wong V, Cooney D, Bar-Yam Y. Beyond Contact Tracing: Community-Based Early Detection for Ebola Response. *PLoS Currents* 2016; **8**: doi:10.1371/currents.outbreaks.322427f4c3cc2b9c1a5b3395e7d20894.
259. Dixon MG, Taylor MM, Dee J, et al. Contact Tracing Activities during the Ebola Virus Disease Epidemic in Kindia and Faranah, Guinea, 2014. *Emerging Infectious Diseases* 2015; **21**(11): 2022-8. doi:10.3201//eid2111.150684.
260. WHO. Ebola haemorrhagic fever in Sudan, 1976. Report of a WHO/International Study Team. *Bulletin of the World Health Organization* 1978; **56**(2): 247-70. https://www.ncbi.nlm.nih.gov/pubmed/307455.
261. Kilmarx PH, Clarke KR, Dietz PM, et al. Ebola virus disease in health care workers--Sierra Leone, 2014. *Morbidity and mortality weekly report* 2014; **63**(49): 1168-71.
262. WHO. Personal protective equipment in the context of filovirus disease outbreak response. http://apps.who.int/iris/bitstream/10665/137410/1/WHO_EVD_Guidance_PPE_14.1_eng.pdf?ua=1 (Accessed 24 February 2015).
263. Pathmanathan I, O'Connor KA, Adams ML, et al. Rapid assessment of Ebola infection prevention and control needs--six districts, Sierra Leone, October 2014. *Morbidity and Mortality Weekly Report* 2014; **63**(49): 1172-4.
264. Nossiter A. Ebola Help for Sierra Leone Is Nearby, but Delayed on the Docks. The New York Times. 2014 05/10/2014. http://www.nytimes.com/2014/10/06/world/africa/sierra-leone-ebola-medical-supplies-delayed-docks.html (Accessed 24 February 2015).
265. Legrand J, Grais RF, Boelle PY, Valleron AJ, Flahault A. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infectection* 2007; **135**(4): 610-21. doi:10.1017/s0950268806007217.
266. Meltzer MI, Atkins CY, Santibanez S, et al. Estimating the future number of cases in the Ebola epidemic--Liberia and Sierra Leone, 2014-2015. *Morbidity and Mortality Weekly Report supplements* 2014; **63**(3): 1-14.
267. White RA, MacDonald E, de Blasio BF,et al. Projected treatment capacity needs in Sierra Leone. *PLoS currents* 2015; **7**: doi:10.1371/currents.outbreaks.3c3477556808e44cf41d2511b21dc29f.

268. WHO. Ebola Situation Report: WHO, 2015. http://iacld.ir/DL/elm/93/whoebolaresponseroadmapsituationreport7january20152.pdf. (Accessed 5 February 2016).
269. Townsend JP, Skrip LA, Galvani AP. Impact of bed capacity on spatiotemporal shifts in Ebola transmission. *Proceedings of the National Academy of Sciences* 2015; **112**(46): 14125-6. doi:10.1073/pnas.1518484112.
270. Kirsch TD, Moseson H, Massaquoi M, et al. Impact of interventions and the incidence of ebola virus disease in Liberia-implications for future epidemics. *Health Policy and Planning* 2016; 10.1093/heapol/czw113: doi:10.1093/heapol/czw113.
271. Chowell G, Viboud C. Controlling Ebola: key role of Ebola treatment centres. *The Lancet Infectious diseases* 2015; **15**(2): 139-41. doi:10.1016/s1473-3099(14)71086-2.
272. Barbisch D, Koenig KL, Shih FY. Is There a Case for Quarantine? Perspectives from SARS to Ebola. *Disaster Medicine and Public Health Preparedness* 2015; **9**(5): 547-53. doi:10.1017/dmp.2015.38.
273. Breman JG, Heymann DL, Lloyd G, et al. Discovery and Description of Ebola Zaire Virus in 1976 and Relevance to the West African Epidemic During 2013-2016. *The Journal of infectious diseases* 2016; **214**(suppl 3): S93-s101. doi:10.1093/infdis/jiw207.
274. Piot P. The Importance of Social and Behavioral Change Response in the Zika Outbreak: Lessons Learned from Ebola. 2016. http://globalhealth.org/wp-content/uploads/Zika-Call-to-Action-10-03-16-gm-1-002.pdf, (Accessed 10 December 2016).
275. Zhu FC, Wurie AH, Hou LH, et al. Safety and immunogenicity of a recombinant adenovirus type-5 vector-based Ebola vaccine in healthy adults in Sierra Leone: a single-centre, randomised, double-blind, placebo-controlled, phase 2 trial. *The Lancet* 2016; 10.1016/s0140-6736(16)32617-4: doi:10.1016/s0140-6736(16)32617-4.
276. Ewer K, Rampling T, Venkatraman N, et al. A Monovalent Chimpanzee Adenovirus Ebola Vaccine Boosted with MVA. *New England Journal of Medicine* 2016; **374**(17): 1635-46. doi:10.1056/NEJMoa1411627.
277. Henao-Restrepo AM, Longini IM, Egger M, et al. Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial. *The Lancet* 2015; **386**(9996): 857-66. doi:10.1016/s0140-6736(15)61117-5.
278. Tong YG, Shi WF, Liu D, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* 2015; **524**(7563): 93-6. doi:10.1038/nature14490.
279. Sissoko D, Keita M, Diallo B, et al. Ebola Virus Persistence in Breast Milk After No Reported Illness: A Likely Source of Virus Transmission From Mother to Child. *Clinical Infectious Diseases* 2017; **64**(4): 513-6. doi:10.1093/cid/ciw793.
280. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 2017; **544**(7650): 309-15. doi:10.1038/nature22040 http://www.nature.com/nature/journal/v544/n7650/abs/nature22040.html#supplementary-information.
281. Stadler T, Kouyous R, von Wyl V. Estimating the basic reporductive number from viral sequence data. *Molecular Biology Evolution* 2012; **29**: 347-57.
282. Stadler T, Kuhnert D, Rasmussen DA, du Plessis L. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. *PLoS currents* 2014; **6**: doi:10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.
283. Volz E, Pond S. Phylodynamic analysis of ebola virus in the 2014 sierra leone epidemic. *PLoS currents* 2014; **6**: doi:10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e.
284. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016; **530**(7589): 228-32. doi:10.1038/nature16996.

285. Camacho A, Kucharski A, Aki-Sawyerr Y, et al. Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-time Modelling Study. *PLoS currents* 2015; **7**: doi:10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2.
286. Kucharski AJ, Camacho A, Checchi F, et al. Evaluation of the benefits and risks of introducing Ebola community care centers, Sierra Leone. *Emerging Infectious Diseases* 2015; **21**(3): 393-9. doi:10.3201/eid2103.141892.
287. Chretien JP, Riley S, George DB. Mathematical modeling of the West Africa Ebola epidemic. *eLife* 2015; **4**: doi:10.7554/eLife.09186.
288. Mari Saez A, Weiss S, Nowak K, et al. Investigating the zoonotic origin of the West African Ebola epidemic. *EMBO Mol Med* 2014; **7**(1): 17-23. doi:10.15252/emmm.201404792.
289. Ansumana R, Jacobsen KH, Sahr F, et al. Ebola in Freetown area, Sierra Leone--a case study of 581 patients. *New England Journal of Medicine* 2015; **372**(6): 587-8. doi:10.1056/NEJMc1413685.
290. Majumder M. Estimating the fatality of the 2014 West African Ebola Outbreak. 10/09/2014 2014. http://www.healthmap.org/site/diseasedaily/article/estimating-fatality-2014-west-african-ebola-outbreak-910142016). (Accessed 26 October 2016)
291. Heymann DL, Barakamfitiye D, Szczeniowski M, et al. Ebola Hemorrhagic Fever: Lessons from Kikwit, Democratic Republic of the Congo. *The Journal of Infectious Diseases* 1999; **179**: S283-S6. http://www.jstor.org/stable/30117635.
292. Rio C. The Ebola Crisis: Lessons in International Cooperation for Global Health 2015. http://www.aahcdc.org/Portals/0/mtgs/if15/DEL%20RIO%20Ebola.pdf, (Accessed 10 October 2016).
293. Boseley S WJ. World Health Organisation declares Zika virus public health emergency. The Guardian. 2016 01/02/2016. (Accessed 10 October 2016). https://www.theguardian.com/world/2016/feb/01/zika-virus-world-health-organisation-declares-global-health-emergency
294. Khan AS, Tshioko FK, Heymann DL, et al. The reemergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995. Commission de Lutte contre les Epidemies a Kikwit. *The Journal of Infectious Diseases* 1999; **179 Suppl 1**: S76-86. doi:10.1086/514306.
295. Casillas AM, Nyamathi AM, Sosa A, et al.A current review of Ebola virus: pathogenesis, clinical presentation, and diagnostic assessment. *Biological Research for Nursing* 2003; **4**(4): 268-75.
296. Tomori O, Bertolli J, Rollin PE, et al. Serologic survey among hospital and health center workers during the Ebola hemorrhagic fever outbreak in Kikwit, Democratic Republic of the Congo, 1995. *The Journal of Infectious Diseases* 1999; **179 Suppl 1**: S98-101. doi:10.1086/514307.
297. Wamala JF, Lukwago L, Malimbo M, et al. Ebola hemorrhagic fever associated with novel virus strain, Uganda, 2007-2008. *Emerging Infectious Diseases* 2010; **16**(7): 1087-92. doi:10.3201/eid1607.091525.
298. CDC. Bioterrorism agents/diseases. http://emergency.cdc.gov/agent/agentlist-category.asp (Accessed 24th February 2015).
299. Borio L, Inglesby T, Peters CJ, et al. Hemorrhagic fever viruses as biological weapons: medical and public health management. *Journal of the American Medical Association* 2002; **287**(18): 2391-405.
300. Marzi A, Feldmann F, Geisbert TW, et al. Vesicular stomatitis virus-based vaccines against Lassa and Ebola viruses. *Emerging Infectious Diseases* 2015; **21**(2): 305-7. doi:10.3201/eid2102.141649.
301. LSHTM. Ebola vaccine trail funding announced by the Innovative Medicines Initiative. http://www.lshtm.ac.uk/newsevents/news/2015/ebola_vaccine_trial_funding.html (Accessed 24 February 2015).

302. Jones SM, Feldmann H, Stroher U, et al. Live attenuated recombinant vaccine protects nonhuman primates against Ebola and Marburg viruses. *Nature Medicine* 2005; **11**(7): 786-90. doi:10.1038/nm1258.
303. Sullivan NJ, Geisbert TW, Geisbert JB, et al. Immune protection of nonhuman primates against Ebola virus with single low-dose adenovirus vectors encoding modified GPs. *PLoS Med* 2006; **3**(6): e177. doi:10.1371/journal.pmed.0030177.
304. Faye O, Boelle PY, Heleze E, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *The Lancet Infectious Diseases* 2015; **15**(3): 320-6. doi:10.1016/s1473-3099(14)71075-8.
305. BBC News. Ebola: Liberia confirms cases, Senegal shuts border. http://www.bbc.com/news/world-africa-26735118 (Accessed 30 January 2015).
306. BBC News. Seven die in Monrovia Ebola outbreak. http://www.bbc.co.uk/news/world-africa-27888363 (Accessed 30 January 2015).
307. Africa Online News. Ebola Deaths Turn Redemption Hospital into Ghost Town. http://frontpageafricaonline.com/index.php/health-sci/1987-ebola-deaths-turn-liberia-s-redemption-hospital-into-ghost-town (Accessed 30 January 2015).
308. All Africa News. Liberia: Ebola Kills Doctor At Redemption Hospital. http://allafrica.com/stories/201407021024.html (Accessed 30 January 2015).
309. All Africa News. Liberia: Four Nurses in Ebola Web At Phebe Hospital. http://allafrica.com/stories/201407211455.html (Accessed 30 January 2015).
310. Schieffelin JS, Shaffer JG, Goba A, et al. Clinical illness and outcomes in patients with Ebola in Sierra Leone. *New England Journal of Medicine* 2014; **371**(22): 2092-100. doi:10.1056/NEJMoa1411680.
311. CDC. Outbreak of Ebola viral hemorrhagic fever--Zaire, 1995. *MMWR Morbidity and mortality weekly report* 1995; **44**(19): 381-2.
312. Hall RC, Hall RC, Chapman MJ. The 1995 Kikwit Ebola outbreak: lessons hospitals and physicians can apply to future viral epidemics. *General Hospital Psychiatry* 2008; **30**(5): 446-52. doi:10.1016/j.genhosppsych.2008.05.003.
313. Muyembe-Tamfum JJ, Mulangu S, Masumu J, et al. Ebola virus outbreaks in Africa: past and present. *The Onderstepoort Journal of Veterinary Research* 2012; **79**(2): 451. doi:10.4102/ojvr.v79i2.451.
314. Reiter P, Turell M, Coleman R, et al. Field investigations of an outbreak of Ebola hemorrhagic fever, Kikwit, Democratic Republic of the Congo, 1995: arthropod studies. *The Journal of Infectious Diseases* 1999; **179 Suppl 1**: S148-54. doi:10.1086/514304.
315. Ndambi R, Akamituna P, Bonnet MJ, et al. Epidemiologic and clinical aspects of the Ebola virus epidemic in Mosango, Democratic Republic of the Congo, 1995. *Journal of Infectious Diseases* 1999; **179 Suppl 1**: S8-10. doi:10.1086/514297.
316. CDC. Ebola virus Disease: Transmission. http://www.cdc.gov/vhf/ebola/transmission (Accessed 24 February 2015).
317. WHO. Global Atlas of the Health Workforce. http://apps.who.int/gho/data/node.main.A1444?lang=en&showonly=HWF (Accessed 30 January 2015).
318. Leroy EM, Baize S, Volchkov VE, et al. Human asymptomatic Ebola infection and strong inflammatory response. *The Lancet* 2000; **355**(9222): 2210-5.
319. WHO. WHO Ebola Sitrep. http://apps.who.int/ebola/ebola-situation-reports (Accessed 13 June 2016).
320. Bah OM, Kamara HB, Bhat P, et al. The influence of the Ebola outbreak on presumptive and active tuberculosis in Bombali District, Sierra Leone. *Public Health Action* 2017; **7**(Suppl 1): S3-S9. doi:10.5588/pha.16.0093.
321. Ministry_of_Health_and_Sanitation. Ebola virus disease situation report 504. Freetown, Sierra Leone, 2015.
322. CDC – Personal communication informal census data. 2017.

323. Gleason B, Redd J, Kilmarx P, et al. Establishment of an Ebola Treatment Unit and Laboratory - Bombali District, Sierra Leone, July 2014-January 2015. *MMWR Morbidity and mortality weekly report* 2015; **64**(39): 1108-11. doi:10.15585/mmwr.mm6439a4.
324. WHO. Clinical management of patients with viral haemorrhagic fever: a pocket guide for the front-line health worker. Geneva: WHO, 2014. http://apps.who.int/iris/bitstream/10665/130883/2/WHO_HSE_PED_AIP_14.05.pdf. (Accessed 17 October 2016).
325. Medecins_Sans_Frontieres. Filovirus haemorrhagic fever guideline. Barcelona: MSF, 2008. https://www.medbox.org/ebola-guidelines/filovirus-haemorrhagic-feverguideline/preview. (Accessed 16 October 2016).
326. Roshania R, Mallow M, Dunbar N, et al. Successful Implementation of a Multicountry Clinical Surveillance and Data Collection System for Ebola Virus Disease in West Africa: Findings and Lessons Learned. *Global Health: Science and Practice* 2016; **4**(3): 394-409. doi:10.9745/ghsp-d-16-00186.
327. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30**(9): 1312-3. doi:10.1093/bioinformatics/btu033.
328. Anisimova M, Gil M, Dufayard J-F,et al. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology* 2011; **60**(5): 685-99. doi:10.1093/sysbio/syr041.
329. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* 2010; **5**(3): e9490. doi:10.1371/journal.pone.0009490.
330. Jombart T, Cori A, Didelot X, et al. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology* 2014; **10**: doi:http://dx.doi.org/10.1371/journal.pcbi.1003457.
331. Eichner M, Dowell SF, Firese N. Incubation Period of Ebola Hemorrhagic Virus Subtype Zaire. *Osong Public Health and Research Perspectives* 2011; **2**(1): 3-7. doi:10.1016/j.phrp.2011.04.001.
332. Khan A, Naveed M, Dur-e-Ahmad M, Imran M. Estimating the basic reproductive ratio for the Ebola outbreak in Liberia and Sierra Leone. *Infectious Diseases of Poverty* 2015; **4**: 13. doi:10.1186/s40249-015-0043-3.
333. Rosenke K, Adjemian J, Munster VJ, et al. Plasmodium Parasitemia Associated With Increased Survival in Ebola Virus–Infected Patients. *Clinical Infectious Diseases* 2016; **63**(8): 1026-33. doi:10.1093/cid/ciw452.
334. Goodfellow I and team. 2016. Personal communication
335. Rettner R. Ebola Virus Lives on Hospital Surfaces for Days. 2017. https://www.livescience.com/50758-ebola-virus-survival-surfaces.html (Accessed 19 September 2017).
336. Manguvo A, Mafuvadze B. The impact of traditional and religious practices on the spread of Ebola in West Africa: time for a strategic shift. *The Pan African Medical Journal* 2015; **22**: 9. doi:http://doi.org/10.11694/pamj.supp.2015.22.1.6190.
337. Chenna R, Sugawara H, Koike T, et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* 2003; **31**(13): 3497-500. doi:10.1093/nar/gkg500.
338. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005; **21**(5): 676-9. doi:10.1093/bioinformatics/bti079.
339. Vandormael A, Dobra A, Bärnighausen T,et al.. Incidence rate estimation, periodic testing and the limitations of the mid-point imputation approach. *International Journal of Epidemiology* 2017; 10.1093/ije/dyx134: doi:10.1093/ije/dyx134.
340. Bouckaert R, Heled J, Kühnert D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology* 2014; **10**(4): e1003537. doi:10.1371/journal.pcbi.1003537.

341. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biology* 2006; **4**(5): e88. doi:10.1371/journal.pbio.0040088.
342. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 2005; **22**(5): 1185-92. doi:10.1093/molbev/msi103.
343. Rambaut A, Suchard MA, Xie D, Drummond A. Tracer v1.6. 2014. http://tree.bio.ed.ac.uk/software/tracer/. (Accessed 10 September 2017).
344. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 2010; **59**(3): 307-21. doi:10.1093/sysbio/syq010.
345. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *Journal of Molecular Biol* ogy 1990; **215**(3): 403-10. doi:10.1016/s0022-2836(05)80360-2.
346. Williams B, Campbell C. Creating alliances for disease management in industrial settings: a case study of HIV/AIDS in workers in South African gold mines. *International Journal of Occupational and Environmental Health* 1998; **4**(4): 257-64. doi:10.1179/oeh.1998.4.4.257.
347. Dobra A, Barnighausen T, Vandormael A, Tanser F. Space-time migration patterns and risk of HIV acquisition in rural South Africa. *AIDS* 2017; **31**(1): 137-45. doi:10.1097/qad.0000000000001292.
348. Williams BG, Taljaard D, Campbell CM, et al. Changing patterns of knowledge, reported behaviour and sexually transmitted infections in a South African gold mining community. *AIDS* 2003; **17**(14): 2099-107. doi:10.1097/01.aids.0000076323.42412.26.
349. Shahmanesh M, Patel V, Mabey D, Cowan F. Effectiveness of interventions for the prevention of HIV and other sexually transmitted infections in female sex workers in resource poor setting: a systematic review. *Tropical Medicine and International Health* 2008; **13**(5): 659-79. doi:10.1111/j.1365-3156.2008.02040.x.
350. Wertheim J. 2017. Personal communication.
351. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences* 2016; **113**(10): 2690-5. doi:10.1073/pnas.1522930113.
352. Haber N, Tanser F, Bor J, et al. From HIV infection to therapeutic response: a population-based longitudinal HIV cascade-of-care study in KwaZulu-Natal, South Africa. *The Lancet HIV* 2017; **4**(5): e223-e30. doi:10.1016/s2352-3018(16)30224-7.
353. CDC. HIV Risk Behaviours. http://www.cdc.gov/hiv/risk/estimates/riskbehaviors.html (accessed 13th August 2016).
354. Iwuji CC, Orne-Gliemann J, Tanser F, et al. Evaluation of the impact of immediate versus WHO recommendations-guided antiretroviral therapy initiation on HIV incidence: the ANRS 12249 TasP (Treatment as Prevention) trial in Hlabisa sub-district, KwaZulu-Natal, South Africa: study protocol for a cluster randomised controlled trial. *Trials* 2013; **14**: 230. doi:10.1186/1745-6215-14-230.
355. Iwuji CC, Orne-Gliemann J, Larmarange J, et al. Uptake of Home-Based HIV Testing, Linkage to Care, and Community Attitudes about ART in Rural KwaZulu-Natal, South Africa: Descriptive Results from the First Phase of the ANRS 12249 TasP Cluster-Randomised Trial. *PLoS Med*icine 2016; **13**(8): e1002107. doi:10.1371/journal.pmed.1002107.
356. Mishra S, Steen R, Gerbase A,et al.. Impact of High-Risk Sex and Focused Interventions in Heterosexual HIV Epidemics: A Systematic Review of Mathematical Models. *PLOS ONE* 2012; **7**(11): e50691. doi:10.1371/journal.pone.0050691.
357. Weine SM, Kashuba AB. Labor Migration and HIV Risk: A Systematic Review of the Literature. *AIDS and Behavior* 2012; **16**(6): 1605-21. doi:10.1007/s10461-012-0183-4.
358. Collinson MA, White MJ, Bocquier P, et al. Migration and the epidemiological transition: insights from the Agincourt sub-district of northeast South Africa. *Global Health Action* 2014; **7**: 10.3402/gha.v7.23514. doi:10.3402/gha.v7.23514.

359. Lurie M, Harrison A, Wilkinson D, Karim SA. Circular migration and sexual networking in rural KwaZulu/Natal: implications for the spread of HIV and other sexually transmitted diseases. *Health Transition Review* 1997; **7**: 17-27. http://www.jstor.org/stable/40608686.
360. Wringe A, Cremin I, Todd J, et al. Comparative assessment of the quality of age-at-event reporting in three HIV cohort studies in sub-Saharan Africa. *Sexually Transmitted Infections* 2009; **85**: doi:10.1136/sti.2008.033423.
361. SASPEN. Migration and health care: The case of HIV and AIDS in Botswana. 2015.
362. Harling G, Newell ML, Tanser F, et al. Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in rural KwaZulu-Natal, South Africa. *Journal of Acquir ed Immune Deficiency Syndromes*2014; **66**(4): 443-51. doi:10.1097/qai.0000000000000198.
363. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data--or tears: an application to educational enrollments in states of India. *Demography* 2001; **38**(1): 115-32.
364. Rutstein SO, Johnson K. The DHS Wealth Index. Calverton, Maryland, 2004.
365. Vandormael A, Dobra A, Barnighausen T, de Oliveira T, F. T. Incidence rate estimation, periodic testing and the limitations of the mid-point imputation approach. *International Journal of Epidemiology* 2017; https://doi.org/10.1093/ije/dyx134: doi:https://doi.org/10.1093/ije/dyx134.
366. Ishikawa S. PASTML Ancestral Reconstruction. 2017. https://github.com/saishikawa/PASTML (Accessed 5 September 2017).
367. Ishikawa S. Ancestral StaTe Reconstruction using mAximum-Likelihood (ASTRAL) Program. 2017. https://github.com/saishikawa/ASTRAL (Accessed 5 September 2017).
368. Pupko T, Pe I, Shamir R, Graur D. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution* 2000; **17**(6): 890-6. doi:10.1093/oxfordjournals.molbev.a026369.
369. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 2007; **24**(8): 1586-91. doi:10.1093/molbev/msm088.
370. Rossum G. Python reference manual: CWI (Centre for Mathematics and Computer Science), 1995.
371. Huèrta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* 2016; **33**(6): 1635-8. doi:10.1093/molbev/msw046.
372. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003; **13**(11): 2498-504. doi:10.1101/gr.1239303.
373. Waldman A. On India's Roads: Cargo and a Deadly Passenger. *New York Times, New York* 2005; **6**. (Accessed 5 September 2017).
374. Amon J, Todrys K. Access to antiretroviral treatment for migrant populations in the global South. *Sur Revista Internacional de Direitos Humanos* 2009; **6**: 162-87. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-64452009000100009&nrm=iso.
375. Tomita A, Vandormael AM, Barnighausen T,et al.. Social Disequilibrium and the Risk of HIV Acquisition: A Multilevel Study in Rural KwaZulu-Natal Province, South Africa. *Journal of Acquired Immune Deficiency Syndromes* 2017; **75**(2): 164-74. doi:10.1097/qai.0000000000001349.
376. Dare OO, Cleland JG. Reliability and validity of survey data on sexual behaviour. *Health Transit Rev* 1994; **4**.
377. Tourangeau R, Yan T. Sensitive questions in surveys. *Psychology Bulletin* 2007; **133**: doi:10.1037/0033-2909.133.5.859.

378. Harling G, Gumede D, Mutevedzi T, et al. The impact of self-interviews on response patterns for sensitive topics: a randomized trial of electronic delivery methods for a sexual behaviour questionnaire in rural South Africa. *BMC Medical Research Methodology* 2017; **17**(1): 125. doi:10.1186/s12874-017-0403-8.
379. Fenton KA, Johnson AM, McManus S, Erens B. Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections* 2001; **77**(2): 84-92.
380. UNAIDS. AIDS by numbers 2015. Geneva: http://www.unaids.org/sites/default/files/media_asset/AIDS_by_the_numbers_2015_en.pdf., (Accessed 17 October 2016).
381. Cohen MS, Chen YQ, McCauley M, et al. Antiretroviral Therapy for the Prevention of HIV-1 Transmission. *New England Journal of Medicine* 2016; **375**(9): 830-9. doi:10.1056/NEJMoa1600693.
382. Moscona R, Ram D, Wax M, et al. Comparison between next-generation and Sanger-based sequencing for the detection of transmitted drug-resistance mutations among recently infected HIV-1 patients in Israel, 2000–2014. *Journal of the International AIDS Society* 2017; **20**(1): 21846. doi:10.7448/IAS.20.1.21846.
383. UNAIDS. Focus on location and population. Geneva, Switzerland: UNAIDS, 2015. http://www.unaids.org/sites/default/files/media_asset/WAD2015_report_en_part01.pdf. (Accessed 17 October 2016).
384. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. 2014, https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (Accessed 17 October 2016).
385. Council for International Organizations of Medical Sciences (CIOMS) and WHO. International Ethical Guidelines for Health Related Research Involving Humans. Geneva, 2016. https://cioms.ch/shop/product/international-ethical-guidelines-for-health-related-research-involving (Accessed 26 October 2016).
386. Geller G, Dvoskin R, Thio CL, et al. Genomics and infectious disease: a call to identify the ethical, legal and social implications for public health and clinical practice. *Genome Medicine* 2014; **6**(11): 106. doi:10.1186/s13073-014-0106-2.
387. Jao I, Kombe F, Mwalukore S, et al. Research Stakeholders' Views on Benefits and Challenges for Public Health Research Data Sharing in Kenya: The Importance of Trust and Social Relations. *PLoS One* 2015; **10**(9): e0135545. doi:10.1371/journal.pone.0135545.
388. Parker M, Kwiatkowski DP. The ethics of sustainable genomic research in Africa. *Genome Biology* 2016; **17**: 44. doi:10.1186/s13059-016-0914-3.
389. de Vries J, Munung SN, Matimba A, et al. Regulation of genomic and biobanking research in Africa: a content analysis of ethics guidelines, policies and procedures from 22 African countries. *BMC Medical Ethics* 2017; **18**(1): 8. doi:10.1186/s12910-016-0165-6.
390. MalariaGEN. https://www.malariagen.net/ethics (Accessed July 2017).
391. H3 Africa. https://h3africa.org/about/ehtics-and-governance (Accessed July 2017).
392. de Vries J, Tindana P, Littler K, et al. The H3Africa policy framework: negotiating fairness in genomics. *Trends in Genetics* 2015; **31**(3): 117-9. doi:10.1016/j.tig.2014.11.004.
393. Tindana P, Bull S, Amenga-Etego L, et al. Seeking consent to genetic and genomic research in a rural Ghanaian setting: a qualitative study of the MalariaGEN experience. *BMC Medical Ethics* 2012; **13**: 15. doi:10.1186/1472-6939-13-15.
394. Emanuel EJ, Wendler D, Killen J, Grady C. What makes clinical research in developing countries ethical? The benchmarks of ethical research. *The Journal of Infectious Diseases* 2004; **189**(5): 930-7. doi:10.1086/381709.
395. Parker M, Bull SJ, Vries J,et al.. Ethical data release in genome-wide association studies in developing countries. *PLoS Medicine* 2009; **6**: doi:10.1371/journal.pmed.1000143.

396. Bull S, Cheah PY, Denny S, et al. Best Practices for Ethical Sharing of Individual-Level Health Research Data From Low- and Middle-Income Settings. *Journal of Empirical Research on Human Research Ethics* 2015; **10**(3): 302-13. doi:10.1177/1556264615594606.
397. Wellcome_Trust. Ethical-sharing-of-health-research-data-in-low-and-middle-income-countries. 2014.((Accessed 20 October 2016).
398. Cohen J. Pinpointing HIV spread in Africa poses risks. *Science* 2017; **356**(6338): 568-9. doi:10.1126/science.356.6338.568.
399. NIH. Ethical, Legal and Social Implications Program. https://www.genome.gov/elsi (Accessed 27 June 2017).
400. Paraskevis D, Pybus O, Magiorkinis G, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. *Retrovirology* 2009; **6**: 49. doi:10.1186/1742-4690-6-49.
401. Brown AE, Gifford RJ, Clewley JP, et al. Phylogenetic reconstruction of transmission events from individuals with acute HIV infection: toward more-rigorous epidemiological definitions. *The Journal of Infectious Diseases* 2009; **199**(3): 427-31. doi:10.1086/596049.
402. Batorsky R, Sergeev RA, Rouzine IM. The route of HIV escape from immune response targeting multiple sites is determined by the cost-benefit tradeoff of escape mutations. *PLoS Computational Biology* 2014; **10**(10): e1003878. doi:10.1371/journal.pcbi.1003878.
403. Dunn D, Pillay D. UK HIV drug resistance database: background and recent outputs. *Journal of HIV Therapy* 2007; **12**(4) 97-98 .
404. Schoeni-Affolter F, Ledergerber B, Rickenbach M, et al. Cohort profile: the Swiss HIV Cohort study. *International Journal of Epidemiology* 2010; **39**(5): 1179-89. doi:10.1093/ije/dyp321.
405. Lemey P, Van Dooren S, Van Laethem K, et al. Molecular testing of multiple HIV-1 transmissions in a criminal case. *AIDS* 2005; **19**(15): 1649-58.
406. Bernard EJ, Azad Y, Vandamme AM,et al. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV medicine* 2007; **8**(6): 382-7. doi:10.1111/j.1468-1293.2007.00486.x.
407. Carlson JM, Du VY, Pfeifer N, et al. Impact of pre-adapted HIV transmission. *Nature Medicine* 2016; **22**(6): 606-13. doi:10.1038/nm.4100.
408. Statistics Botswana. Population and Housing Census - Analytical Report. Gaborone, 2014. www.statsbots.org (Accessed 12 January 2016)
409. Jacques G, Mmatli TO. Addressing Ethical Non-Sequiturs in Botswana's HIV and AIDS Policies: Harmonising the Halo Effect. *Ethics and Social Welfare* 2013; **7**(4): 342-58. doi:10.1080/17496535.2013.768071.
410. Munthe C. The Goals of Public Health: An Integrated, Multidimensional Model. *Public Health Ethics* 2008; **1**(1): 39-52. doi:10.1093/phe/phn006.
411. Sox HC. Resolving the tension between population health and individual health care. *Journal of the American Medical Association* 2013; **310**(18): 1933-4. doi:10.1001/jama.2013.281998.
412. Gostin LO, Bayer R, Fairchild AL. Ethical and legal challenges posed by severe acute respiratory syndrome: implications for the control of severe infectious disease threats. *Journal of the American Medical Association* 2003; **290**(24): 3229-37. doi:10.1001/jama.290.24.3229.
413. Cohen MS, McCauley M, Sugarman J. Establishing HIV treatment as prevention in the HIV Prevention Trials Network 052 randomized trial: an ethical odyssey. *Clinical Trials* 2012; **9**(3): 340-7. doi:10.1177/1740774512443594.
414. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 Infection with Early Antiretroviral Therapy. *New England Journal of Medicine* 2011; **365**(6): 493-505. doi:10.1056/NEJMoa1105243.

415. Rutstein SE, Ananworanich J, Fidler S, et al. Clinical and public health implications of acute and early HIV detection and treatment: a scoping review. *Journal of the International AIDS Society* 2017; **20**(1): 1-13. doi:10.7448/ias.20.1.21579.
416. Cohen MS, McCauley M, Sugarman J. The Ethical Odyssey in Testing HIV Treatment as Prevention. *Clinical trials* 2012; **9**(3): 340-7. doi:10.1177/1740774512443594.
417. Shannon K, Strathdee SA, Goldenberg SM, et al. Global epidemiology of HIV among female sex workers: influence of structural determinants. *The Lancet* 2015; **385**(9962): 55-71. doi:10.1016/s0140-6736(14)60931-4.
418. Human Rights Watch. Tell me where I can be safe? The impact of Nigeria's same sex marriage (prohibition) act. 2016. https://www.hrw.org/report/2016/10/20/tell-me-where-i-can-be-safe/impact-nigerias-same-sex-marriage-prohibition-act (Accessed 25 May 2017).
419. Cloete A, Simbayi L, Zuma K, et al. The People Living With HIV Stigma Index: South Africa 2014: Human Sciences Research Council, South Africa, UNAIDS, Eastern Cape AIDS Council, 2014. (Accessed 25 May 2017).
420. News 24. Tanzania threatens to publish list of gay people. 2017. http://www.news24.com/Africa/News/tanzania-threatens-to-publish-list-of-gay-people-20170218 (Accessed 25 August 2017).
421. Amon J. Personal communication with UNAIDS. 2017.
422. WHO. Health and human rights. 2015. http://www.who.int/mediacentre/factsheets/fs323/en/ (Accessed 23 August 2017).
423. Bernard E, Cameron S. Advancing HIV Justice and Advancing HIV Justice 2: Building momentum in global advocacy against HIV criminalisation. Brighton/Amsterdam: HIV Justice Network and GNP+. 2016. https://www.scribd.com/doc/312008825/Advancing-HIV-Justice-2-Building-momentum-in-global-advocacy-against-HIV-criminalisation#download&from_embed. (Accessed 25 August 2017).
424. O'Byrne P, Willmore J, Bryan A, et al. Nondisclosure prosecutions and population health outcomes: examining HIV testing, HIV diagnoses, and the attitudes of men who have sex with men following nondisclosure prosecution media releases in Ottawa, Canada. *BMC Public Health* 2013; **13**: 94. doi:10.1186/1471-2458-13-94.
425. Amon JJ, Baral SD, Beyrer C, Kass N. Human rights research and ethics review: protecting individuals or protecting the state? *PLoS Medicine* 2012; **9**(10): e1001325. doi:10.1371/journal.pmed.1001325.
426. Abecasis AB, Geretti AM, Albert J, Power L, Weait M, Vandamme AM. Science in court: the myth of HIV fingerprinting. *The Lancet Infectious diseases* 2011; **11**(2): 78-9. doi:10.1016/s1473-3099(10)70283-8.
427. Leitner T. Guidelines for HIV in court cases. *Nature* 2011; **473**(7347): 284. doi:10.1038/473284a.
428. Department of Health, Education and Welfare; National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. 2015/05/09 ed; 2014. p. 4-13.
429. Barchi F, Matlhagela K, Jones N, Kebaabetswe PM, Merz JF. "The keeping is the problem": A qualitative study of IRB-member perspectives in Botswana on the collection, use, and storage of human biological samples for research. *BMC Medical Ethics* 2015; **16**: 54. doi:10.1186/s12910-015-0047-3.
430. UNAIDS. Good participatory practice: Guidelines for HIV biomedical prevention trials. Geneva: UNAIDS, 2011. (Accessed 12 October 2016).
431. WHO U. Ethical considerations in biomedical HIV prevention trials: UNAIDS-WHO guidance document. Geneva, Switzerland: UNAIDS & WHO, 2012.
432. Singh JA, Mills EJ. The abandoned trials of pre-exposure prophylaxis for HIV: what went wrong? *PLoS Medicine* 2005; **2**(9): e234. doi:10.1371/journal.pmed.0020234.

433. Vermund SH, Blevins M, Moon TD, et al. Poor clinical outcomes for HIV infected children on antiretroviral therapy in rural Mozambique: need for program quality improvement and community engagement. *PLoS One* 2014; **9**(10): e110116. doi:10.1371/journal.pone.0110116.
434. PopART. PopART website. 2017. https://www.lshtm.ac.uk/popart (Accessed 26 September 2017).
435. Simwinga M, Mwanza F. PopART study, Zambia, Community Engagement study. 2017.
436. Boyson AR, Zimmerman RS, Shoemaker S. Exemplification of HAART and HIV/AIDS: A News Experiment. *Health Communication* 2015; **30**(9): 901-10. doi:10.1080/10410236.2014.903222.
437. Zimmerman RS, Kirschbaum AL. News of Biomedical Advances in HIV: Relationship to Treatment Optimism and Expected Risk Behavior in US MSM. *AIDS Behaviour* 2017; 10.1007/s10461-017-1744-3: doi:10.1007/s10461-017-1744-3.
438. De Maio N, Wu C-H, Wilson DJ. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Computational Biology* 2016; **12**(9): e1005130. doi:10.1371/journal.pcbi.1005130.
439. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology* 2016; **11**(12): e1004613. doi:10.1371/journal.pcbi.1004613.
440. Didelot X, Gardy J, Colijn C. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution* 2014; **31**(7): 1869-79. doi:10.1093/molbev/msu121.
441. Whitty CJM, Mundel T, Farrar J, et al. Providing incentives to share data early in health emergencies: the role of journal editors. *The Lancet*; 2015 **386**(10006): 1797-8. doi:10.1016/S0140-6736(15)00758-8.
442. Heymann DL. The international response to the outbreak of SARS in 2003. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2004; **359**(1447): 1127-9. doi:10.1098/rstb.2004.1484.
443. Ebola Response Anthropology Platform. 2014. http://www.ebola-anthropology.net/ (Accessed 9 October 2017).
444. Stanley DA, Honko AN, Asiedu C, et al. Chimpanzee adenovirus vaccine generates acute and durable protective immunity against ebolavirus challenge. *Nature Medicine* 2014; **20**(10): 1126-9. doi:10.1038/nm.3702 http://www.nature.com/nm/journal/v20/n10/abs/nm.3702.html#supplementary-information.
445. Groves T. How research data sharing can save lives. *British Medical Journal Opinion* 2015, http://blogs.bmj.com/bmj/2015/09/08/trish-groves-how-research-data-sharing-can-save-lives/.
446. Wellcome_Trust. Sharing data during Zika and other global health emergencies. 2016. https://wellcome.ac.uk/news/sharing-data-during-zika-and-other-global-health-emergencies (Accessed 9 October 2017).
447. Van Noorden R. Gates Foundation announces world's strongest policy on open access research *Nature News Blog* 2015, http://blogs.nature.com/news/2014/11/gatesfoundation-announces-worlds-strongest-policy-on-open-access-research.html http://www.gatesfoundation.org/how-we-work/generalinformation/open-access-policy.
448. WHO. Developing global norms for sharing data and results during public helath emergencies. 2015. http://www.who.int/medicines/ebola-treatment/data-sharing_phe/en/ (Accessed 9 October 2017).
449. Joint Global Statement. Statement on Data Sharing in Public Health Emergencies. *Multiple* 2016, https://www.biomedcentral.com/about/press-centre/business-press-releases/11-02-16. (Accessed 9 October 2017).

450. Haynes CL, Cook GA, Jones MA. Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learnt from developing a disease register. *Journal of Medical Ethics* 2007; **33**(5): 302-7. doi:10.1136/jme.2006.016907.
451. Greenough A, Graham H. Protecting and using patient information: the role of the Caldicott Guardian. *Clinical Medicine (Northfield Il)* 2004; **4**(3): 246-9.
452. Adeloye D, Thompson JY, Akanbi MA, et al. The burden of road traffic crashes, injuries and deaths in Africa: a systematic review and meta-analysis. *Bulletin of the World Health Organization* 2016; **94**(7): 510-21a. doi:10.2471/blt.15.163121.
453. WHO. Global status report on road safety. Geneva: WHO, 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. (Accessed 12 October 2017).
454. Dean AG, Arner TG, Sunki GG, et al. Epi Info™, a database and statistics program for public health professionals. . Atlanta, GA, USA: CDC, 2011.
455. McCormack JE, Hird SM, Zellmer AJ, et al. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* 2013; **66**(2): 526-38. doi:https://doi.org/10.1016/j.ympev.2011.12.007.
456. Lemmon EM, Lemmon AR. High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 2013; **44**(1): 99-121. doi:10.1146/annurev-ecolsys-110512-135822.
457. Elm Ev, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Brtish Medical Journal* 2007; **335**(7624): 806. http://www.bmj.com/content/335/7624/806.abstract.