Using a phylogenetic approach that combines laboratory and clinical data to enhance understanding of HIV transmission events among men who have sex with men

This work is presented by

Alison Elizabeth Brown

as a thesis for the degree of

Doctor of Philosophy

in the

Division of Population Health

University College London

2009

Acknowledgements

This work was funded by a Medical Research Council Special Training Fellowship in Health Services and Health of the Public Research (G106/1219).

Many thanks to my supervisors for their encouragement, support, interest and ideas: Jonathan Clewley, Noel Gill, Anne Johnson and Deenan Pillay. Thanks also to Pam Sonnenberg for providing thoughtful insights.

There are several people without whose generosity this work would not have been possible: Stephane Hue, Oliver Pybus and Robert Gifford for providing phylogenetics training; Martin Fisher, David Pao and Darshan Sudarshi for allowing me access to the Brighton data; Kholoud Porter for providing the CASCADE data; and Fraser Lewis and Andrew Leigh-Brown for providing data for chapter four. Thanks to Caroline Sabin and Andrew Buckton for help with obtaining UK CHIC and sequence data respectively. Thanks also to Caroline for her contributions towards the sensitivity analysis in chapter six.

Thanks to Barry Evans and Valerie Delpech for their support and for allowing me to write the thesis when I should have been doing my proper job. I would also like to thank Louise Logan, Caterina Hill, Susie Huntington, Alicia Thornton and Brian Rice for their advice and encouragement.

On a personal level, thanks to Anna Scott and Karen Troup for their ill-advised belief that the time would come when I would submit. Many, many thanks to Mary Brown (my mum) for proof reading (it is thanks to her that there definitely are no split infinitives in this thesis). Finally, thank you to Ian Bruce for his support, and the sacrifices he has made over the past four years. (I suspect he only proposed to give me something else to talk about).

Candidate's contribution

I 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.

I was provided with the Brighton, CASCADE and the London and Manchester sequence datasets for the purpose of the thesis. I was substantially involved with the data collation for the Brighton dataset. Between 2003-5, I was the survey co-ordinator of the UA STI survey. This involved co-ordinating all aspects of the survey, and specifically for this thesis, included the selection of HIV positive samples to be DNA sequenced for epidemiological purposes.

I prepared and cleaned the data from all four datasets, and designed and conducted all of the analyses, (including the phylogenetic analyses). For the Brighton analysis, statistical assistance was provided in undertaking the multi-variate and univariable analyses, including the sensitivity analyses. The writing of the thesis was undertaken by me.

Alison Brown

Abstract

A phylogenetic approach combining HIV *pol* sequences with laboratory and clinical data was undertaken to explore HIV transmissions between men who have sex with men (MSM). Combining putative transmission events (reconstructed through phylogenetic analyses of *pol* sequences) with clinical (e.g. viral load) and diagnostic (e.g. recently-acquired infection) data can enhance understanding of HIV transmission more than can be gleaned from each individual source.

The thesis: assessed the consistency of phylogenetic reconstructions of HIV transmission events; explored transmissions from recently HIV-infected MSM at diagnosis and critiqued such analyses; and ascertained which groups of diagnosed HIV-infected MSM are generating HIV transmissions.

Sensitivity analyses demonstrated that phylogenetic reconstructions of transmission events were 80% consistent as sample sizes were varied. Previous phylogenetic reconstructions overestimated transmission from recently HIV-infected MSM through failing to recognize that this infection stage is transitory. Comparison of infection dates between recently HIV-infected MSM involved in transmission events revealed only half of the transmissions were generated during recent infection. Through allowing infection stage (and other markers of transmission risk) to reflect the course of HIV infection it was established that the recently HIV-infected have a transmission risk of 3.04 (compared to the chronically HIV-infected population). Transmission rates were elevated among the untreated population; 72% (28/39) were generated from

treatment-naïve MSM and 23% (9/39) from MSM interrupting treatment. Overall, 69% (27/39) of transmissions occurred from MSM with CD4 counts >350 cells/mm^3.

BHIVA guidelines recommend treatment discussions start when patients' CD4 counts reach 200-350mm^3. This work contributes to the debate on the public health benefit of treating all HIV-diagnosed individuals, regardless of clinical need. Behavioural interventions need to increase awareness of recent HIV infection, and the elevated transmission risk from untreated populations. Phylogenetics has enormous potential to contribute to public health, but remains in its infancy; methods need rigorous assessment and results require cautious interpretation.

List of related peer-reviewed journals and abstracts
Papers:

1) Brown, A.E., Gifford, R.J., Clewley, J.P., et al. Concerted Action on Seroconversion to AIDS and Death in Europe (CASCADE) Collaboration., Phylogenetic Reconstruction of Transmission Events from Individuals with Acute HIV Infection: Toward More-Rigorous Epidemiological Definitions. Journal of Infectious Disease, 2009a. 199(3): p. 427-431.

2) Brown, A.E., Murphy, G., Rinck, G., et al. Implications for HIV testing policy derived from combining data on voluntary confidential testing with viral sequences and serological analyses. Sex Transm Infect, 2009b. 85(1): p. 4-9.

Abstracts:

1) Brown, A.E., Sudarshi, D., Fisher, M., *et al*. *Transmitted HIV Drug Resistance May Originate from Untreated and Undiagnosed Individuals: A Phylogenetic Exploration.* CROI 2009. 16th Conference on Retroviruses and Opportunistic Infections. February 8-11, 2009 Palais des Congrès de Montréal, Montréal, Canada.

2) Fisher, M.J., Sudarshi, D., Brown, A.E., *et al. HIV Transmission among Men Who Have Sex with Men: Association with ART, Infection Stage, Viremia, and Sexually Transmitted Diseases, a Longitudinal Phylogenetic Study.* CROI 2009. 16th Conference on Retroviruses and Opportunistic Infections. February 8-11, 2009 Palais des Congrès de Montréal, Montréal, Canada.

3) Brown, A.E., Gifford, R.J., Porter K., *et al. A phylogenetic exploration of the role of seroconverters in transmitting HIV drug resistant viruses within Europe 2007.* The Fourteenth International Workshop on HIV Dynamics and Evolution will be held at the Paradore de Segovia in Segovia, Spain on April 17 – 20.

4) Brown, A.E., Clewley, J.P., Hue, S., *et al. Do primary HIV-1 infections amongst MSM contribute disproportionately to onward transmissions in the UK: a phylogenetic approach?* 12th Annual Conference of the British HIV Association (BHIVA) Brighton Dome · Brighton · 29 March – 1 April 2006.

Contents

Acronyms

AIDS	Acquired immune deficiency syndrome
ARV	Anti-retroviral therapy
BASHH	British Association of Sexual Health and HIV
BEAST	Bayesian evolutionary analysis by sampling trees
BHIVA	British HIV Association
BS	Bootstrap
CASCADE	Concerted Action on Seroconversion and Death in Europe
CI	Confidence intervals
DH	Department of Health
DNA	Deoxyribonucleic acid
F81	Felsenstein 81 evolutionary model
GTR	General time reversible evolutionary model
GUM	Genito-urinary medicine
HIV	Human immunodeficiency virus
HKY85	Hasegawa, Kishino and Yano 1985 evolutionary model
HPA	Health Protection Agency
IDU(s)	Injecting drug use(rs)
JC	Jukes Cantor evolutionary model
K2P	Kimura's 2 parameter evolutionary model
ML	Maximum likelihood
MSM	Men who have sex with men
NJ	Neighbour joining tree
NNRTI	Non-nucleotide reverse transcriptase inhibitor
NRTI	Nucleotide reverse transcriptase inhibitor
PI	Protease inhibitor
PR	Protease
PYFU	Person years of follow up
RNA	Ribonucleic acid
RT	Reverse transcriptase
SOPHID	Survey of prevalent HIV infections diagnosed
SPREAD	Strategy to control spread of HIV drug resistance
STARHS	Serological testing algorithm for recent HIV seroconversion
STI(s)	Sexually transmitted infection(s)
TDR	Transmitted HIV drug resistance
VCT	Voluntary confidential HIV testing
UA STI survey	Unlinked anonymous survey of attendees of sentinel STI clinics
UAI	Unprotected anal intercourse
UK CHIC	UK Collaborative HIV cohort
UPGMA	Unweighted pair group method with arithmetic means
WRB	World region of birth

1. Chapter One: Introduction

This chapter provides an introduction to HIV infection. The UK HIV epidemic is then described alongside a summary of current transmission and prevention strategies. The basic concepts of phylogenetics are introduced within the context of how they can be applied to enhance understanding of HIV transmission. The gaps in the research are outlined. The chapter ends with the research objectives and a description of the thesis structure.

1.1. About HIV

1.1.1 Human immunodeficiency virus

Infection with Human Immunodeficiency Virus (HIV) is fatal if left untreated. HIV infection causes progressive deterioration of the host's immune system to levels that make an individual susceptible to the development of a range of opportunistic infections – otherwise known as AIDS (Acquired Immune Deficiency Syndrome) (Adler 1987). HIV is transmitted from human to human: through sex (anal, vaginal and oral); vertically (from mother to child); through injecting drug use; and from blood contact with contaminated blood products. Without treatment, HIV-infected individuals typically survive for a median of 10 years following infection (Porter, Babiker et al. 2003).

Whilst antiretroviral therapy (ARV) treatment has transformed HIV from a fatal to a chronic infection, treatment is expensive and only effective provided an assiduous routine of medication is followed indefinitely (Palella, Delaney et al. 1998). ARVs suppress, but do not eradicate the virus. Viral resistance to ARVs can develop, increasing the risk of HIV-related death (Beinker, Mayers et al. 2001; Zaccarelli, Tozzi et al. 2005). Therefore HIV remains a serious, life-long infection.

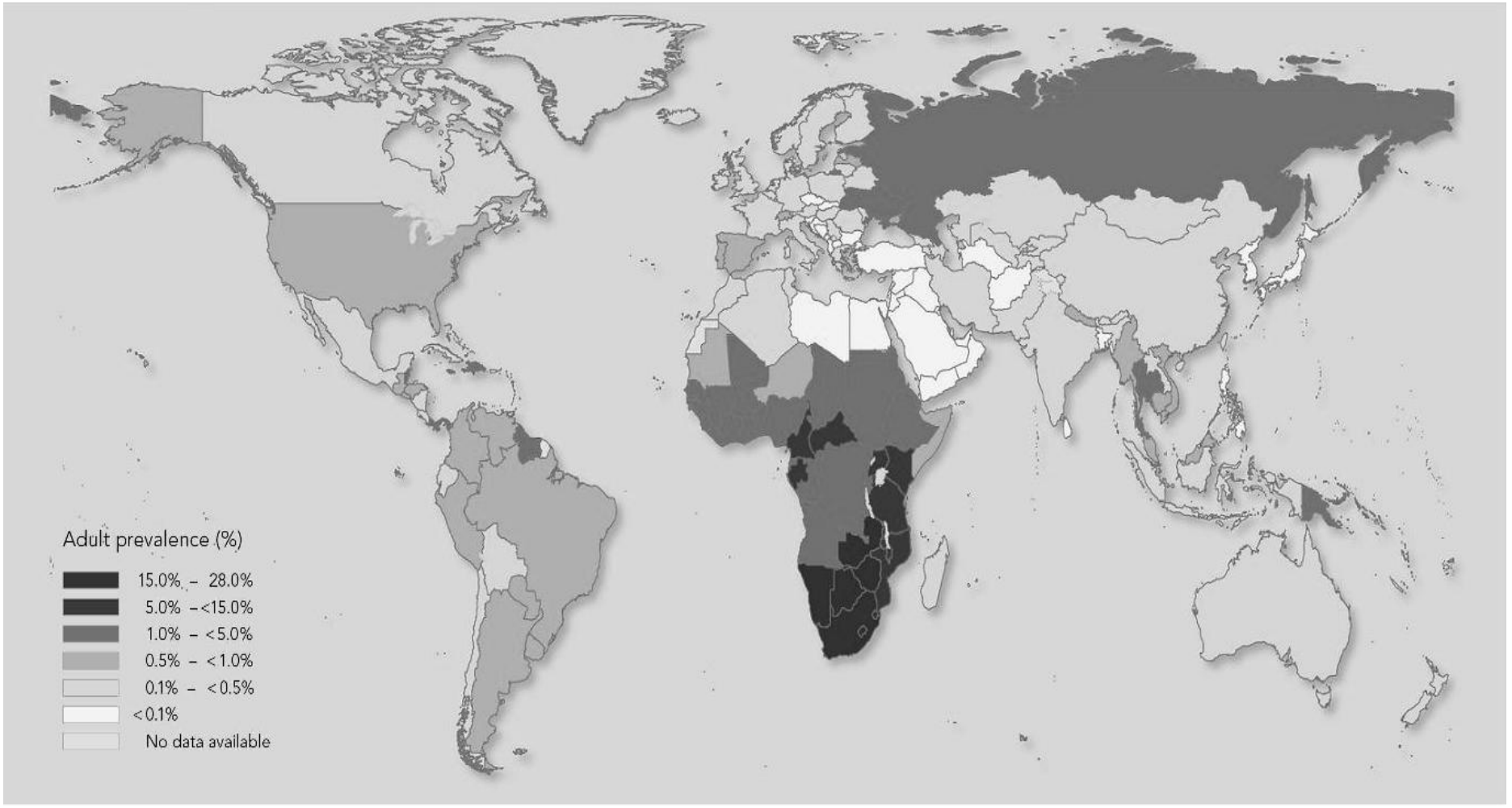
1.1.2 The global HIV epidemic

The first AIDS cases were recognized in the early 1980s (Berridge 1996) among men who have sex with men (MSM) in the United States (Brennan and Durack 1981) and Western Europe (Dubois 1981). Further cases indicated that AIDS could also be acquired through injecting drug use (Masur, Michelis et al.

1981), from mother to child (MMWR 1982) and through heterosexual sex (Clumeck, Sonnet et al. 1984). Subsequently, AIDS was found within Africa (Kamradt, Niese et al. 1985; Mann, Bila et al. 1986). HIV itself was isolated in 1983 (Barre-Sinoussi, Chermann et al. 1983), and found to be the cause of AIDS (Blattner, Gallo et al. 1988).

Almost 30 years later, an estimated 33 million people are living with HIV worldwide **(Figure 1.1**), with the virus responsible for approximately 25 million cumulative deaths (UNAIDS 2008). In 2007, it was estimated that 2.7 million people became infected with HIV and that two million died from HIV-related illnesses. Sub-Saharan Africa remains the focus of the global HIV epidemic. This region accounts for over two-thirds of the people living with HIV, and 75% of HIV-related deaths (UNAIDS 2008). Within sub-Saharan Africa, HIV is primarily transmitted heterosexually, although transmission between MSM also is becoming recognized (van Griensven, de Lind van Wijngaarden et al. 2009). Elsewhere, the key groups affected by HIV remain MSM, injecting drug users and heterosexual migrants who acquired their infection within sub-Saharan Africa.

Figure 1.1: Estimated global HIV prevalence among individuals aged 15-49



Source: adapted from UNAIDS (UNAIDS 2008)

1.1.3 Genetic diversity of HIV

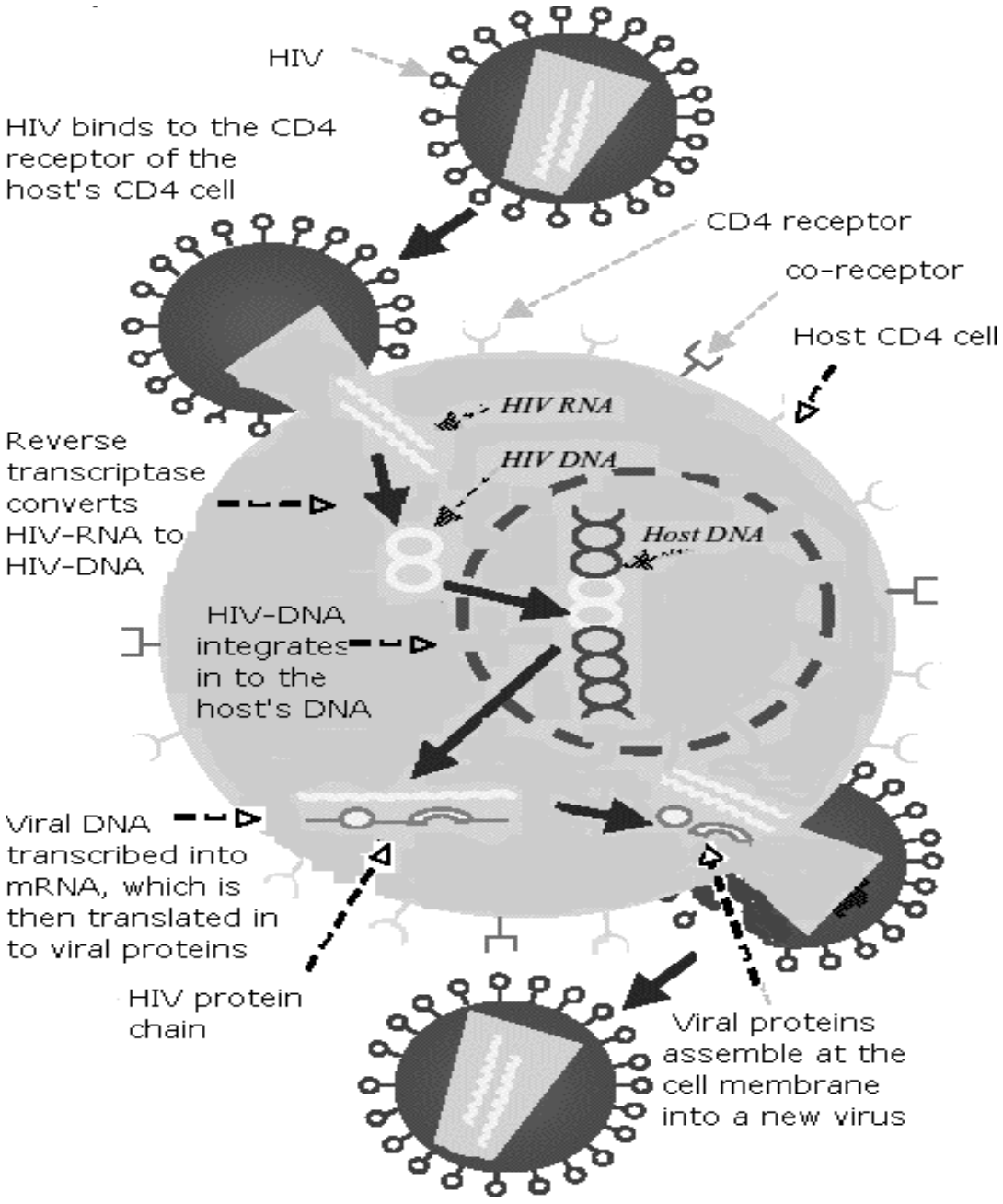
Two types of HIV have been distinguished to date: HIV-1 and HIV-2. HIV-1 accounts for the majority of infections globally, whilst HIV-2 is the more prevalent in west African countries (Poulsen, Aaby et al. 1993). HIV-1 can be further categorized into three groups: M (main), N (new) and O (out-group) based on their genetic similarity (Sharp, Bailes et al. 2001). Group M is responsible for the most global HIV infections. Groups O and N remain restricted to West-Central Africa (Gurtler, Zekeng et al. 1996). Group M viruses are further subdivided into nine subtypes (A-K) due to the substantial genetic diversity within this group (Louwagie, McCutchan et al. 1993). There are also mosaics of subtypes in circulation, named as "circulating recombinant forms" (CRFs) and "unique recombinant forms" (URFs) according to their frequency. HIV subtypes are associated with geographic areas and specific risk groups. On a global scale, the most prevalent HIV-1 clades are subtypes C (47%), A (27.2%), B (12.3%), and D (5.3%) (Osmanov, Pattou et al. 2002). HIV infection has long been associated with subtype B in Western Europe and North America, and most frequently found among MSM and injecting drug users (IDUs). For the rest of this thesis, "HIV" refers to "HIV-1".

1.1.4 Life cycle of HIV

This section describes the life cycle of HIV and is summarized in **Figure 1.2**. Once the virus has been transmitted into the host's body, HIV binds to CD4 receptors on the host's CD4+ T-lymphocyte cells (hereafter called CD4 cells) (Turner and Summers 1999). These cells are a sub-class of T-lymphocytes and have a key role in host immunity. Supplementary interaction with two co-

receptors allows the virus to fuse with host's cell membrane and enter the CD4 cell. The HIV-RNA (ribonucleic acid) is converted into HIV-DNA (deoxyribonucleic acid), generating a provirus, through the release of the reverse transcriptase enzyme.

Figure 1.2: Life cycle of HIV



Source: adapted from

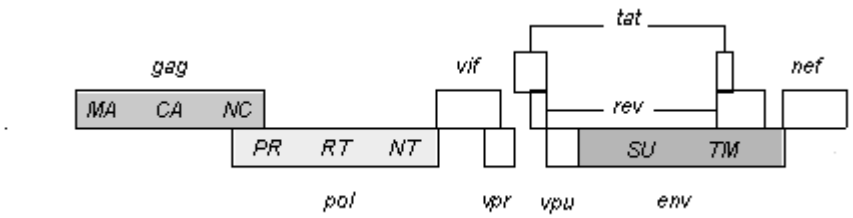
http://aidsinfo.nih.gov/contentfiles/HIVLifeCycle_FS_en.pdf)

The HIV-DNA enters the host cell nucleus where it is integrated into the host's DNA (facilitated by the integrase enzyme). The provirus can then remain latent or be active, generating products for the generation of new virions. Activation of the host cell results in the transcription of integrated viral DNA into messenger RNA (mRNA), which is then translated into viral proteins. Amongst these is HIV protease, which is required to process other HIV proteins into their functional forms. The viral RNA and viral proteins assemble at the cell membrane into a new virus, which is then released and capable of infecting another host cell. The originally infected host cell dies.

1.1.5 HIV genome

The HIV genome consists of approximately 9200 nucleotides of RNA. The RNA contains nine genes encoding for 14 viral proteins **(Figure 1.3**). Three of the genes are major: *gag* (encodes for internal structural proteins); *env* (encodes for transmembrane proteins); and *pol* (encodes enzymatic proteins e.g. reverse transcriptase (RT), protease (PR), and integrase). The remaining six genes are accessory, encoding for regulatory (*tat* and *nef*) and auxiliary (*vif*, *nef*, *vpr* and *vpu*) factors.

Figure 1.3: Organization of the HIV-1 genome



The three main genes are shaded: *gag*; *pol*; and *env.*

Source: adapted from (Watts, Dang et al. 2009)

1.1.6 The course of HIV infection

Following transmission, free virus is released which infect further CD4 cells, in turn leading to increased viral load (a measure of how much HIV is in the body). As more CD4 cells become targeted, the host's immune capability gradually diminishes, leaving the individual susceptible to opportunistic infections. From initial infection to death, the course of HIV infection within an untreated individual can be summarized into four main stages **(Figure 1.4**):

Recently-acquired HIV Infection

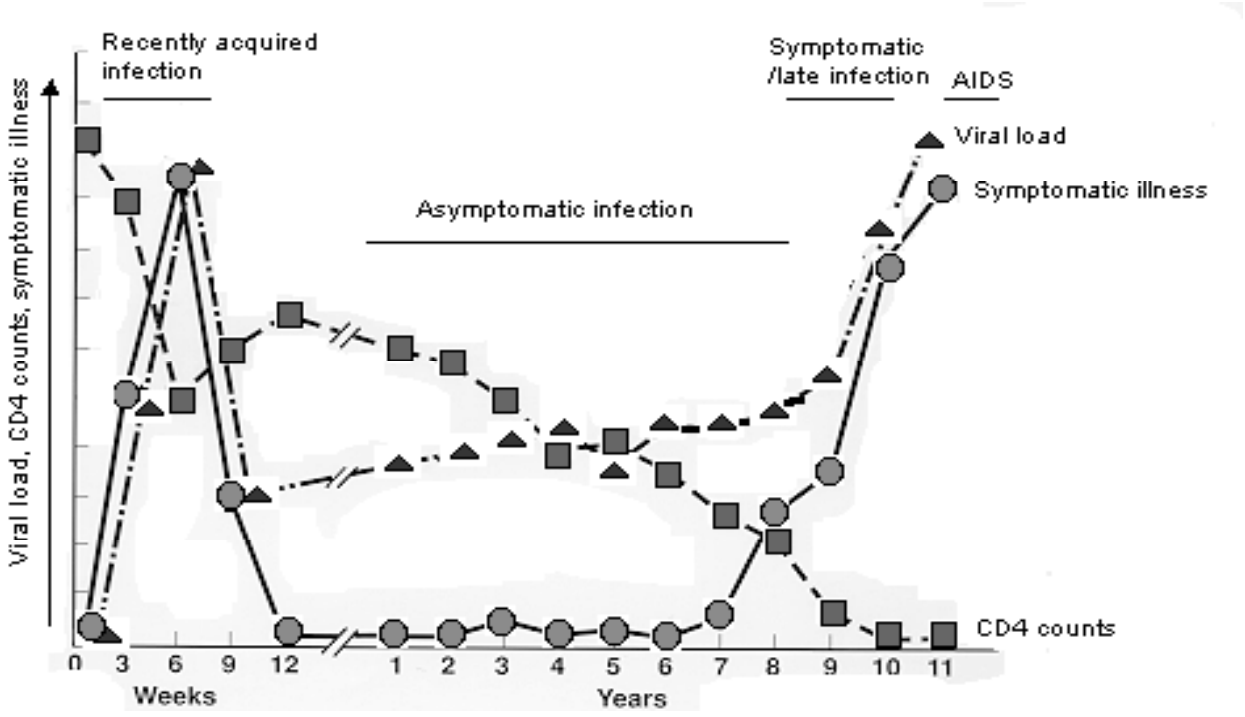
Also known as primary, early or acute infection, recently acquired HIV infection (recent infection) describes the first few weeks following infection. During this time, HIV is present in blood serum and plasma, but a specific antibody response may not have been developed (Pilcher, McPherson et al. 2002) (section 2.3.1). Consequently, this period is characterized by high levels of virus, reaching levels of up to 100 million copies of HIV (RNA/mL). At the same time CD4 cell counts temporarily decrease as a result of cell death following viral replication. Blood virus levels are known to peak at around 17 days, with semen virus levels peaking at around 30 days following infection (Pilcher, Joaki et al. 2007). Between 40-90% of patients experience symptoms of early HIV infection (Kahn and Walker 1998). Once an antibody response has been generated by the host, HIV replication is restrained and CD4 cell counts return to normal levels.

Asymptomatic infection

Asymptomatic or chronic infection describes the period where viral replication is restrained by the antibody response. The infected individual will not usually

have clinical symptoms. Viral replication and CD4 cell turnover remain active, but at low levels, and the immune system gradually weakens over time (Ho, Neumann et al. 1995). The time interval between HIV infection to clinical AIDS has a median of 10 years, but varies considerably between individuals.

Figure 1.4: Schematic diagram of the typical course of HIV infection



Source: adapted from (Pantaleo, Graziosi et al. 1993)

Symptomatic/late infection

As CD4 cells reach very low levels (e.g. under 200 cells/mm^3) patients develop clinical symptoms. It is at this stage that individuals become vulnerable to developing "opportunistic infections" – i.e. infections that do not cause illness in individuals with good immunity. Patients diagnosed with CD4 counts under 200 cells/mm^3 are described as being diagnosed with late stage infection (Phillips and Pezzotti 2004). Such groups have a higher risk of HIV-related mortality

(Chadborn, Baster et al. 2005; Chadborn, Delpech et al. 2006) compared with those diagnosed at an earlier stage of infection.

AIDS

The host's immune system fails allowing the development of opportunistic infections (Adler 1987) – eventually leading to death. In the absence of treatment, survival time after diagnosis of AIDS is 10-12 months on average. The definition of AIDS includes one of a number of indicator diseases in persons with an HIV infection, including Kaposi's sarcoma, extra-pulmonary tuberculosis and *Pneumocystis pneumonia:*

(http://www.eurohiv.org/reports/report_37/aids_euro_definition_eng.pdf accessed 12th August 2009).

1.1.7 Monitoring disease progression

Patients may be infected for many years before their HIV infection is diagnosed. This is because of the long duration of asymptomatic infection. The host's immune response generates antibodies to HIV which are in sufficient quantity for HIV antibody tests to be positive about six weeks following infection (Busch and Courouce 1997). Recently, other diagnostic tools have been used in conjunction with improved antibody tests to recognize HIV as early as two weeks following infection (Busch, Glynn et al. 2005). Diagnostic techniques and markers that can recognize HIV during recently acquired infection are described in more detail in section 2.3.2.

Once diagnosed, two main clinical markers are used to monitor disease progression: blood CD4 cell counts and viral load. CD4 cell counts per mm^3 of

the blood provide an indication of how well an individual's immune system is functioning, since CD4 cells gradually decline over the course of an HIV infection **(Figure 1.4**). In 2003, guidelines recommended that an individual should begin ARV treatment once the CD4 cell count drop to under 200 cells/mm^3 (Pozniak, Gazzard et al. 2003). In 2008, guidelines were updated to recommend treatment discussions should start sooner, once CD4 counts reach between 200-350 cells/mm^3 (Gazzard 2008).

Viral load measures how much HIV is in the body and increases as the host's immune function decreases **(Figure 1.4).** It is strongly associated with transmission risk: the higher the viral load, the higher the risk of transmitting virus per exposure (Castilla, Del Romero et al. 2005) (section 1.1.10).

1.1.8 Antiretroviral therapy (ARVs)

ARV drugs prevent HIV infection causing progressive damage to the host immune system through inhibiting viral replication and thus clinical progression. The drugs inhibit key stages of the HIV lifecycle and are classified accordingly: fusion inhibitors; reverse transcriptase inhibitors; integrase inhibitors; and protease inhibitors (Zeniga 2008).

Fusion inhibitors block HIV from fusing to the host cell's membrane. RT inhibitors include nucleotide reverse transcriptase inhibitors (NRTIs) and non nucleotide reverse transcriptase inhibitors (NNRTIs). Both work by blocking 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. Provirus is

prevented from integrating into the DNA of the host cell through the inhibition of integrase. Protease inhibitors (PIs) bind to the active site of PR, thereby preventing the production of viral proteins for the final assembly of new virions.

Since 1996, in North America and Western Europe, the provision of ARVs has been widespread. Patients treated with ARVs have been shown to achieve reductions in viral load to very low levels, over prolonged periods of time, provided they have good adherence to treatment (Montaner, Reiss et al. 1998; Friedland and Williams 1999). The advent of treatment has substantially reduced HIV-related mortality in these settings (Mocroft, Vella et al. 1998) (Palella, Delaney et al. 1998; Bhaskaran, Hamouda et al. 2008). However, such medication is only successful if followed indefinitely (Palella, Delaney et al. 1998) and is not effective for all patients due to the adverse side effects (Lucas, Chaisson et al. 1999) and the presence of viruses resistant to the treatment (Beinker, Mayers et al. 2001).

1.1.9 HIV drug resistant viruses

HIV drug resistant viruses contain mutations that reduce the susceptibility of the HIV to ARV. ARV inhibits key stages of the HIV replication cycle (Zeniga 2008) delaying or even preventing disease progression (Hammer, Eron et al. 2008). Consequently, HIV-infected individuals with viruses resistant to ARVs are less likely to be treated successfully (Hirsch, Brun-Vezinet et al. 2003), and have an increased risk of HIV-related death (Beinker, Mayers et al. 2001), compared to those with viruses that did not contain drug resistant mutations (wild-type viruses).

Drug resistance to ARV can be “acquired” or “transmitted”. Acquired resistance develops following treatment failure in patients with suboptimal adherence to ARV. Transmitted drug resistance occurs through infection with a resistant strain, referred to as TDR. Both acquired and TDR have adverse consequences for the success of treatment. However, TDR has the potential to reverse the effectiveness of ARV more rapidly (Little, Holte et al. 2002). The difficulties associated with measuring HIV drug resistance and definitions of drug resistance are detailed in section 2.5.2.

1.1.10 Transmission rates from HIV-infected patients.

Transmission rates have been estimated through analyzing transmission between monogamous serodiscordant couples (where one partner is HIV positive and the other HIV negative) followed up over time. The probability of transmission per sexual act has been found to be closely correlated to the HIV-infected partner’s plasma viral load, and correspondingly, their infection stage and treatment status. Such studies are summarized in **Table 1.1.**

In 2000, Quinn *et al*. (Quinn, Wawer et al. 2000) first demonstrated the relationship between plasma viral load and HIV transmission. Overall, 90 HIV negative partners of the 415 heterosexual serodiscordant couples became HIV-infected over a period of 30 months. The mean plasma viral load in the HIV positive partner was found to be significantly higher among couples whose initially HIV negative partners became infected during the study period compared to those where partners remained negative (90,254 copies/mL vs. 38,029 copies/mL, p=0.01). Importantly, no transmission was observed within couples where the HIV positive partner had a plasma viral load under 1500

copies/mL. Similar results were later presented by Wawer *et al*. (Wawer, Gray et al. 2005) and Gray *et al*., (Gray, Wawer et al. 2001). The latter found that, out of 43 serodiscordant couples where the HIV positive partner's plasma viral load was <1700 copies/mL, one initially HIV negative partner became infected.

In Thailand, (Tovanabutra, Robison et al. 2002) it was found that each log increment in plasma viral load was associated with an 81% increase in the risk of transmission. No transmission occurred between 17 couples where the positive partner had a plasma viral load <1094 copies/mL. Castilla *et al*. (Castilla, Del Romero et al. 2005) studied 393 HIV serodiscordant heterosexual couples between 1991-2003 to identify the proportion of transmission that occurred from ARV-treated patients (specific regimen not detailed). While no transmissions occurred when the positive partner was ARV-treated, HIV was found to persist in genital secretions. A review was undertaken of heterosexual transmissions occurring as a result of natural pregnancies among 62 HIV serodiscordant couples in Spain between 1998-2005 (Barreiro, del Romero et al. 2006). All the HIV-infected partners received ARV and their median plasma viral load at around conception was <500 copies/mL. None of the initially HIV negative partners became infected.

While the studies indicate minimal transmission among the virally suppressed, it is important not to over interpret the results due to the small numbers of HIV serodiscordant couples followed up; this is reflected in the upper 95% confidence limits (**Table 1.1**). Additionally, whilst no transmission occurred where the positive partner was treated, the viral presence within genital

secretions means the potential for HIV transmission cannot be discounted. Finally, patients fully adherent to ARVs may experience transient low-level viral rebound (blips) (Garcia-Gasco, Maida et al. 2008). The transmission risk from blips is not clear.

However, the relationship between viral load, treatment and transmission is so strong that a recent report controversially suggested that HIV serodiscordant heterosexual couples may have unprotected sex, provided the HIV-infected partner is successfully treated with plasma viral load suppressed to <40 copies/mL (Vernazza 2008). The report has been criticized as inconclusive and dangerous, jointly by the World Health Organization and UNAIDS (http://data.unaids.org/pub/PressStatement/2008/080201_hivtransmission_en.pdf, accessed 12th August 2009), and the American Centers for Disease Control (http://www.cdc.gov/hiv/resources/press/020108.htm, accessed 12th August 2009). Additionally, other factors can increase the risk of transmission, including the presence of sexually transmitted infections (STIs) (Wasserheit 1992) the type of sexual contact (anal, (Lane, Pettifor et al. 2006) vaginal or oral (Gilbart, Evans et al. 2004)).

Table 1.1: Summary of studies examining the effect of viral load and/or ARV on sexual HIV transmission between serodiscordant couples

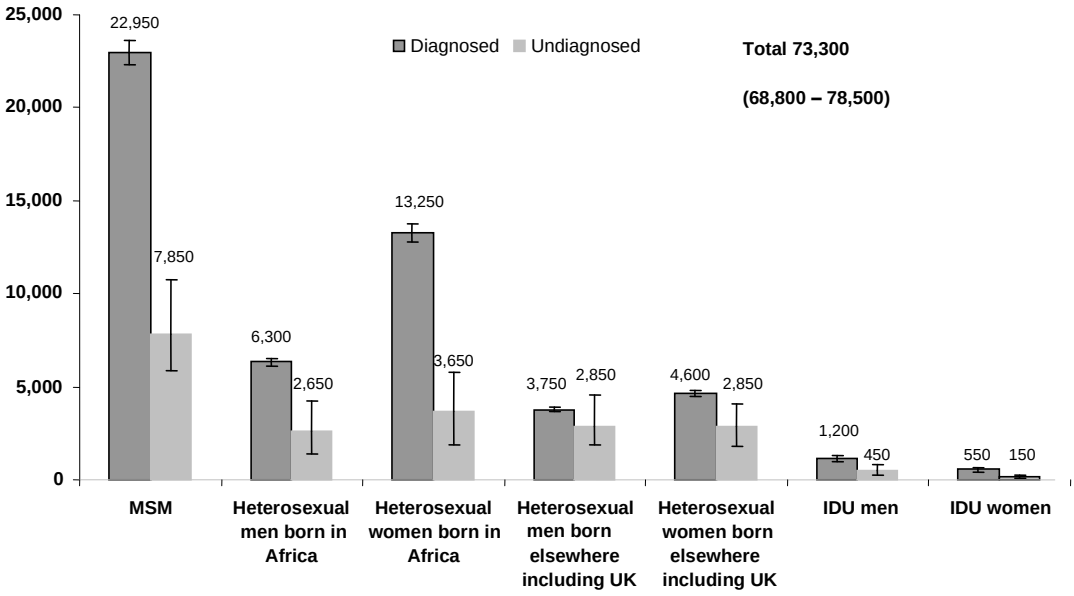
Author, year, reference	Study aim	Setting: Country, Years sampled	Number of HIV serodiscordant couples	"Viral load suppression" definition	ARV details	Total proportion of HIV transmissions	Proportion of transmissions from virally suppressed/ARV compliant HIV+ partner (n/N*, 95% CI))	Comments
Impact of viral load on HIV transmission								
Quinn, 2000, (Quinn, Wawer et al. 2000)	Examine impact of viral load on HIV transmission	Rakai, Uganda, 1994-1998	415	<1500 copies/mL	Untreated	21.7% (90/415)	0% (0/51, 0-0.07)	76.7% of transmissions occurred from couples where infected partner had a viral load >10,000
Gray, 2001,(Gray, Wawer et al. 2001)	Calculate probability of HIV transmission per coital act	Rakai, Uganda, 1994-1998	174	<1700 copies/mL	Untreated	21.8% (38/174)	2.3% (1/43, 0.0041-0.12)	Genital ulceration and viral load major determinates of HIV transmission
Fideli, 2001,(Fideli, Allen et al. 2001)	Compare biological determinants between transmitting and non-transmitting serodiscordant couples	Lusaka, Zambia, 1994-2000	311	<10,000 copies/mL	Untreated	33.4% (104/311)	15.2% (8/52, 0.08-0.28)	Viral load more important determining factor in female to male transmission than vice versa
Tovanabutra, 2002,(Tovanabutra, Robison et al. 2002)	Evaluate association between HIV viral load and transmission	Northern Thailand1992-1998	493	<1094 copies/mL	Untreated	44.2% (218/493)	5.9% (1/17, 0.011-0.27)	STI diagnosis and earlier first sex associated with transmission
Wawer, 2005,(Wawer, Gray et al. 2005)	Estimate rates of HIV transmission per coital act	Rakai, Uganda, 1994-1999	235	<3.45 log	Untreated	28.9% (68/235)	0% (0/57, 0-0.63)	Rate of transmission highest during early- and late-stage infection
Impact of ARVs on HIV transmission								
Musicco, 1994 (Musicco, Lazzarin et al. 1994)	Determine effect of ZDV on HIV transmission	Italy, 1987-1992	436	Not collected	ZDV daily	6.2% (21/436)	9.4% (6/64, 0.044-0.19)	ZDV used more frequently among patients with advanced disease
Castilla, 2005(Castilla, Del Romero et al. 2005)	Determine if ARV can diminish risk of HIV transmission within serodiscordant couples	Madrid, Spain, 1991-2003	393	Not collected	a) No ARV 1991-1995 b) ARV 1999-2003	a) 7.4 (29/393) 1991-2003	b) 0% (0/52, 0-0.0069) 1999-2003	HIV persisted in genital secretions among ARV treated
Barreiro (Barreiro, del Romero et al. 2006)	Review natural pregnancies among HIV serodiscordant couples	Spain, 1998-2005	62	<500 copies/mL	ARV (various regimens)	N/A	0% (0/62, 0-0.058)	Length of follow-up not specified

1.2. HIV in the UK

1.2.1 The UK HIV epidemic

In the UK, HIV infection remains a significant public health concern more than 25 years after the first AIDS cases were reported (O'Connor, McEvoy et al. 1983). An estimated 77,400 people were estimated to be living with HIV in the UK in 2007, of whom over one quarter were unaware of their infection **(Figure 1.5**). The largest groups living with HIV in the UK are MSM and heterosexuals born in sub-Saharan Africa (HPA 2008).

Figure 1.5: Estimated number of adults (15-59) living with HIV, UK: 2007

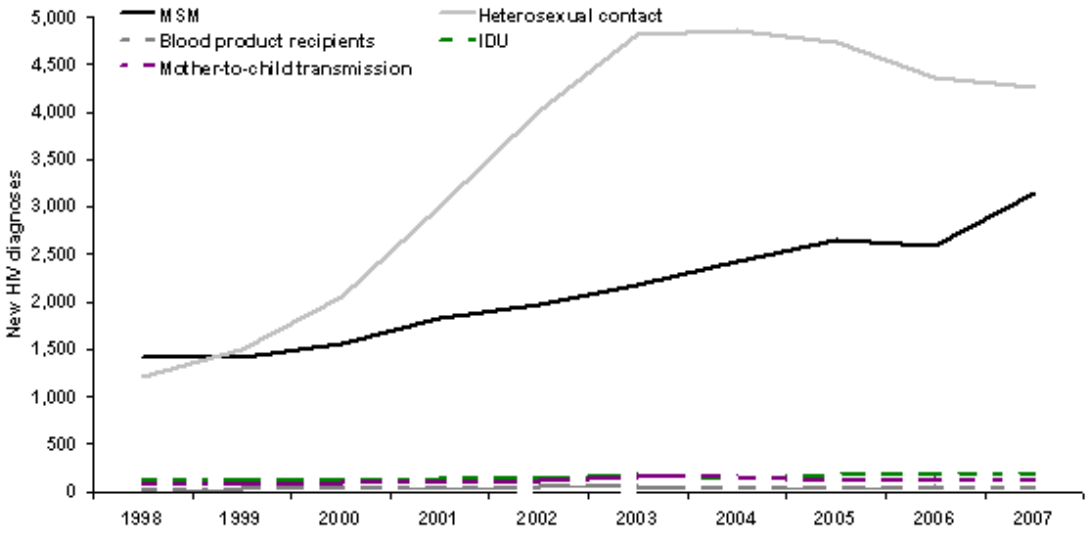


Source: (HPA 2008)

The number of new HIV diagnoses reported in the UK continues to rise, with 7,734 new diagnoses reported in 2007. While MSM are the group most affected by HIV in the UK, over the last decade, the number of new HIV diagnoses reported among heterosexuals has exceeded the number reported among MSM

(**Figure 1.6**). Of the 4260 diagnoses reported among heterosexuals in 2007, 3300 are thought to have acquired their infection in Africa.

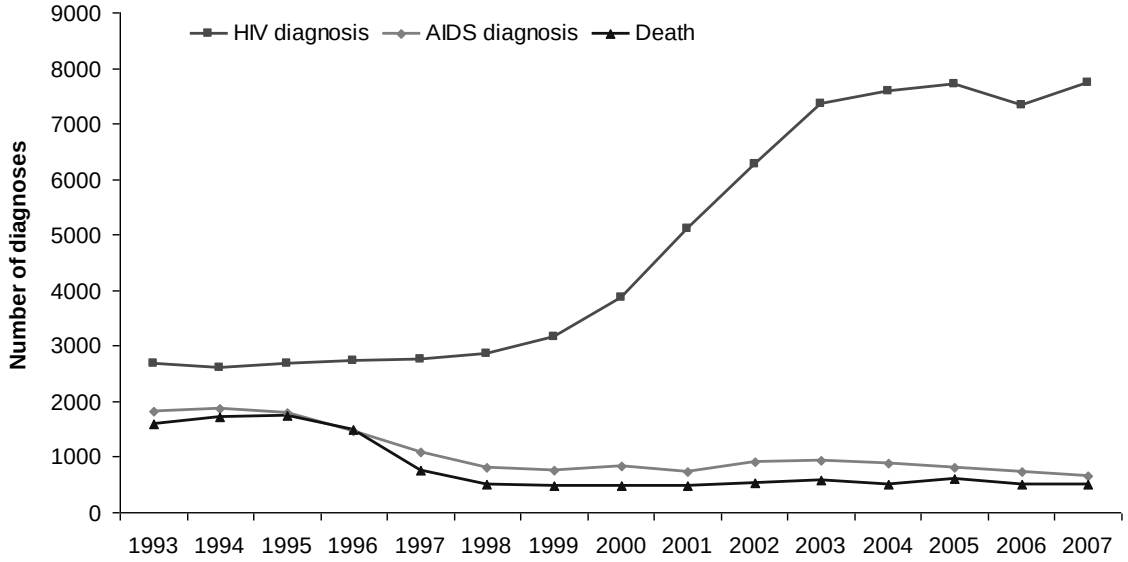
Figure 1.6: Adjusted number of new HIV diagnoses by risk group, UK: 1998-2007



Source: (HPA 2008)

In the UK, ARVs have been widely available as an effective HIV treatment since 1996 (Montaner, Reiss et al. 1998). The late 1990s consequently saw a sharp decrease in the number of AIDS diagnoses and deaths **(Figure 1.7)**. Since this time, the number of deaths has remained stable, and the number of AIDS diagnoses continued to decline. This, combined with the rising number of new HIV diagnoses, means an increasing number of people are living with diagnosed HIV infection in the UK. In 2007, there were 56,556 diagnosed patients accessing care in the UK. This is a three-fold increase since 1998.

Figure 1.7: HIV diagnoses, AIDS case reports and deaths in HIV-infected individuals, UK: 1993-2007



Source: (HPA 2008)

1.2.2 Resistance to antiretroviral therapy

Studies of drug resistance in the UK have found a wide range in the prevalence of transmitted (TDR) and acquired drug resistance. This is due to: differences in the definitions of drug resistance used; population sampling, and the time between infection and the time the sample was taken for resistance testing (see section 2.5.2).

Estimates of the prevalence of TDR among drug naïve patients diagnosed in the UK have varied between 10-20%. More recent data suggest that the prevalence of TDR may be declining (Cane, Chrystie et al. 2005; UKCollaborativeGroup 2007). Prevalence of TDR was measured at 8% in 2004. However trends are difficult to interpret, since increasing numbers of

patients have their virus tested for resistance mutations at around diagnosis due to national recommendations (Pozniak, Gazzard et al. 2003). Consequently, the decrease could be a consequence of the changing attributes of the population denominator.

1.2.3 HIV diversity in the UK

In the UK, subtype B is the most frequent HIV subtype, corresponding to that circulating among MSM and IDUs in Western Europe and North America. However, the increasing number of new HIV diagnoses among heterosexuals infected in Africa has led to greater subtype diversity within the UK. Among HIV-infected heterosexuals attending sentinel STI clinics in England, Wales and Northern Ireland between 1997-2000, subtype C accounted for the majority (32%). Subtypes B (29%), and A (12%), CRFs (9%), and URFs (8%) and subtypes D-H (8%) were also detected. Heterosexuals with non-B subtypes were more likely to have been infected within sub-Saharan Africa, reflecting the distribution of subtypes found outside the UK (Tatt, Barlow et al. 2004).

More recently, a large scale phylogenetic analysis (see section 1.4) of HIV *pol* sequences among 5675 patients diagnosed between 1996-2004 in the UK (Gifford 2007) found that 74% were infected with subtype B, with subtypes A and C occurring at 6% and 10% respectively (HIV exposure by subtype was not presented).

1.2.4 HIV transmission in the UK

It is unclear whether HIV transmission is increasing within the UK. This is because surveillance data are difficult to interpret. For instance, the number of

new HIV diagnoses is determined not only by HIV transmission rates, but also the relative uptake of HIV testing among high risk populations and migration of individuals from high prevalence countries. The relative contribution of these factors is difficult to disentangle (Dougan, Elford et al. 2007). The annual HIV incidence among MSM attending sentinel STI clinics in England, Wales and Northern Ireland has been estimated at around 3% since 1995 (Murphy, Charlett et al. 2004). However, the representativeness of this high risk population and its associated biases (e.g. people with recent risk behaviour are more likely to be included in the denominator of patients attending STI clinics (Burchell, Calzavara et al. 2003)) mean such data need to be interpreted with caution (see section 2.3.5). Nevertheless, the continued high rates of STIs diagnosed among MSM, including among those with a diagnosed HIV infection, all suggest that if transmission is not increasing, it is certainly continuing even within the context of high uptake of ARVs.

1.3. Preventing HIV transmission: the UK response

1.3.1 Preventing HIV transmission

Prevention has remained an essential component of the public health response to the UK HIV epidemic. The sexual health strategy of 2001 (DH 2001) had two major initiatives: to increase HIV testing (to reduce undiagnosed HIV infection) and to change sexual risk behaviour (to prevent transmission from the HIV-infected population).

Voluntary confidential HIV testing (VCT) aims to reduce HIV transmission through reducing the proportion of HIV-infected individuals who are unaware of

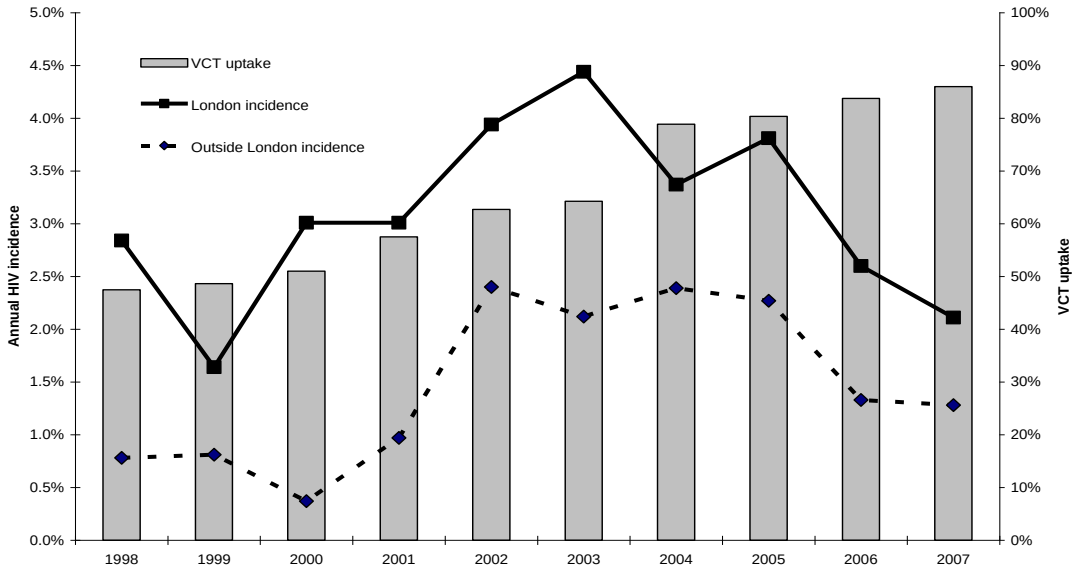
their HIV infection. An HIV diagnosis provides the opportunity for ARV for those with low CD4 counts (which reduces viral load and consequently infectivity) (Vernazza, Troiani et al. 2000), partner notification, and behaviour change counselling. Best practice guidance suggest VCT should be offered to all at their first attendance at a sexual health clinic, and subsequently according to risk (BHIVA 2006). Updated guidelines in 2008 (BHIVA 2008) also suggest those with recent risk exposure be targeted in conjunction with testing with fourth generation diagnostic tests.

Prevention initiatives aim to reduce HIV transmission through promoting safer sex. This means reducing: unprotected sex; number of sexual partners; and concurrent partnerships. Large scale health promotion campaigns have been launched by England's Department of Health and media focused such as: the "Don't die of ignorance" campaign in the 1980s (DH 1987) and more recently, the Sex Lottery (http://www.dh.gov.uk/en/Publicationsandstatistics/Pressreleases/DH_4025977 accessed 17th August 2009). Other initiatives are commissioned at local levels within the NHS, or funded through charities or non-governmental organizations.

1.3.2 Impact of HIV prevention initiatives in the UK

Superficially, the promotion of VCT has been successful. Between 1998-2007, the proportion of MSM attending sentinel STI clinics receiving HIV tests increased from 46 to 86%. However, this has not translated directly into a prevention success. **Figure 1.8** plots the uptake of VCT against annual HIV incidence; the increase in VCT has not been mirrored by a significant decrease in HIV incidence. This may be due to five reasons which are discussed in turn.

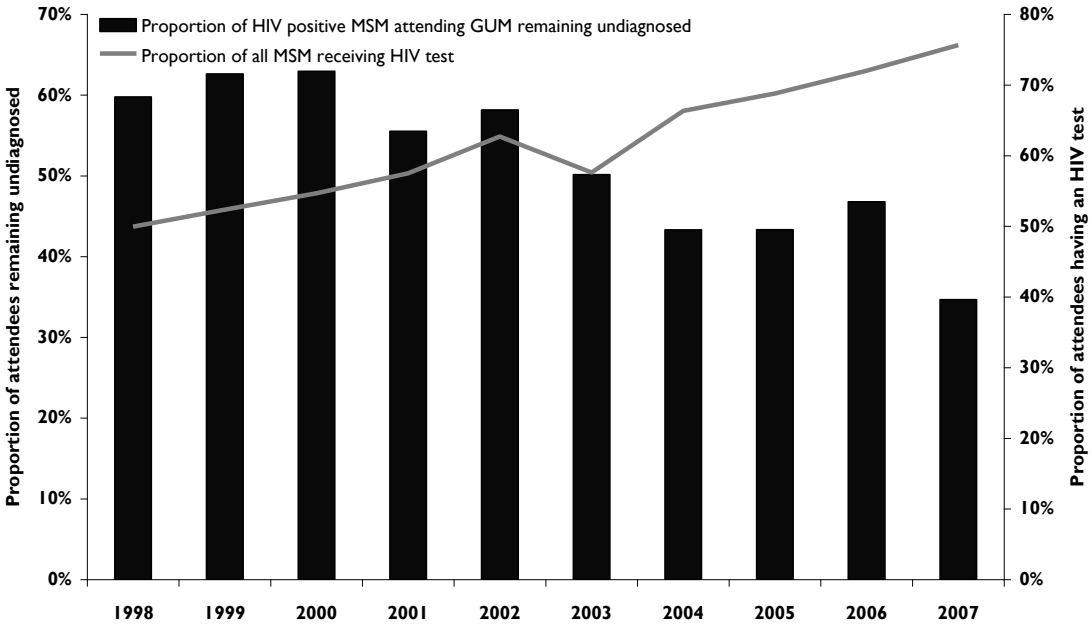
Figure 1.8: The proportion receiving HIV tests and the annual HIV incidence among MSM attending sentinel STI clinics, England, Wales and Northern Ireland: 1998-2007



Source: Personal communication, Unlinked Anonymous STI survey.

Firstly, those receiving HIV tests may not necessarily be those at highest risk of having an HIV infection. Of HIV-infected MSM who arrived at sentinel STI clinic unaware of their infection in 2007, 30% (223/749) left the clinic without an HIV diagnosis because they were either not offered, or had declined an HIV test **(Figure 1.9**). Importantly, an audit found that those arriving at clinics with a recent risk exposure were more likely to defer VCT due to concerns that testing soon after exposure may not provide accurate results (Munro 2007); reattendance rates were low. The failure to pick up those most at risk limits the success of an overall increase in uptake, and is also largely unnecessary due to improvements in testing technology (section 2.3.2).

Figure 1.9: The proportion of MSM attending sentinel STI clinics receiving HIV tests and the fraction of HIV-infected MSM remaining undiagnosed, UK: 1998-2007



Source: (HPA 2008)

Secondly, VCT will only work as a prevention strategy if patients modify their sexual behaviour once they have become diagnosed. However, evidence suggests that the risk behaviour continues even after diagnosis. For instance, Dodds *et al* (Dodds, Mercey et al. 2004) found approximately half of MSM with a diagnosed HIV infection had had unprotected anal intercourse (UAI) within the past 12 months. Risky sexual behaviour among the HIV-diagnosed is thought to be facilitated by websites that allow MSM to seek partners for UAI (Elford, Bolding et al. 2007). The continuation of STI diagnoses among diagnosed HIV-infected men is further evidence that unprotected sex is continuing within this population (Simms, Fenton et al. 2005; Dodds, Johnson et al. 2007).

It is speculated that risk behaviour among the HIV diagnosed population may continue as a consequence of widespread ARVs (Elford and Hart 2005). A meta-analysis found that while being on ARVs and achieving an undetectable plasma viral load did not increase sexual behaviour risk (Crepaz, Hart et al. 2004), peoples' beliefs about ARV (e.g. that it has changed HIV from a fatal disease to a chronic, treatable infection) may have facilitated the continuation of unprotected sex within this population.

Thirdly, while patients with a diagnosed HIV infection have the opportunity to receive ARV treatment, which reduces their infectivity, ARV is prescribed on the basis of individual clinical need and not for public health benefit (section 1.3.1). It was recommended that patients commenced treatment once CD4 counts reached under 200 cells/mm^3 in 2003 (Pozniak, Gazzard et al. 2003), and updated to under 350 cells/mm^3 during 2008 (Gazzard 2008). In 2007, it was estimated that 30% of the diagnosed HIV-infected UK population receiving care were not prescribed ARV (personal communication – SOPHID). Therefore, HIV diagnosis does not necessarily lead to viral load reducing therapy, and many patients with diagnosed HIV infection may continue to have elevated levels of virus.

Fourthly, if the recently HIV-infected are disproportionately generating HIV transmission, then the impact of VCT on reducing transmission will be limited. This is because those with recently acquired HIV infection will not yet have had the opportunity to receive an HIV test.

Finally, VCT promotion may not be implemented on a large enough scale. The prevalence of an infection drives onward transmission, and increases in the prevalent pool of diagnosed and undiagnosed infection may limit the ability HIV testing to make an impact. For instance while the proportion of HIV-infected MSM leaving the clinic remaining unaware of their infection has reduced, the increase in the number of MSM living with HIV has meant the absolute number has increased (Brown, Tomkins et al. 2006). Therefore many more successfully targeted tests are needed to order to make an impact.

1.3.3 Improving interventions to prevent HIV transmission

Whilst the promotion of VCT remains essential for the clinical outcome of individual patients, its potential to reduce current levels of HIV transmission appears uncertain. The potential barriers to VCT's success provide important clues as to why HIV transmission may be continuing in the UK.

In order for prevention interventions to be better targeted, it is important to determine:

- whether individuals with recently-acquired HIV infection are disproportionately driving new infections (including TDR);
- the extent that new HIV infections are generated by the diagnosed HIV-infected population;
- whether reducing the viral load of the HIV diagnosed population through treatment could substantially impact upon population level transmission.

This may be achievable through application of phylogenetic techniques that can reconstruct transmission events at the population level. If such events can be linked to groups/risk factors that may be important in generating transmission (e.g. the recently HIV-infected, and/or diagnosed population/viral load and/or presence of STIs) then the groups/risk factors important in spreading HIV within the UK could be correctly identified and targeted for prevention initiatives.

1.4. Combining phylogenetics with clinical, diagnostic and surveillance data for public health purpose

Phylogenetic analyses involve exploiting small differences in DNA which are combined with computational methods to calculate the degree of relatedness between organisms (see section 2.2). To date, phylogenetic analyses of HIV sequences have been conducted to: explore genetic diversity (Perrin, Kaiser et al. 2003); to aid vaccine development (Novitsky, Smith et al. 2002); to supplement evidence in criminal investigations of HIV transmission (Leitner 2000), (de Oliveira, Pybus et al. 2006); and to time calibrate phylogenies to date the introduction of specific HIV strains into populations or countries (Hue, Pillay et al. 2005; Drummond, Ho et al. 2006) through calculating the rate of mutations. Phylogenetics has also been used to reconstruct transmission events between HIV-infected individuals (Leitner, Escanilla et al. 1996), by assuming that sequences very closely related to each other may have been transmitted between those infected with these strains (thereby gauging “who infected who” (Leitner, Escanilla et al. 1996). This thesis concentrates upon the latter application of phylogenetics.

Within the past decade, such analyses have become more frequent. This is because since 2003, guidelines (Pozniak, Gazzard et al. 2003) recommended that patients newly diagnosed with HIV infection have their virus DNA sequenced to identify drug resistance mutations to inform treatment options (Pozniak, Gazzard et al. 2003). This has resulted in the accumulation of HIV *pol* sequences, which are believed to be appropriate for phylogenetic reconstruction (Hue, Clewley et al. 2004).

Phylogenetic reconstructions of HIV transmission events can become more powerful when they are combined with clinical and diagnostic data. Demographic, laboratory and clinical data are collected routinely around the time of HIV diagnoses for medical and surveillance purposes. Demographic data (e.g. ethnicity, age-group) and clinical data (e.g. sexual orientation, co-infection with sexually transmitted infections (STIs) etc) allow the ascertainment of risk factors associated with infection. Additionally, recently-acquired HIV infections among MSM can be distinguished through laboratory algorithms such as the Serological Testing Algorithm for Recent HIV Seroconversion (STARHS) (Janssen, Satten et al. 1998; Murphy and Parry 2008). The combination of such data has the potential to enhance our understanding of HIV transmission and its associated risk factors to a greater extent than can be gleaned from each data source when considered individually. This thesis combines HIV *pol* sequences with demographic, laboratory and clinical data to enhance understanding of HIV transmission between MSM within the UK, with the ultimate aim of better targeting prevention interventions.

1.5. The current literature

1.5.1 Studies that explore transmission from patients with recently-acquired infection

Phylogenetic reconstructions of HIV transmission events have frequently been undertaken among populations of MSM who were diagnosed soon after HIV infection. This is because of the interest in this group with a high onward transmission potential (Pinkerton 2007). Additionally, such datasets are relatively easy to obtain since they are frequently generated as a by-product of studies of transmitted drug resistance (Pozniak, Gazzard et al. 2003; Hue, Clewley et al. 2004) (section 1.5.3).

Previous phylogenetic reconstructions of HIV transmission events using data exclusively from patients with recent HIV infection have concluded that such populations have an elevated transmission potential. For instance, Pao *et al.* (Pao, Fisher et al. 2005) used phylogenetic analysis of HIV *pol* sequences to identify possible transmissions between recently HIV-infected MSM. Out of 103 individuals with recent infection, 35 possible transmission clusters were observed. Yerly *et al*. (Yerly, Kaiser et al. 1999) performed an exploratory phylogenetic analysis of individuals with recent HIV infection diagnosed between 1996-9 in Switzerland. Of 197 sequences, 29% formed a transmission cluster with at least one other sequence.

1.5.2 Analyses that "measure" the level of transmission

Brenner *et al.* attempted to measure the extent that the recently HIV-infected were generating new transmission events in Canada (Brenner, Roger et al.

2007). Phylogenetic analysis was performed to identify instances of possible transmission events. The proportion of new "transmission events" from the recently HIV-infected (n=696) and the chronically HIV-infected (n=795) were compared: they concluded that 49% of transmission events were generated by the recently HIV-infected compared with 27% from the chronically HIV-infected. In contrast, the recently HIV-infected were estimated to make up 10% of the HIV-infected population. This provided support that the recently HIV-infected are disproportionately generating transmission.

1.5.3 Analyses of the transmission of drug resistant viruses

The majority of phylogenetic explorations have focussed upon transmitted drug resistance and consequently, the recently HIV-infected populations. This ensures that the study population is truly drug naïve and increases the detection of resistant viruses. This is because patients are infected with a small population of viruses; resistant strains are rapidly outgrown by wild-type viruses shortly following infection due to a reduced reproductive fitness of resistant strains (section 2.5.2) (Devereux, Youle et al. 1999; Yerly, Junier et al. 2009).

Phylogenetic analyses of sequences from recently HIV-infected individuals have been conducted to identify possible transmission events between persons with viruses that have identical drug resistance mutations – indicating transmission from HIV-infected individuals with drug resistant viruses. Through this technique, Pao *et al*. (Pao, Fisher et al. 2005) found two clustering sequences from recently HIV-infected MSM that shared the same drug resistance mutation. Yerly *et al* found that whilst 29% of sequences were phylogenetically linked to at

least one other sequence, and 8.8% (17) of individuals had a drug resistance mutation, none of the possible transmission events contained sequences with identical mutations (Yerly, Kaiser et al. 1999). Later, Yerly *et al* (Yerly, Junier et al. 2009) found that patients whose sequences had drug resistant mutations were more likely to form transmission events than those without, and that the prevalence of NNRTI mutations increased between 2000-2008.

However, the extent to which the acquired and transmitted resistant populations are generating onward infections is not known. Leigh-Brown *et al* estimated through modelling methods that 30% of those with acquired resistance may pass their infection onwards (Leigh Brown, Frost et al. 2003). Kearney *et al*. (Kearney, Maldarelli et al. 2009) compared the spectrum of drug resistance mutations between the transmitted and acquired populations and found a greater distribution in the latter group; this may suggest that the populations with acquired resistance are not the only source of TDR, and persistent mutations in the TDR population may continue to spread.

More recently, studies have examined how transmission rates of specific drug resistance mutations vary. Examining *pol* sequences from patients recently HIV-infected at diagnosis, Brenner *et al*. (Brenner, Roger et al. 2008) demonstrated that while the overall frequency of drug resistant viruses between non-clustered and clustered transmission was similar – 14.3% and 16.5% respectively, the frequency of specific classes of mutations differed between clustering sequences and non-clustering sequences. For instance, virus harbouring mutations to NRTIs were less like to cluster than those viruses with

NNRTIs. This further illustrates that it may be the specific mutations, and not necessarily the infection category of the bearer that determines onward transmission of TDR.

1.6. Limitations

Phylogenetic analyses that combine clinical, diagnostic, and demographic data are reliant on many assumptions that require further consideration.

1.6.1 Accuracy of phylogenetics

There is much debate about the suitability of phylogenetics for reconstructing HIV transmission events: it is argued that phylogenetic reconstructions are insufficient to be used as evidence in criminal cases (Bernard 2007). This is because viral genomes can be extremely similar, through parallel or convergent evolution, making it difficult to prove definitively that two viruses have a recent common origin. Additionally, even if two individuals shared a closely related virus, phylogenetic analysis could not verify the direction of transmission and it is difficult to rule out both individuals being infected via a third partner, or that a third party was an intermediary partner between the two. In Leitner *et al.*'s phylogenetic analysis of transmission events, one out of 13 transmission events was incorrectly reconstructed (Leitner, Escanilla et al. 1996).

Epidemiological data on sexual partnerships between HIV-infected patients is rarely available to validate the results empirically (and if available would arguably render phylogenetic analysis unnecessary). It is assumed that phylogenetic reconstructions are sufficiently accurate for analyses for public health purposes, particularly if conservative cut-offs for possible transmission

events are applied (e.g. bootstrap (a measure of statistical support – see section 2.2.7) support of 99% or more and a genetic distance of less than 0.015 nucleotide substitutions per site (Hue, Clewley et al. 2004), see section 2.2.11). Studies to date have not necessarily used such conservative cut-offs (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007).

However, the robustness of phylogenetic methods for population level, public health applications has not been rigorously assessed. Population level studies frequently use samples derived from MSM with diagnosed HIV infection attending a particular health care site for HIV treatment and care to assess transmission events within an HIV-infected population (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007). By definition, these "samples" will not contain an HIV sequence from every single HIV-infected MSM in the population. In the UK for example, approximately one third of individuals living with HIV are estimated to be undiagnosed (HPA 2008). Further reasons why MSM may only be partially represented in population-based studies include HIV diagnosed MSM being sexually active in geographic areas where they do not attend clinics for treatment and care, HIV diagnosed MSM not attending clinics, or samples from HIV diagnosed MSM being unavailable (due to amplification problems etc). The consequences of not including every sequence within a local population for the purposes of phylogenetic reconstructions are not known. If the formation of robust clusters (section 2.2.11) is affected by the diversity, representativeness and completeness of the entire sample, these factors could be crucial.

It is possible that the identification of some transmission events may be affected by the relative genetic diversity of the population sample. This means that any observed phylogenetic clusters may reflect relative relationships between sequences rather than demonstrate absolute relationships. Brenner *et al* (Brenner, Roger et al. 2007) undertook a phylogenetic reconstruction of transmission events using sequences entirely obtained from patients diagnosed during recent infection and found 75 possible phylogenetic relationships according to cut-off criteria described above. A second phylogenetic analysis included an additional number of sequences from patients chronically HIV-infected at diagnosis; they found that several of the original "transmission events" between individuals with recent HIV infection were disrupted. Sample size was found to be an important determinant of accuracy in a study (Wiens and Servedio 1998) that compared the results from parsimony, likelihood and distance based phylogenetic methods (see section 2.2.6). Assessment is therefore required of the extent that phylogenetic reconstructions can produce reproducible and consistent results.

1.6.2 Diversity of *pol*

The *pol* region of the HIV genome (section 1.1.5) is generally used for phylogenetic analyses out of convenience, since it is a by-product of the routine HIV drug resistance testing. There are relatively few full length (i.e. entire genome) HIV sequences available from patients involved in a known transmission event, and large scale phylogenetic analyses would be restricted by the sequencing costs of producing such a dataset and the computing power needed to analyse it.

However, there are concerns that the *pol* region is too conserved since it codes for regulatory genes involved in viral replication and consequently has insufficient genetic variability (Palmer, Vuitton et al. 2002). This has led to debate about its suitability for phylogenetic reconstructions (Palmer, Vuitton et al. 2002; Sturmer, Preiser et al. 2004). While *gag* and *env* genes have been preferred due to their greater genetic variability, *pol* remains attractive due to its greater accessibility through routine drug resistance testing. Hue *et al* (Hue, Clewley et al. 2004) demonstrated that phylogenetic analyses of 140 HIV *pol* sequences produced the same results as analyses using HIV *env* and *gag* sequences from the same patients.

1.6.3 Sampling considerations

It is likely that the specific selection of HIV-infected individuals from whom the sequences were obtained for phylogenetic analysis will affect the likelihood of recognizing transmission events. For instance, sequences collected from one geographical region and taken around the same date may be more likely to be drawn from individuals within the same transmission network.

Brenner *et al.* selected three populations within Canada (Brenner, Roger et al. 2007): a Quebec cohort of recently HIV-infected MSM (n=215) and a provincial genotyping programme (n=502). A chronically HIV-infected population was also taken entirely from the provincial genotyping programme (n=660). The presence of individuals drawn from the Quebec cohort increases the chance that individuals will be selected from the same transmission networks compared to the provincial sample. This generates a bias; more transmission events will be identified between the recently HIV-infected sample from Quebec compared

with the sample of chronically infected individuals. Furthermore, while the chronically HIV-infected were sampled between 2001-2005; the dates of infection for these sequences were not known. The interval between infection dates within this population may span tens of years, increasing the likelihood that they had been obtained from different transmission networks. Consequently it is essential that population samples used for phylogenetic analyses are selected with the consideration that no group under investigation has an increased likelihood of being involved in transmission networks.

1.6.4 Infection stage considerations

Interpretation of and comparison between phylogenetic studies of HIV transmission using datasets derived from patients with recent infection is problematic. This is due to the definitions of recent HIV infection employed. Importantly, if generous definitions of the latter are applied, it may serve to overestimate the extent that this population are generating transmission events. Viral load is known to peak less than a month following infection, but remains elevated for approximately 10 weeks (Fiscus, Pilcher et al. 2007). However, definitions of recent infection and the laboratory techniques used to identify it vary (Pilcher, Fiscus et al. 2005; Fiscus, Pilcher et al. 2007), and those frequently adopted mean that patients can be so categorized up to (Pao, Fisher et al. 2005) or even more than six months after infection (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007).

Moreover, infection stages do not remain static over the course of an individual's infection (section 1.1.6). However, many studies do not consider the transient nature of recent infection. Phylogenetic analyses frequently

categorize patients as being recently or chronically HIV-infected according to their infection stage at diagnosis (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007). However, all patients will experience recent and chronic infection, and the observation of possible transmission events (ascertained through phylogenetic reconstruction) between patients diagnosed during recent infection does not necessarily mean transmission occurred while the patients were recently HIV-infected.

To illustrate this, the mean time interval between infection dates in Brenner's "transmission events" identified between patients recently HIV-infected at diagnosis was 15 months (±9.5 months) – indicating the transmissions could have occurred many months after HIV acquisition. Specifically, 27 of the phylogenetically linked sequences from recently HIV-infected individuals had time intervals of over two years between their infection dates. These were interpreted as transmissions occurring from patients with recent HIV infection. It is similarly difficult to interpret possible transmission events found between sequences taken from individuals with chronic infection: all individuals with chronic HIV infection were once "recently HIV-infected". For studies that explore the impact of infection stage on transmission risk, it is essential to account for the transient nature of the risk factor.

1.6.5 Sexual behaviour

While factors such as viral load, STIs and infection stage are likely to affect population level HIV transmission, the sexual behaviour of patients from whom a sequence was derived will also be important. A phylogenetic reconstruction revealing many transmission events close together in time and geography may

reflect the action of one or two individuals with high rates of sexual partner change. This would produce results that it would be difficult to extrapolate to the population level. It is extremely challenging to include these data in phylogenetic analyses since such data are hard to capture, and difficult to link with reconstructed transmission events. While Pao *et al* captured indicators of unsafe sex (Pao, Fisher et al. 2005) with their reconstruction, the method could not precisely match behaviour to actual transmission events. Consequently results need to be interpreted with caution.

1.6.6 Analyses of transmitted drug resistance

Studies that measure the prevalence of TDR and transmission from patients with drug resistance mutations are difficult to interpret (section 1.2.2 and 2.5.2). Studies of TDR have found a range of prevalence levels (Little, Holte et al. 2002) (Grant, Hecht et al. 2002; Cane, Chrystie et al. 2005). However, it is impossible to exclude the possibility that phylogenetically linked sequences, sharing identical mutations were generated by one individual with acquired resistance, rather than onward transmission between individuals with TDR.

1.7. Gaps in the research

Phylogenetic analyses to date have highlighted the heightened transmission potential of the recently HIV-infected. However, such studies have remained exploratory, and have focused almost exclusively on sequences from patients diagnosed during recent HIV infection. While Brenner *et al.* (Brenner, Roger et al. 2007) attempted to measure the extent that the recently HIV-infected cause new transmission events relative to the chronically HIV-infected, their methods lacked rigour. Differences between the sample populations of the recently and

chronically HIV-infected, and the failure to recognize the transient nature of infection stage, mean that the extent that the recently HIV-infected are generating transmission has not been measured with adequate precision.

Phylogenetic analyses conducted for public health purposes require more rigorous design, application and interpretation. From an epidemiological perspective, appropriate consideration should be given to sample selection (especially with regard to geography and diagnosis dates). In order to obtain a measure of HIV transmission at the population level from certain groups, datasets should be extended to include sequences from individuals who were chronically HIV-infected at diagnosis. The latter group needs to be broadly comparable with the population of the recently HIV-infected with regard to their likelihood of being included in transmission networks (as far as is possible). It is also important in analyses that infection stages are considered to be transient properties that change over the course of an individual's infection.

The published evidence that indicates individuals with recent HIV infection have a high risk of transmission per sexual act (Wawer, Gray et al. 2005). This and the relative ease with which this population can be studied may mean that other groups that have potentially important roles in generating HIV are insufficiently investigated. Phylogenetic analyses need not be restricted to studying the recently HIV-infected; analyses can be extended to integrate other relevant risk factors, such as concurrent STIs, and chronically-infected (both the ARV-treated and untreated individuals).

The prevalence of TDR has been studied and its transmission reconstructed through phylogenetic analyses. However, the extent that acquired and transmitted drug resistant populations are generating TDR is not known. Additionally, the sources of TDR have not been explored in relation to viral load and infection stage.

There are uncertainties with regard to the ability of phylogenetic analysis to identify accurate transmission events to be used for public health purposes. The concerns include: whether the variability of *pol* is sufficient for accurate reconstructions; the effect of the overall genetic diversity of the population samples have on the likelihood of recognizing phylogenetic relationships; the inability to determine the direction of transmission; and the possibility that observed phylogenetic linked relationships could have been generated or mediated through an unsampled third party. While it is not possible to verify the transmission events identified through phylogenetic reconstructions through linkage to sexual histories at a population level, it is possible to perform sensitivity analyses to assess the consistency of phylogenetic approaches.

1.8. Research objectives

This thesis uses a phylogenetic approach that combines clinical, diagnostic and demographic data to enhance understanding of HIV transmission events among MSM. Specifically, it uses combined datasets to perform phylogenetic reconstructions of HIV transmission events to:

1) assess the consistency of results from phylogenetic reconstructions of HIV transmission events;
2) explore transmissions from the recently HIV-infected population and critique the methods used in such analyses;
3) ascertain the risk factors associated with transmission using a novel approach;
4) explore the transmission sources of patients diagnosed with TDR.

1.9. Thesis outline

Chapter two outlines the concepts, methods and definitions applied to phylogenetics, and sequence-based, clinical, diagnostic and demographic data. **Chapter three** describes and compares the datasets used in the thesis. **Chapter four** explores the consistency of phylogenetic approaches in reconstructing HIV transmission events. **Chapter five** explores the transmission events from the recently HIV-infected and critiques the methods used in such analyses. **Chapter six** ascertains the sources of new HIV transmissions and the factors associated with these sources. **Chapter seven** explores the sources of TDR. **Chapter eight** describes how the thesis contributes to research and links it with the current literature. It also outlines the thesis limitations and describes the future research needed in this field.

2. Chapter Two: Sequence-based, diagnostic and clinical HIV data: principles and definitions

Sequence-based, laboratory and clinical data are collected routinely around the time of HIV diagnoses for medical and surveillance purposes. The data sources available and the techniques used to obtain them are described in this chapter. The chapter also outlines how the sources can be combined and used for public health analyses, and their limitations. Finally, the definitions used throughout this thesis are provided.

2.1. Introduction

Sequence-based, laboratory and clinical data are collected routinely around the time of HIV diagnoses for medical and surveillance purposes. The combination of such data has the potential to enhance our understanding of HIV transmission (section 1.4). This chapter aims to describe the data sources available and the techniques used to obtain them. It also outlines how the sources can be combined and analysed, and their limitations. The specific data sources used in the thesis are detailed in chapter three.

2.2. Phylogenetic analysis of sequence-based data

The principles of phylogenetics are outlined. The applications of phylogenetic analysis to HIV sequences are then described. Phylogenetic methodologies are explained and the techniques used in this thesis are defined.

2.2.1 Phylogenetic principles

Phylogenetic analyses investigate the evolutionary relationship between organisms. Over time, through successive generations, changes occur in genetic make-up of organisms. Phylogenetic reconstructions exploit the changes that have accumulated between these organisms. The technique is derived from evolutionary theory, which states that the genetic similarity between species is attributable to a descent from common ancestry. It assumes the more similar the genetic make up of two organisms is, the closer they are in time to having a common ancestor (Page 1998).

2.2.2 DNA, nucleotides, amino acids and proteins

Genetic material is contained within DNA (deoxyribonucleic acid). DNA comprises two long strands of nucleotides arranged into a double helix. The nucleotides consist of two purines (Adenine (A), and Guanine (G)), and two pyrimidines (Thymine (T), and Cytosine (C)). Phylogenies are routinely reconstructed from nucleotide, or amino acid sequences which are aligned so that homologous bases are compared. In this thesis, only evolutionary changes between nucleotide sequences are considered, since this contains the highest resolution of genetic variability.

2.2.3 Nucleotide sequence alignments

Before phylogenetic analyses can be conducted, the sequences need to be aligned. This is because the analyses exploit differences found at specific nucleotide positions. In order for these analyses to be meaningful, it is essential that comparisons are made between equivalent (homologous) positions. The nucleotide sequences are aligned in-frame i.e. as triplets of three nucleotides (codons) coding for specific amino acids.

A representation of a nucleotide sequence alignment is shown in **Figure 2.1a**. Differences that accumulate between sequences over time can take the form of point mutations/substitutions (whereby the nucleotide has been replaced by another base), insertion (one or more nucleotides have been added into the sequence), or deletion (one or more nucleotides have been removed). During a sequence alignment, each nucleotide position will be compared at equivalent positions and categorized as either a “match” (where the nucleotides are identical), a “mismatch” (where the nucleotides differ), or a “gap” (where one or

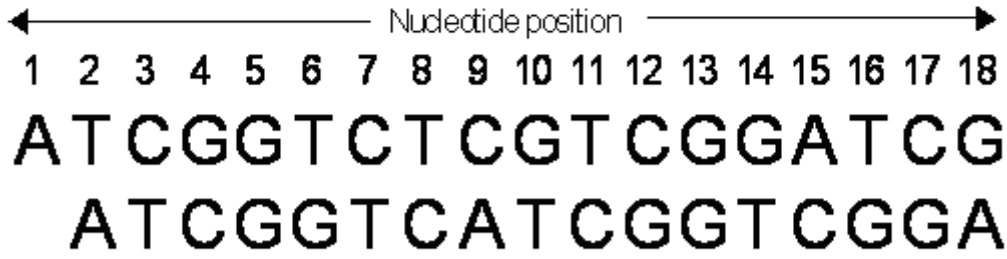
more of the sequences has had an insertion/deletion) **(Figure 2.1b**). Sequence alignments work to take account of insertions or deletions **(Figure 2.1c**). It is important to limit the number of gaps so that the resultant alignment makes biological sense. Provided the sequences are sufficiently similar, sequence alignments can be performed manually using software such as Sequence Analyzer (SE-AL). Otherwise, programmes such as Clustal-X (Thompson, Gibson et al. 1997) can be used to align sequences crudely according to a scoring system based on the frequency and extent of gaps. A heuristic search is then conducted to find the best alignment according to the scoring system, which should then be checked manually before phylogenetic reconstruction.

2.2.4 Evolutionary models

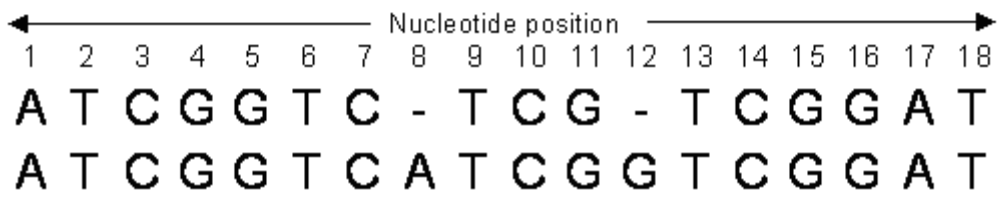
The majority of phylogenetic methods are dependent on the user defining an explicit evolutionary model. These are necessary since the rate of substitutions that accumulate between two sequences over time is not linear (Yang 1996). Some substitutions occur more frequently than others (Kimura 1980). For instance transitions (substitutions between purines, or between pyrimidines) are more frequent than transversions (substitutions between purines and pyrimidines) (Kimura 1980). Additionally, multiple substitutions may occur at the same nucleotide position, with the saturation of specific positions making it difficult to ascertain the true evolutionary distance (Holmes 1998). Evolutionary models convert the genetic distances that exist between sequences into measures of estimated evolutionary distance.

Figure 2.1: Graphical representation of sequence alignments

i) Unaligned sequences:



ii) Aligned sequences:



Several models of molecular evolution have been developed. Each model varies in its complexity and the extent to which it incorporates biochemical and evolutionary knowledge. Models can include three parameters: nucleotide base

frequency, base exchangeability and the rate of heterogeneity between sites. Examples of common models of evolution are given in **Table 2.1**.

The base frequency parameter accounts for the respective frequency of the four nucleotides throughout the sequences being compared. Base exchangeability describes the relative tendency of bases to be substituted for one another.

Table 2.1: Summary of the main evolutionary models

Evolutionary model	Description
Jukes Cantor (JC) (Jukes 1969)	Assumes an equal frequency of the four bases and that all substitutions are equally likely.
Kimura's two parameter model (K2P) (Kimura 1980)	Assumes base frequency is equal, but that transitions occur more frequently than transversions.
Felsenstein (F81) (Felsenstein 1981)	This model allows an unequal frequency of base substitutions.
Hasegawa, Kishino and Yano (HKY85) (Hasegawa, Kishino et al. 1985)	Allows both the base frequency and transition/transversion rate to differ.
General Time Reversible Model (GTR) (Lanave, Preparata et al. 1984)	Allows all substitutions to occur at a different rate and permits the user to define initial base frequency rate of heterogeneity between sites. Also allows substitution rates to be reversible.

Substitutions are not equally likely between the four bases. The rate of heterogeneity summarises the difference in substitution rates across the DNA sequence, typically described by a gamma distribution. The rate variation among sites is described by the shape parameter (α). Small values of α will result in L-shape distributions, indicating extreme rate variation across the sequences, whereas high values of α will reflect a bell shape distribution, demonstrating that most of the sites remain invariable (Holmes 1998). Models featuring a gamma distribution of rate heterogeneity are conventionally given the suffix + Γ.

The simplest model is the Jukes Cantor model since it only includes one parameter, the base substitution rate, and considers the rate to be equal between all four bases. The most complex is the GTR model since it encompasses all three parameters and allows these to be user-defined.

Software such as ModelTest can identify the most appropriate evolutionary model, which is informed by the sequence data itself. The ModelTest algorithm with the program PAUP calculates the hierarchical likelihood ratios of 56 evolutionary models using a chi-square distribution and selects the model that best fits the sequence data (i.e. that with the highest likelihood score) and estimates the corresponding parameters.

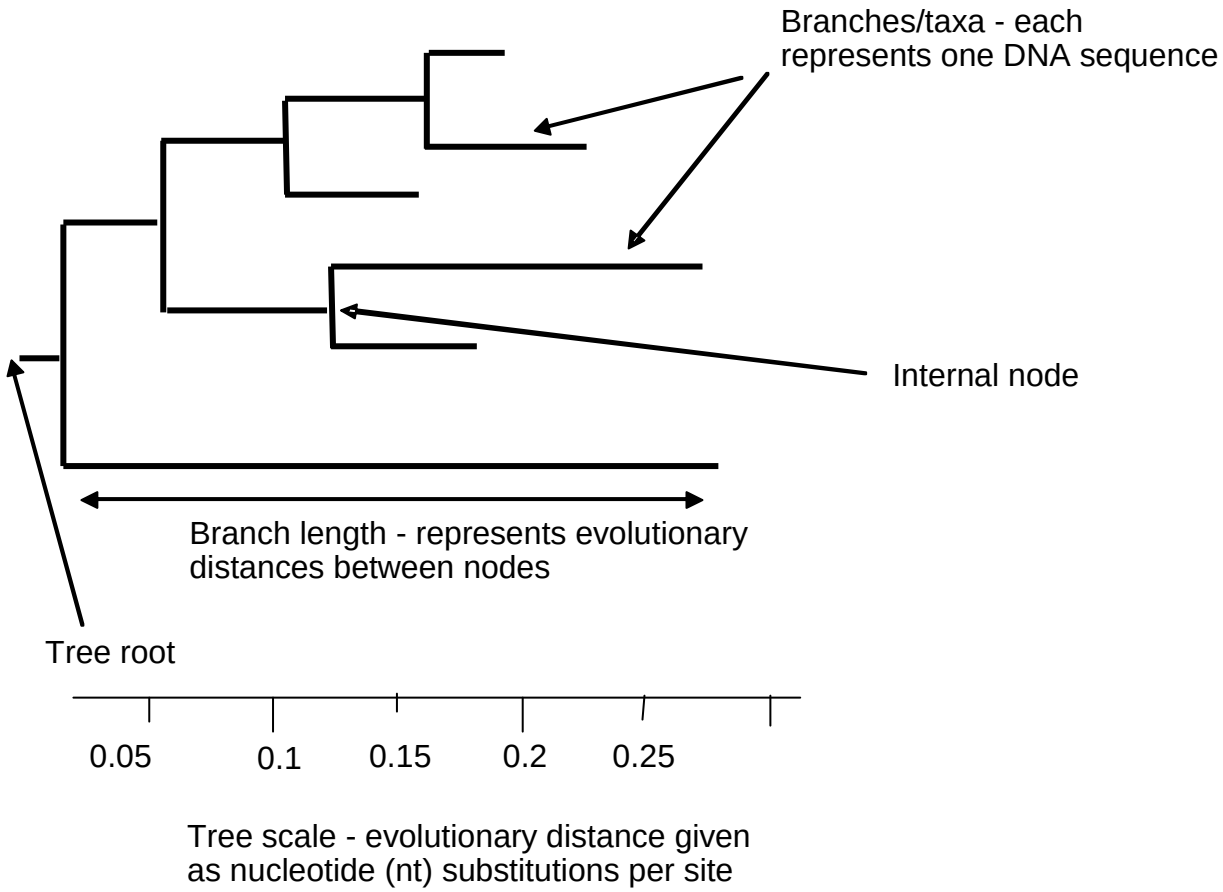
2.2.5 Phylogenetic trees

A phylogenetic tree is a graphical representation of the evolutionary relationships that exist between aligned sequences. Examples of different types of tree representations are given in **Figure 2.2**. 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". The branching pattern, (the order of the nodes and branches), is referred to as the "topology".

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; the branch lengths indicate the extent of genetic divergence 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 common ancestry of the sequences, but not the extent of genetic divergence.

Figure 2.2: Schematic diagram of a rooted phylogenetic tree



2.2.6 Tree building methods

There are different methods of creating phylogenetic trees. These vary in their complexity, computational demands and their assumptions about evolutionary theory. There is no "one-fit" method to suit all data and no tree is guaranteed to find the true evolutionary relationships between sequences. The most appropriate tree building method (and, where relevant, the evolutionary model upon which it is created) is informed by the specific data set. Tree-building methods are generally categorized into "distance-based" and "character-based" methods. These are summarized in **Table 2.2.**

Distance-based methods

Distance-based methods involve two steps. Firstly, the evolutionary distance is calculated between every possible sequence pair in the given alignment, creating a distance matrix. The proportion of nucleotide positions in which the two sequences differ is calculated, creating a measure of dissimilarity. Secondly, the tree is constructed on the basis of the relationship between the distance values. There are two main distance based methods: the Unweighted Pair Group Method with Arithmetic Means (UPGMA) and Neighbour-joining (NJ).

UPGMA

This method identifies the two sequences with the smallest genetic distance between them, and then halves that distance to define the branching patterns. These two sequences are then combined and considered as one unit. A new distance matrix is computed, calculating the distance from the newly formed unit to each of the remaining sequences. The process is repeated until there is only one entry in the matrix. A tree is then built on the basis of the distance matrix obtained.

Table 2.2: Summary of phylogenetic tree construction methods

Method	Description	Advantages	Disadvantages
Distance-based			
UPGMA	Tree constructed through pair-wise distance matrix.	Quick.	Assumes rate of molecular evolution is constant.
Neighbour-joining	As above, but distances can be adjusted through application of appropriate evolutionary model.	Quick, allows application of evolutionary model.	Has been known to build trees incorrectly.
Character-based			
Maximum likelihood	Uses Maximum Likelihood to calculate the probability of a tree by describing the patterns of the sequence alignment, given a specific model of nucleotide substitution. The method calculates the likelihood of all possible trees for the specified alignment, and selects the one associated with the maximum likelihood.	Exhaustive.	Computationally demanding.
Bayesian	Uses Bayesian theory and a user defined evolutionary model to calculate "posterior" probabilities of all possible trees for the specified alignment, and selects the one with the highest probability.	Allows genetic distances obtained to be time calibrated.	Fairly time intensive.
Parsimony	Tree constructed on the basis that the minimum number of evolutionary differences between sequences is the most likely.	Quick.	Two or more trees can be selected as best tree. Does allow for substitutions occurring at same site multiple times. It only uses a small fraction of nucleotide positions to inform tree.

The UPGMA method assumes that the rate of evolution is linear (i.e. that there is a global molecular clock). Consequently all taxa are equidistant to the root. However, this assumption is no longer accepted; this is a frequent criticism of this method's reliability (Li 1993).

Neighbour-joining

The NJ method works in the same way as the UPGMA method, but rather than assuming the rate of evolution is constant, it permits the adoption of a specific evolutionary model.

Like the UPGMA method, it is quick, and computationally undemanding. However, it cannot guarantee to find the right tree: cases have been reported where the true phylogeny was known, and the NJ tree building algorithm has provided inaccurate reconstructions (Hillis D 1996). It has been found to have inferior robustness compared to character based methods (Huelsenbeck 1995).

A further criticism of distance methods is that they do not make use all of the available information. Through constructing distance matrices, information about individual sites is lost, and only an overall estimate of relative distances between the sequences is given. Additionally the branch lengths are not necessarily biologically meaningful. Programmes that can be used to build NJ trees include ClustalW (Thompson, Gibson et al. 2002) and PAUP (Swofford 2001).

Character-based methods

Character-based methods are informed directly from the sequences, rather than from the proportion of nucleotide differences. There are two main character-based methods: Maximum likelihood and Bayesian analysis.

Maximum likelihood

Maximum likelihood (ML) tree construction exploits the concept of the statistical theory of likelihood probabilities (Felsenstein 1996). In the context of phylogenetic reconstructions, maximum likelihood methods calculate the probability of a tree by describing the patterns of the sequence alignment, given a specific model of nucleotide substitution. The method calculates the likelihood

of all possible trees for the specified alignment, and selects the one associated with the maximum likelihood.

Specifically, the likelihood (L) that the constructed tree represents the actual evolutionary relationship is given by:

$$L=P\,(D|H)$$

Where P is the probability that the observation D (the sequences) is true, given the hypothesis, H (the phylogenetic tree).

The ML approach is rigorous, makes use of each nucleotide position, allows the user to adopt a specific evolutionary model that fits the specified data and permits statistical testing of evolutionary hypotheses (Holmes 1998). However, it is computationally demanding, particularly as the number of taxa increases. In order to reduce the time used to build the trees, often an initial NJ tree is used as a starting topology, and likelihood scores are computed for each rearrangement of the initial topology. Programs that can be used for ML tree construction include PAUP (Swofford 2001) and Phylip.

Bayesian Analysis

Bayesian phylogenetic analysis is based on Bayes' theorem. Like the Maximum likelihood method, the Bayesian method incorporates an explicit evolutionary model, and looks for trees that correlate best to the sequence alignment under the specified model. However, unlike the ML method, it calculates the so-called

“posterior probability” of distribution of trees, i.e. the probability of a tree, given the observed data (Drummond and Rambaut 2007). The most frequently used software for phylogenetic Bayesian analysis is called Mr.Bayes (Ronquist and Huelsenbeck 2003).

Bayesian analysis is becoming more popular (Lewis, Hughes et al. 2008) due to computational advances. Specifically, a simulation technique called Markov Chain Monte Carlo (MCMC) has increased the method’s efficiency. This algorithm searches at random for the tree with the highest probability. It constructs a random tree “T1”, and compares it to another randomly generated tree “T2”. If the likelihood of T1 is lower compared to the likelihood of T2, then T2 replaces the current T1. If not T1 is held in 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 during the MCMC simulation (Mau, Newton et al. 1999).

This method is more efficient than the ML method since it permits simultaneous independent searches that can exchange information. Also, the MCMC method provides a measure of statistical support for each tree constructed, negating the need for techniques to check tree robustness, such as bootstrapping (see section 2.2.7). Bayesian analysis also permits the calibration of phylogenies, allowing the introduction of specific lineages to be dated (Hue, Pillay et al. 2005; Lewis, Hughes et al. 2008).

2.2.7 Tree robustness

It is impossible to guarantee that phylogenetic reconstructions will represent the true evolutionary relationship between sequences (Hillis D 1996). Consequently, it is important to obtain a measure of tree robustness: this is frequently estimated through a process called bootstrapping (Efron, Halloran et al. 1996). Bootstrapping is a statistical resampling method that works through repeated, random sampling of columns of nucleotide positions of sequence alignments. A new tree is created on the basis of each randomly sampled column. This process is repeated many times (usually up to 1000). The proportion of times the original tree matches the reconstructions of the randomly sampled columns is expressed as a percentage, or the bootstrap value. Bootstrap values are shown on the branches of the tree. A bootstrap value of 100% on a particular branch would demonstrate that it was present in all sampled trees.

2.2.8 Applications of HIV phylogenetics

The applications of phylogenetic analyses of HIV sequences are described in section 1.4.

2.2.9 Reconstructing HIV transmission events through phylogenetic analysis

For the purposes of reconstructing “who infected who” phylogenetic methods have tended to use maximum likelihood (Brenner, Roger et al. 2008) and Bayesian (Lewis, Hughes et al. 2008) approaches. Those that use maximum likelihood approaches frequently use neighbour joining methods to identify an initial topology (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007), and use it

as the starting tree. Since the actual transmission histories are not known, it is difficult to validate whether reconstructions represent true transmission events, and efforts have focused upon developing a definition of what may be considered a “robust” indicator of such a transmission event.

The cut-offs used to identify a possible transmission event comprise the observation of a cluster with a bootstrap of at least 99% and a genetic distance <0.015 nucleotide substitutions per site (Hue, Clewley et al. 2004). These cut-offs are conservative and have been shown to reduce the proportion of false-positive clusters among a phylogenetic analysis of UK HIV sequences derived from individuals with a known transmission history (Hue, Clewley et al. 2004). The need for such definitions to be interpreted with caution remains since these cut-offs are not absolute indicators of transmission.

2.2.10 Limitations of phylogenetic reconstructions of HIV transmission events

The limitations of using phylogenetic reconstructions of HIV transmission events for public health purposes are described in section 1.6.1 and 1.6.2.

2.2.11 Thesis methods and definitions

Phylogenetic methods

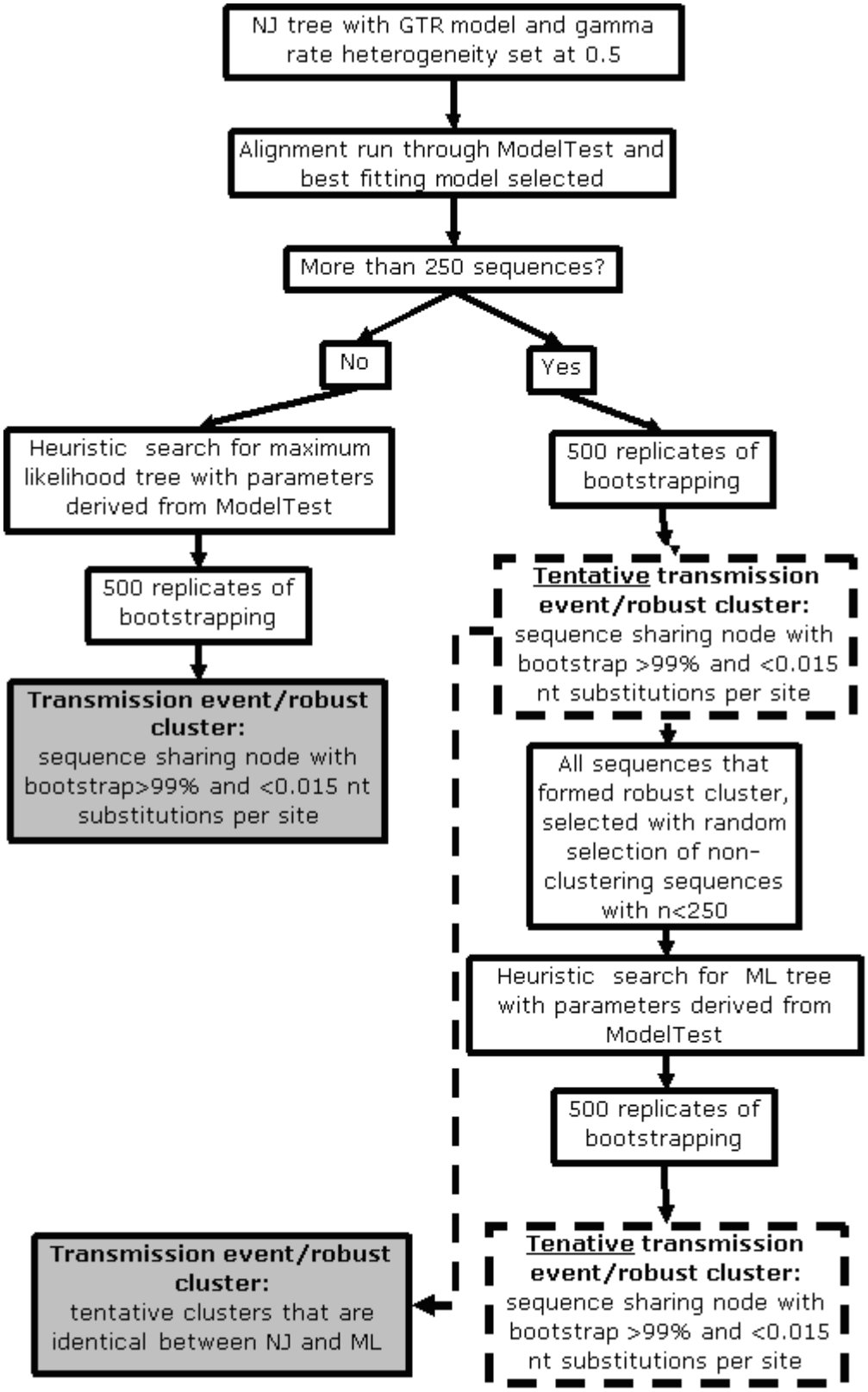
The phylogenetic methods and definitions used in this thesis are summarized in **Figure 2.3**. The sequences were aligned manually in Se-Al (Sequence Analyzer version 2.0). The nucleotide positions associated with drug resistance were deleted to prevent bias arising from convergent evolution (Shafer, Rhee et al. 2007) (whereby organisms have a similar range of mutations due to shared

environmental pressures rather than a shared ancestry). A NJ tree was constructed using the General Time Reversible model with gamma rate heterogeneity estimated at 0.5. The alignment was then run through PAUP (version 4.0b10) with ModelTest (version 3.7) to select the best fitting model.

If there were fewer than 250 sequences in the sample, a maximum likelihood tree was constructed using the initial NJ tree as the starting tree, and the best fitting model and the parameters derived through PAUP with ModelTest. The resultant tree was then bootstrapped with 500 replications.

If there were more than 250 sequences in the sample, a NJ tree was constructed using the most appropriate model selected by PAUP with ModelTest and the parameters derived. This is because the computational demands of constructing a maximum likelihood tree for larger sample sizes were not feasible. Sequences that formed a cluster within the NJ tree were then selected, along with a random sample of sequences that did not form a robust cluster. This smaller sample of sequences was then treated as a sample of fewer than 250 sequences and re-run through PAUP with ModelTest and a maximum likelihood tree was constructed accordingly and bootstrapping analysis repeated. The topology and relationship between the sequences that clustered in the original NJ tree were then compared. Provided that the sequences formed the same relationship as in the original NJ tree, they were interpreted as a robust cluster.

Figure 2.3: Schematic diagram of the phylogenetic methods used throughout thesis



Definitions

In all instances, a cluster was defined as two or more sequences that shared a common node, a bootstrap support of 99% or more, and a genetic distance of under 0.015 nucleotide substitutions per site (Hue, Clewley et al. 2004). This was interpreted as a “robust cluster” and also a “possible transmission event”

2.3. HIV diagnostic data and identifying recent infection

HIV infection is characterized by an extended asymptomatic period that can last 10-12 years **(Figure 1.4)**. While a proportion of patients may experience “seroconversion illness” shortly after infection, which, in combination with a recent risk exposure may prompt them to come forward for testing (Remis, Alary et al. 2000) others may remain unaware of their infection for many years. Many epidemiological studies of HIV infection are consequently dependent upon HIV-infected patients coming forward for HIV diagnostic tests. In addition to identifying antibodies to HIV, viral proteins, or the virus itself, diagnostic tests have also been adapted so that patients who acquired their infection recently can be distinguished from those with long-standing infection.

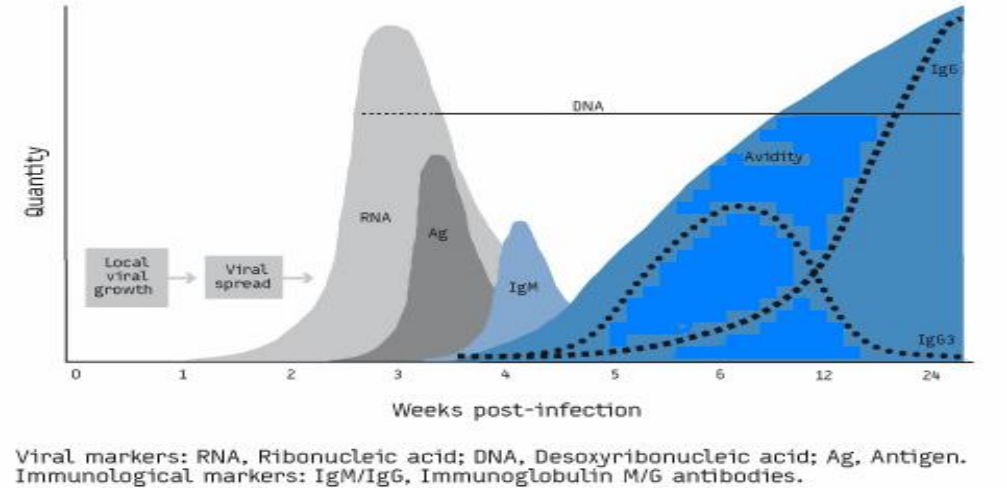
These diagnostic markers, and their application in ascertaining HIV infection stage and calculating HIV incidence, are described below. First the immune response to HIV is outlined. Secondly, the diagnostic markers that relate to the specific stages of the immune response are described. Thirdly, methods used to identify recent infection are summarized (including diagnostic markers, laboratory algorithms and testing histories). Fourthly, the public benefits of linking infection stage to population-level data are summarised. The limitations

of measuring recent infection and incidence estimates at the population level are described. The definitions of recent infection used in the thesis are outlined.

2.3.1 The HIV immune response

The HIV immune response following exposure is shown in **Figure 2.4.**

Figure 2.4: Schematic diagram of viral load and immunological markers during the first weeks of HIV infection



Source: adapted from (Murphy and Parry 2008)

The very early stages of HIV infection are characterized by local viral growth, followed by viral spread two-three weeks after infection (Busch and Courouce 1997; Busch, Glynn et al. 2005). Approximately 10 days after infection, HIV-RNA and the p24 antigen (a protein component of the virus core) may become detectable (Busch, Glynn et al. 2005), peaking at around 16 days post infection and declining over the following 10 weeks as the antibody response develops.

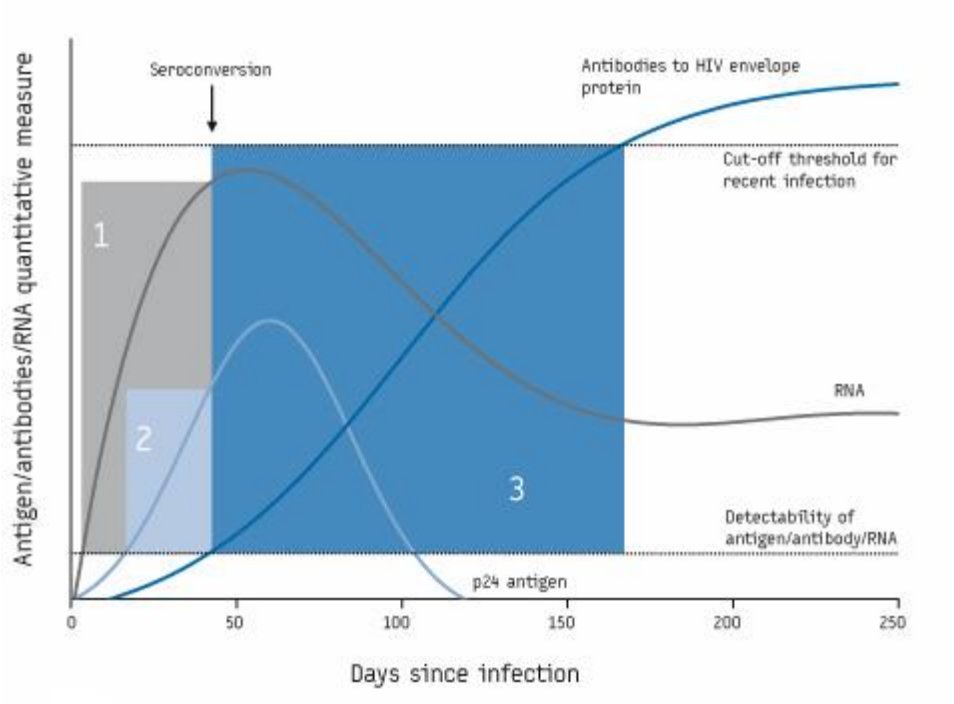
The antibody response to HIV normally occurs between three to twelve weeks following an HIV transmission. This period, describing the change from antibody negative to antibody positive, is called “seroconversion”. There is considerable variation in the time of seroconversion between individuals (Busch, Glynn et al.

2005). The initial immune response is typically characterized with a virus specific IgM response (Gaines, von Sydow et al. 1988). It generally peaks within 1-2 weeks after infection falling back to background levels 1-2 weeks later. Shortly afterwards the IgG response is developed – this response is long lived with IgG antibody levels detectable at a high level throughout infection.

2.3.2 Diagnostic tests

The period of time from HIV infection to that when diagnostic tests can detect HIV infection, is characterized in **Figure 2.5**. This period is referred to as the "diagnostic window" (Busch and Courouce 1997).

Figure 2.5: Schematic diagram showing what stage of infection diagnostic markers can detect HIV infection



Key:
1) HIV-RNA detectable
2) p24 detectable
3) IgG detectable

Source: adapted from (Le Vu, Pillonel et al. 2008)

Most patients are diagnosed with HIV through the detection of virus specific antibodies to HIV (anti-HIV IgG) (Gurtler 1996). The HIV enzyme immunoassay (EIA or ELISA) exploits the binding of the patients' antibodies against HIV protein to antigens immobilized typically on the surface of the wells of microtitre plates. Patient serum, which may contain HIV antibodies, is added to the plate. Non HIV antigen binding antibodies are washed free of the plate. A second antibody, a conjugate, is then added to determine whether the human serum contains antibodies. Excess conjugate is then washed from the plate. A substrate (chromagen) is added, and the enzyme of the conjugate acts on the substrate to give a colour change if the serum sample contained anti-HIV. The intensity of the colour is calibrated against a scale known as Standardized Optical Density (SOD). A SOD score greater than 1.0 is interpreted as a positive reaction. Confirmatory testing follows before HIV infection is diagnosed, using methods such as the western blot (see below).

The western blot is a method of detecting specific host protein products and is often used as a confirmatory HIV test following a reactive EIA. It uses gel electrophoresis to separate denatured HIV proteins on the basis of their relative molecular mass. The proteins are then transferred to a nitrocellulose membrane where they are probed with the patient serum that is suspected of containing HIV antibodies. For HIV, the antigens detected are derived from *gag* (p53/55, p24/5 p17/18,) *pol* (p31/32) and *env* (p160, p41 p45). HIV can also be diagnosed through nucleic acid testing (Pilcher, Fiscus et al. 2005) conducted through amplification techniques such as the polymerase chain reaction (PCR) and through identifying the virus through cell culture.

The diagnostic window can be reduced through combining HIV antibody tests with other diagnostic markers. Recent improvements in HIV test sensitivity mean the latest ("4th generation" tests) can detect both anti-HIV and HIV p24 antigen within four weeks of infection (Brust, Duttmann et al. 2000; Parry, Mortimer et al. 2003; Perry, Ramskill et al. 2008).

2.3.3 Identifying recent infection

Recent HIV infection can be ascertained in three ways: through diagnostic markers; through laboratory algorithms; and through clinic records of HIV testing history.

Diagnostic markers

Specific diagnostic markers are capable of identifying very early stages of HIV infection. These include detection of p24 antigen and HIV-RNA approximately 10 days after infection, declining over the following 10 weeks as the antibody response develops. Additionally, if a western blot detects high p24 and low p41 levels, this also indicates early infection. Conversely if a western blot detects high *pol* protein products, this may indicate long standing infection.

Laboratory algorithms

Laboratory algorithms have been developed that exploit the early immunological response to HIV infection in order to identify recent infection. Such developments have been necessary because while p24 antigen and western blots are capable of identifying early stage infection, the extremely short duration of these markers serves to under-detect recent infection among individuals recently exposed.

The Serological Testing Algorithm of Recent HIV Seroconversion, or STARHS, (Janssen, Satten et al. 1998) works through taking advantage of the gradual increase in anti-HIV over the first few months of recent HIV infection and assumes this increase occurs at a similar rate between individuals. Three main assays can be integrated into STARHS; the detuned assay; the BED assay; and the avidity assay.

A detuned, (a less sensitive assay), can be applied to samples that were found to be anti-HIV positive using a more sensitive EIA. Through deliberately combining highly sensitive tests with weakly sensitive tests, these strategies are able to estimate the proportion of patients who have only recently undergone seroconversion (Parekh, Pau et al. 2001). An anti-HIV negative test in the less sensitive EIA indicates the sample was derived from someone who acquired their infection within an average of 170 days (95% confidence interval 162-183 days) prior to the specimen collection.

Alternatively, the detuned assay can be replaced by the BED-CEIA assay (Parekh, Kennedy et al. 2002). This is a class-specific EIA that detects and compares the level of anti-HIV-IgG to total IgG. It works on the assumption that the ratio of anti-HIV IgG to total IgG increases after infection. If a specimen is positive on the sensitive EIA, but has a SOD of <0.8 on the BED assay it is considered to be recently infected. Avidity assays measure the strength of bonding between IgGs and the corresponding antigens (Chawla, Murphy et al. 2007). This increases over a period of months following infection.

The STARHS technique used the detuned assay up until fairly recently, but due to concerns about its appropriateness for non-B subtypes, its application was restricted to MSM populations. The BED and avidity assay were developed as an alternative that could be used for both B and non B subtypes, with the avidity assay being most successful (Suligoi, Butto et al. 2008).

Testing histories

As a patient's HIV testing pattern is often well documented in clinic notes, an alternative to laboratory markers can be obtained from testing histories. For instance, a short time interval (e.g. three months or less) between the date of the last HIV negative test and the date of the HIV positive can be used to indicate recently acquired HIV infection. However, this is more likely to identify recent HIV infection among those with frequent HIV testing patterns.

2.3.4 Applying recent infection markers to population level data

The development of techniques to ascertain recent HIV-infection has facilitated epidemiological studies including the monitoring of HIV incidence (Murphy, Charlett et al. 2004) and partner notification exercises (Pilcher, McPherson et al. 2002). They can also be used for assessment of transmitted drug resistance (TDR (for instance, patients recently HIV-infected at diagnosis are almost certainly drug naïve)). Before the advent of STARHS, HIV incidence estimates were limited to calculating rate of HIV diagnoses occurring among a cohort of HIV negative individuals followed up over many years. Current techniques have many advantages. They are quicker and cheaper than following large scale cohorts, and tests can be applied both to diagnostic samples and retrospectively to stored samples.

The ability to ascertain whether a patient was diagnosed during recent or chronic infection from a single diagnostic sample has enabled the categorization of viral sequences as being from patients who were recently or chronically infected at diagnosis. In turn this has enabled phylogenetic reconstructions of HIV transmission events by patients with recent HIV-infection at diagnosis (see section 1.5.1 and 1.5.2).

2.3.5 Limitations associated with applying markers that can identify recent infection to population level data

There are two issues that require consideration when applying techniques to identify recent HIV-infection at a population level. These relate to the techniques themselves and the selection of the population to which the techniques are applied. These will be discussed in turn.

The specific markers or techniques used to identify recent HIV infection vary between studies, making it difficult to draw comparisons (section 2.3.3). Frequently, the markers of choice are the STARHS, or utilizing testing history, data due to the short window of opportunity that other techniques have to recognize recent infection (e.g. HIV RNA and the p24 antigen). However, the selection of these more inclusive markers may serve to over-estimate the length of “recent infection”. Consequently, a sample from a patient ascertained through STARHS as being recently HIV-infected does not necessarily mean that the patient was experiencing elevated viral load at the time the sample was taken.

Variation exists in the immune response to HIV between individuals; consequently there is also variation between individuals in the time interval between infection and the time specific markers can detect recent HIV infection. (Busch and Courouce 1997; Fiebig, Wright et al. 2003). The number of days following infection that each test can define a recent HIV infection is not an absolute measure, but represents a population average.

There have been attempts to measure the extent of this variation for each marker. However, these have tended to focus on the markers that identify very early infection for purposes of accurate screening for blood donation (Fiebig, Wright et al. 2003; Busch, Glynn et al. 2005). Fewer studies have focused on the variation for markers that identify later stages of recent HIV infection (e.g. STARHS etc). One study (Reed 2004) examined the extent variation from infection to the time STARHS could identify recent HIV infection, between individuals. Nearly 400 follow-up specimens were collected from 60 individuals with recently acquired infection. The specific time interval within which the STARHS algorithm would define a recent infection was ascertained. Whilst the mean window period was 183 days following infection for the 60 individuals, the range of window periods extended from 63 to 404 days, demonstrating considerable variation between individuals. Consequently laboratory markers and techniques need to be interpreted with caution when they are applied to individual samples, for instance, to calculate an infection date from a diagnostic sample.

There are a number of other factors that can affect the accuracy of algorithms such as STARHS. These include late stage infection and treatment with ARV (Janssen, Satten et al. 1998). Whilst most infection stage tests are carried out on diagnostic sera, within anonymised datasets, the treatment status of patients from which the samples were derived cannot be known definitively; this may increase the risk of inadvertently including false-positive results. Finally, the tests can be affected by virus subtype. Algorithms were developed using assays for use on patients with sub-type B (Janssen, Satten et al. 1998) which give inconsistent results for non-B subtypes.

The application of algorithms to patient's diagnostic samples may not produce results representative of the populations at risk of HIV infection. This is because HIV test seeking behaviour may bias results (Remis and Palmer 2009). The resultant bias can go in either direction. For instance, recently HIV-infected patients may experience symptoms of seroconversion illness, prompting them to come forward for testing (Schacker, Collier et al. 1996). Those with recent risk exposure may also be more likely to present for testing (Remis, Alary et al. 2000). Conversely, those with lower sexual risk behaviour may also be more likely to seek regular health check ups. Such biases have been studied using modelling techniques in order to enable adjustment of population-based HIV incidence measurements (Remis and Palmer 2009).

2.3.6 Thesis methods and definitions

"Recent infection" is the term used throughout the thesis, since the methods used to identify it in each of the datasets vary depending on the available data. Since some datasets use STARHS, labels such as "primary" and "acute"

infection or “seroconverters” are inappropriate. Recent infection will not refer to a period of more than six months from infection throughout the thesis. However, the specific methods used to ascertain recent infection are made explicit throughout the thesis, and are defined for each data set in chapter three.

2.4. Clinical and demographic data

In the UK, laboratory and clinical reports of HIV diagnoses, and patients living with diagnosed HIV infection, are collected routinely on a voluntary basis by the Health Protection Agency and Health Protection Scotland for surveillance purposes (HPA 2008). Clinical and demographic data are usually collected at clinical consultation around the time of diagnosis. Data include: exposure (likely transmission route); age; country of infection; country of birth and (where applicable) year of arrival in the UK; ethnicity; injecting drug use; plasma viral load and CD4 count nearest to the time of diagnosis.

2.4.1 Geographic region of infection

The UK has a voluntary reporting system for new diagnoses of HIV, AIDS and deaths. This system attempts to ascertain the country of HIV infection through the follow up of likely HIV exposure events in conjunction with patient travel/migration history (Dougan, Gilbart et al. 2005). However, for many studies, the country, region or city of infection is not always possible to ascertain: often the clinic or country of diagnosis is taken as a proxy for “region” of infection (Masquelier, Bhaskaran et al. 2005).

2.4.2 Risk group and age-group

Data concerning risk group (transmission route) and age (or age-group) are routinely collected at the clinic consultation around the time of diagnosis.

2.4.3 Sexually Transmitted Infections

STI diagnostic data are usually collected around diagnosis and for subsequent attendances post diagnosis. However, the specific STIs may not be reported, and such reports may be categorized as “acute STI” or “other” with definitions of these categories varying between studies.

2.4.4 Viral load

The quantitative detection of HIV RNA in plasma can be used as a prognostic marker to monitor the success of therapy (Berger *et al* 2002) and estimate infectiousness. Various commercial and in house methods are available with the most sensitive tests capable of detecting approximately 50 copies/mL. These are based on polymerase chain reaction (PCR), branched DNA (b-DNA) nucleic acid sequence based amplification (NASBA), ligase chain reaction (LCR) or real time PCR measurements.

2.4.5 CD4 count

CD4 counts provide a measure of how many CD4 cells are present and are consequently used as an indicator of the functioning of the host’s immune system (section 1.1.7). CD4 counts are consequently collected routinely around diagnosis and at subsequent clinic attendances. They are used to monitor infection progression and inform decisions on when it is appropriate to begin therapy. Up until recently, treatment has commenced once CD4 cells counts reached below 200 cells/mm^3 (Pozniak, Gazzard et al. 2003). In Autumn 2008, guidelines were revised to recommend treatment discussions commence when CD4 counts reached between 200-350 cells/mm^3 (Gazzard 2008).

2.4.6 ARV treatment

Patients are very likely to be drug naïve at diagnosis (exceptions include patients treated with pre- and post-exposure prophylaxis before and after a risk exposure respectively, and those not reporting previous treatment). Treatment related data collected for HIV diagnosed patients at each clinic visit can document the occurrence and length of treatment interruptions. The treatment regimen may not be specified. In 2003, BHIVA guidelines recommended that patients without drug resistance mutations initially commence therapy with two nucleotide reverse transcriptase inhibitors (NRTIs), plus either a protease inhibitor (PI) or a non-nucleotide reverse transcriptase inhibitor (NNRTI).

2.4.7 Limitations associated with linkage to phylogenetic analysis

Phylogenetic reconstructions of HIV transmission have linked sequences to clinical and demographic data to see which patient groups have an important role in generating HIV transmission. Most studies to date have used clinical and demographic data obtained at diagnosis to permanently categorise patients by risk group. For data concerning transmission route, ethnicity and, to a lesser extent, age-group, this is reasonable practice. For other variables (e.g. STI presence, viral load, CD4 count), this may be less appropriate, since the risk factors vary considerably over the time of infection. The presence of an STI, (or viral load, CD4 count, treatment status or infection category) at around diagnosis will not necessarily reflect the status of that patient at the time of an HIV transmission event (section 1.6.4). With reference to STI presence, it is not always specified whether patients were screened for STIs. If not, there is likely to be an underestimation of STI reporting, and a bias towards symptomatic infections being recognized. If data are also collected on whether

patients were screened for an STI during the clinical consultation, this gives a better denominator for the calculation of prevalence.

The treatment history of a population used in phylogenetic analysis is important. Sequences from patients who have experienced treatment may have altered considerably to the sequences present within the same patient at around diagnosis. Consequently, for the purposes of phylogenetic reconstructions, it is important that sequences are obtained from drug naïve patients, and are taken as soon as possible following diagnosis.

2.4.8 Thesis methods and definitions

The clinical and demographic data available varies considerably between datasets used in this thesis. Consequently, the specific data used are described in chapter three.

2.5. Drug resistant viruses

2.5.1 About HIV drug resistance

HIV drug resistance is described in brief in section 1.1.9.

2.5.2 Limitations associated with monitoring prevalence and transmission of drug resistant viruses

Measuring the prevalence and monitoring the transmission of drug resistant viruses is difficult for several inter-related reasons: the complex dynamics of viral populations; differences in the sampling strategies; and varied definitions of drug resistance. These will be discussed in turn.

Dynamics of viral populations

Within an HIV-infected individual a number of slightly different viral species (quasi-species) exist. These variants are generated as random errors during the HIV life cycle. Replication errors that generate mutations can occur throughout the HIV genome (the most usual or common variant is known as “wild-type”). Each viral species may have a different level of fitness (i.e. the measure of the virus ability to survive and replicate) which will depend on two factors: the specific mutations it contains and its current environment.

Whilst some mutations are “neutral” and will not affect the ability of the virus to survive and replicate, others are not. By chance, some mutations generated as replication errors will confer resistance to ARV (HIV drug resistant species): such species will be better able to replicate within a patient taking ARV. However, in a selectively neutral environment (for instance, untreated patients), wild-type variants generally have a higher reproductive fitness and will become the dominant strain (Devereux, Youle et al. 1999).

For prevalence and transmission studies of drug resistance, the likelihood of finding drug resistance mutations in the virus from a patient’s blood sample depend upon the time elapsed between HIV infection, and the nature and duration of ARV. Among drug naïve patients, if a substantial amount of time has elapsed since infection, then the likelihood that a drug resistance mutation is found is decreased due the reduced fitness of the resistant strain. This will cause an underestimation of the prevalence of TDR. Similarly, the likelihood of finding drug resistance mutations among patients with acquired resistance will

depend upon whether the patient is currently treated and, if not, the time elapsed since treatment was interrupted. However, all viral variants are archived within the host and have the potential to re-emerge as selective pressures change.

Sampling strategies

The selection of populations for studies examining TDR requires careful consideration. The majority of studies use samples taken from patients at around the time of diagnosis: this population is likely to contain a relatively high but uncertain proportion of patients with recently acquired HIV infection (Burchell, Calzavara et al. 2003). Additionally, some studies use populations exclusively drawn from patients with recent infection to facilitate comparison, and to increase the chance that drug resistance mutations are recognized (Yerly, Kaiser et al. 1999; Masquelier, Bhaskaran et al. 2005) – the selection of such populations can contain inherent biases (Remis and Palmer 2009).

Definitions of HIV drug resistance

There is no unambiguous definition of HIV drug resistance mutations. This is for several reasons. Firstly, drug resistance mutations arise as a response to circulating ARV drugs: consequently new mutations appear as new drugs are introduced. Secondly, drug resistance mutations require different definitions for different purposes. Some are needed to inform clinical treatment options and others for surveillance purposes. However, some clinically relevant mutations are also polymorphic (i.e. the nucleotide position has two or more variants circulating at a high frequency in the population). Consequently, the inclusion of polymorphic mutations (that also confer drug resistance) in definitions used for

surveillance purposes will lead to an overestimation of the prevalence of drug resistance mutations. Thirdly, the definitions have to be derived from the sequences that are available. The majority of definitions have been derived from subtype B viruses. Preliminary observations suggest that mutations that cause drug resistance in subtype B viruses are also the main mutations that cause drug resistance in non-B viruses (Kantor, Katzenstein et al. 2005).

There are at least five expert lists of HIV drug resistance definitions: ANRS drug resistance interpretation algorithm; HIVdb drug resistance interpretation algorithm (Rhee, Gonzales et al. 2003); IAS-USA Mutations Associated With Drug Resistance (Johnson, Brun-Vezinet et al. 2008); Los Alamos National Laboratories HIV Sequence database (Clark 2003) and the Rega Institute Drug Resistance Interpretation Algorithm (Van Laethem, De Luca et al. 2002). Each varies to the extent that they list polymorphic mutations and/or mutations that are associated with reduced response to ARV. The complete list of mutations associated with each of these lists can be found on the Surveillance Drug Resistance Mutation (SDRM) worksheet (http://hivdb.stanford.edu/cgi-bin/AgMutPrev.cgi accessed 20th August 2009).

In 2007, Shafer *et al.* defined a list of mutations suitable for epidemiological studies (Shafer, Rhee et al. 2007). Mutations are included provided they meet the following criteria: commonly recognized as causing or contributing to resistance; non-polymorphic in untreated persons; applicable to all HIV subtypes. However, the list does not include all clinically relevant mutations. This list has recently been updated (Bennett, Camacho et al. 2009).

2.5.3 Thesis methods and definitions

The list of mutations deemed appropriate for surveillance purposes were used to define drug resistance mutations (Shafer, Rhee et al. 2007). These mutations are described in **Table 2.3.**

Table 2.3: HIV drug resistance mutations suitable for use for surveillance purposes

Drug class	Drug resistance mutations for surveillance
PI	L24I, D30N, V32I, M46I, I47A, I47V, G48V, I50V, I50L, F53L, I54V, I54L, I54M, I54A, I54T, I54S, G73C, G73S, G73T, G73A, V82A, V82F, V82T, V82S, V82M, I84V, I84A, I84C, N88D, N88S, L90M
NRTI	M41L, K65R, D67N, D67G, D67del, T69D, T69ins, K70R, L74V, V75A, V75M, V75T, V75S, F77L, Y115F, F116Y, Q151M, M184V, M184I, L210W, T215Y, T215F, T215C, T215D, T215E, T215S, T215I, T215V, K219Q, K219E, K219R
NNRTI	L100I, K101E, K103N, K103S, V106A, V106M, Y181C, Y181I, Y188L, Y188H, Y188C, G190A, G190S, G190E, G190Q, P225H, M230L, P236L

Source: (Shafer, Rhee et al. 2007)

The drug resistance mutations were identified directly from the HIV *pol* sequences. After identification of specific drug resistance mutations, the nucleotide positions associated with drug resistance were deleted prior to phylogenetic analysis to prevent bias from convergent evolution.

All patients with drug resistance mutations who were also recently HIV-infected at diagnosis and/or drug naïve were categorized as having TDR. All other patients with drug resistance mutations were categorized into “acquired” or “transmitted” drug resistance using relevant clinical and diagnostic data. Where data were not available to categorize patients in this way, this is explicitly stated.

2.6. Conclusion

This chapter has described the principles of phylogenetic analysis. It also summarized the laboratory and clinical data that can be linked to HIV *pol* sequences for the phylogenetic reconstruction of transmission events. The methods and limitations of obtaining, linking, analyzing and interpreting these data sources have been described and thesis definitions, where applicable, have been outlined.

In summary, a range of data sources are available that can be combined with HIV *pol* sequences to interpret phylogenetic reconstructions of transmission events. Combining data sources creatively has the potential to enhance our understanding about different aspects of HIV transmission. Specifically, it can help identify which groups are generating HIV transmission, the role of infection stage, viral load and other risk factors, and the transmission of drug resistant viruses. The availability of such datasets is likely to enhance the targeting of public health interventions to reduce HIV transmission. However, two main considerations are apparent: the likely biases resulting from accessing a sequence-based dataset and the appropriateness of linking laboratory and clinical data to interpret phylogenetic reconstructions of HIV transmission events.

3. Chapter Three: Four datasets

This chapter describes and compares the four data sets used in the thesis and outlines the specific considerations in interpreting the data from each.

3.1. Introduction

This thesis uses datasets derived from four populations: the Concerted Action on Seroconversion to AIDS and Death in Europe (CASCADE) study; the Unlinked Anonymous Survey of Sexually Transmitted Infection (STI) Clinic Attendees (the UA STI survey); a population of men who have sex with men (MSM) attending the Brighton HIV clinic; and a random selection of anonymised HIV *pol* sequences taken from MSM attending clinics in Manchester and London. The first two datasets were derived from subsets of established, ongoing surveys. The third population was created for a phylogenetic study on transmission potential of the recently HIV-infected in 2005 (Masquelier, Bhaskaran et al. 2005; Pao, Fisher et al. 2005), but was updated and extended for this thesis. The fourth is an extract taken from an established dataset. This chapter aims to describe and compare the four datasets used in the thesis. It also outlines the specific considerations in interpreting results from each source.

3.2. CASCADE

3.2.1 Background

Established in 1997, CASCADE is an international multi-centre study that monitors events among over 17,000 HIV-infected individuals with well estimated infection dates, throughout the course of their infection (Porter, Babiker et al. 2003). Its aims are to (CASCADE 2009):

1) "Estimate the survival expectations and assess changes over time;
2) Examine the characteristics of recently acquired HIV infection in the population and changes over time;
3) Assess the impact of adverse drug reactions on survival, particularly if

therapy is started close to seroconversion, and monitor changes in the cause of death over time;

4) Examine any changes in the characteristics of the HIV virus over time;
5) Determine the impact of transmitted resistance, virus subtype, and host genetic factors on response to therapy and clinical outcome;
6) Characterise the foci of recent HIV epidemics in Eastern Europe;
7) Characterise initial disease progression in new epidemic areas;
8) Examine the effect of co-infection on HIV disease and response to therapy;
9) Develop new techniques to facilitate the co-ordination of HIV clinical research across European cohorts."

There are 23 cohorts collaborating in CASCADE, taken from 15 European countries, and Australia and Canada. HIV-infected patients are eligible for inclusion provided they are over 15 years old and have a reliably estimated date of infection. Patients are enrolled locally and nationally, and submitted to the CASCADE database, which is maintained in the UK. Patients are typically followed-up throughout the course of their infection. For this thesis, a subset of data was provided comprising drug naive HIV-infected patients with accessible HIV *pol* sequences.

3.2.2 Demographic and clinical data

Clinical and demographic data are collected locally, and submitted to the CASCADE dataset. Clinical data available in the thesis included: exposure category; estimated infection date; and diagnosis date. Demographic data included: age; sex; and country of HIV diagnosis.

3.2.3 Infection stage

For CASCADE, definitions of recently HIV-infected patient included: a) an HIV positive test within three years of an HIV negative test; b) an antibody negative test with polymerase chain reaction (PCR) positivity; or c) an evolving antibody response (Masquelier, Bhaskaran et al. 2005). The estimated infection date was taken as the midpoint between the diagnosis date and the last HIV negative test date. For the laboratory marker, the date the sample was taken was used as a proxy.

3.2.4 Sequences

In each participating cohort, the first plasma sample available after the estimated infection date was taken for genotypic resistance testing. The sequencing methodology has been described previously (Masquelier, Bhaskaran et al. 2005). Participating centres were asked to provide *pol* nucleotide sequence data (the entire protease (PR) gene and at least codons 41-236 of reverse transcriptase (RT)).

3.2.5 Ethics

Ethics approval for cohorts contributing to CASCADE is maintained according to the respective national regulations in the countries concerned.

3.2.6 Dataset specifics

From data pooled in December 2004, 8993 patients from 10 European cohorts, with a reliably estimated date of infection (ranging from 1989-2004) were submitted to the CASCADE dataset. Of these, 677 had complete *pol* sequences, and were from drug naïve patients, taken fewer than 18 months after the estimated date of infection.

In order to ensure the definitions of recent infection were broadly consistent between the three datasets, for the purposes of this thesis, the CASCADE definition for recent infection using the date algorithm was modified. The date of the last negative test and the date of the HIV positive test were available, so it was possible to select a sub-sample for which the interval between the two dates was shorter. Unless stated otherwise in this thesis, patients were defined as being recently infected provided they had an HIV positive test within *90 days* of an HIV negative test. Using this definition, the sample size was reduced from 677 patients defined as seroconverters using the CASCADE definition to 165 patients with recent infection. Thirty four sequences were anti-HIV negative, but PCR positive, and 131 had fewer than 90 days between their HIV positive test date and their last HIV negative test date.

3.3. Unlinked Anonymous Survey of STI clinic attendees

3.3.1 Background

The UA survey (Catchpole, McGarrigle et al. 2000) is part of a family of surveys that make up the Unlinked Anonymous Prevalence Monitoring Programme (Nicoll, Gill et al. 2000). This programme measures the prevalence of HIV using samples left-over from tests undertaken for clinical purposes in accessible population groups. The residual samples are unlinked and anonymised from any patient identifiers before HIV testing. It is impossible to link back an HIV test result to the patient from whom the sample was derived.

Specifically, the UA STI survey has been continuing from 1989 (Gill, Adler et al. 1989) and measures the prevalence of HIV infection (including undiagnosed

HIV infection) among attendees at sentinel STI clinics. It also monitors the uptake of voluntary confidential HIV testing (VCT) in this population (Brown, Tomkins et al. 2006).

Sentinel clinics participate in the survey in England, Wales and Northern Ireland. In each clinic, all attendees undergoing a blood test for syphilis are included in the survey. Patients can be included in the survey up to four times per year, but only once during each calendar quarter. Blood left over from syphilis testing is irreversibly unlinked from patient identifiers and anonymously anti-HIV tested. Limited demographic and clinical data are retained with the sample to assist with interpretation.

Demographic and clinical data

Clinical data consisted of: STI diagnosis; known HIV positive; and VCT received. Demographic data consist of: age-group; world region of birth; year and quarter of attendance.

Using clinical data, patients are categorized by diagnosis status. Samples from patients found to be anti-HIV positive through UA testing are categorized as “previously diagnosed” (diagnosed before the clinic attendance) or “previously undiagnosed”. Samples from “previously undiagnosed” patients are further subdivided into “newly diagnosed” (diagnosed during the clinic attendance) and “undiagnosed” (left clinic attendance unaware of their HIV infection).

3.3.2 Infection stage

Since 1995, samples from MSM found to be HIV-infected with an undiagnosed or newly diagnosed HIV infection have undergone further laboratory testing with the Serological Testing Algorithm for Recent HIV Seroconversion (STARHS) (Janssen, Satten et al. 1998). This has been conducted to provide an annual measure of incidence among this population (Murphy, Parry et al. 2001; Murphy, Charlett et al. 2004).

3.3.3 Sequences

A subset of samples obtained during 1999-2002 from MSM ascertained as recently HIV-infected through STARHS were genotyped for the *pol* gene for a study on transmitted drug resistance (unpublished). The PR region of *pol* was sequenced, along with the first 230 codons of RT (Tatt, Barlow et al. 2004).

3.3.4 Anonymisation

The anonymous nature of the UA technique means no patient identifiers are retained with the sample or the sequence.

3.3.5 Ethics

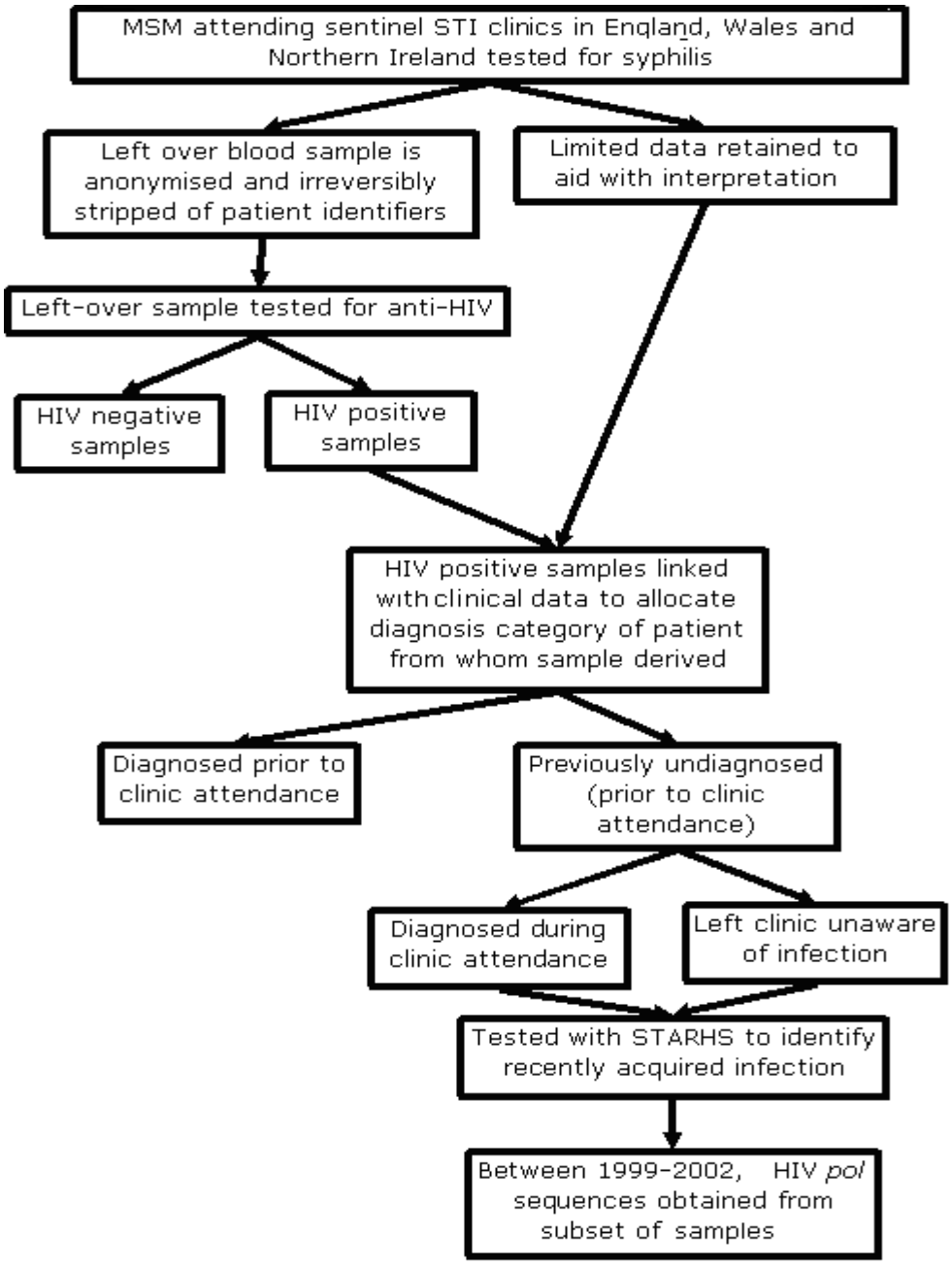
The UA survey has had ethical approval since its inception in 1989 and is compliant with the Human Tissue Act (2004).

3.3.6 Dataset specifics

A flow diagram showing how data were obtained for this thesis is given in **Figure 3.1**. While the UA survey has been continuing since 1989, only the years for which sequences are available (1999-2002) are included in this thesis. Between 1999 and 2002, 28530 MSM attending the sentinel STI clinics were

included in the UA survey **(Figure 3.1).** Of these, 3692 were found to be HIV positive, of whom 1072 had a previously undiagnosed HIV infection (newly diagnosed or undiagnosed HIV infection). Among this group, 21% (229) were found to have been recently-infected through STARHS. Complete PR and RT sequences were obtained for 127 recently HIV-infected patients (samples from the remaining 102 could not be amplified).

Figure 3.1: Flow diagram showing collation of UA STI survey data



3.4. Brighton

3.4.1 Background

Between 2000-2007, the Brighton HIV/STI clinic managed over 90% of the local diagnosed HIV-infected population (personal communication, SOPHID, HPA). The population of clinic attendees therefore represents a relatively comprehensive insight in to a local sexual network.

Since 2000, samples from MSM newly diagnosed with HIV have been referred to be tested for recent infection and sequence-based antiretroviral therapy (ARV) resistance testing. This has enabled the creation of a large *pol* sequence database linkable to information on patient infection stage. This data set was created for a study of the transmission among the recently infected population in 2005 (Pao, Fisher et al. 2005). For this thesis, the database was updated to also include patients attending between 2004-6 and clinical data extended to include plasma viral load and CD4 counts.

3.4.2 Demographic and clinical data

Clinical data were obtained through clinic case notes and electronic patient records. The fields included: evidence of an STI; plasma viral load; CD4 counts; treatment; drug resistance mutations; date of diagnosis. Demographic data collected included: age-group; ethnicity; world region of birth.

Brighton HIV/STI clinic is a collaborator of the UK Collaborative HIV Cohort Study (UK CHIC) (UK-CHIC 2004). This is a UK research programme that follows up clinical details of HIV-infected individuals attending key HIV clinics

over the course of their infection. Clinical details for each diagnosed HIV-infected patient attending Brighton HIV/STI clinic are submitted to UK CHIC annually. Consequently, clinical details were available through linkage to UK CHIC for the vast majority of the HIV-infected attendees at diagnosis and subsequent attendances. This allowed the creation of a longitudinal database capable of following up HIV diagnosed MSM attending the clinic, throughout the study period. Data were updated for patients for each calendar quarter following diagnosis **(Figure 3.2**).

3.4.3 Infection stage

All patients newly diagnosed with HIV since 2000 were referred for laboratory testing: p24 antigen; western blot; and STARHS. STARHS testing was performed using the bioMerieux Vironostika HIV assay as previously described (Pao, Fisher et al. 2005). From clinic notes, HIV testing history was available: patients with 183 days or fewer between their last HIV negative test date and their HIV positive test date could therefore be identified. The precise method for estimating infection dates for each patient from whom a sample was obtained is detailed in section 6.2.1.

3.4.4 Sequences

Brighton HIV/STI clinic collaborates with and submits HIV *pol* sequence data to UK HIV drug resistance database (MRC-Resistance-Database 2009). This national database collates HIV *pol* sequence data annually to estimate the prevalence of drug resistant viruses within untreated patients, and describes the pattern of drug resistance among treated patients. Sequences submitted to the MRC HIV drug resistance database from patients who attended Brighton clinic

between 2000-2006 were extracted for this study. Patients for whom complete sequences were initially unavailable were actively followed up for sequences from the laboratory where the sequence was originally obtained. Where no sequence was available, stored RNA samples taken from the patients around diagnosis were sent to collaborating laboratories for HIV *pol* sequencing. Where no stored RNA sample was available, for the purposes of this study, fresh samples were collected, with patient consent, during routine clinical consultation and sent to collaborating laboratories for sequencing.

For each patient, details were available on the timing of treatment history in relation to the date that the sample was taken for drug resistance testing. Patients with viruses resistant to ARV therefore could be categorized as having acquired or transmitted drug resistance.

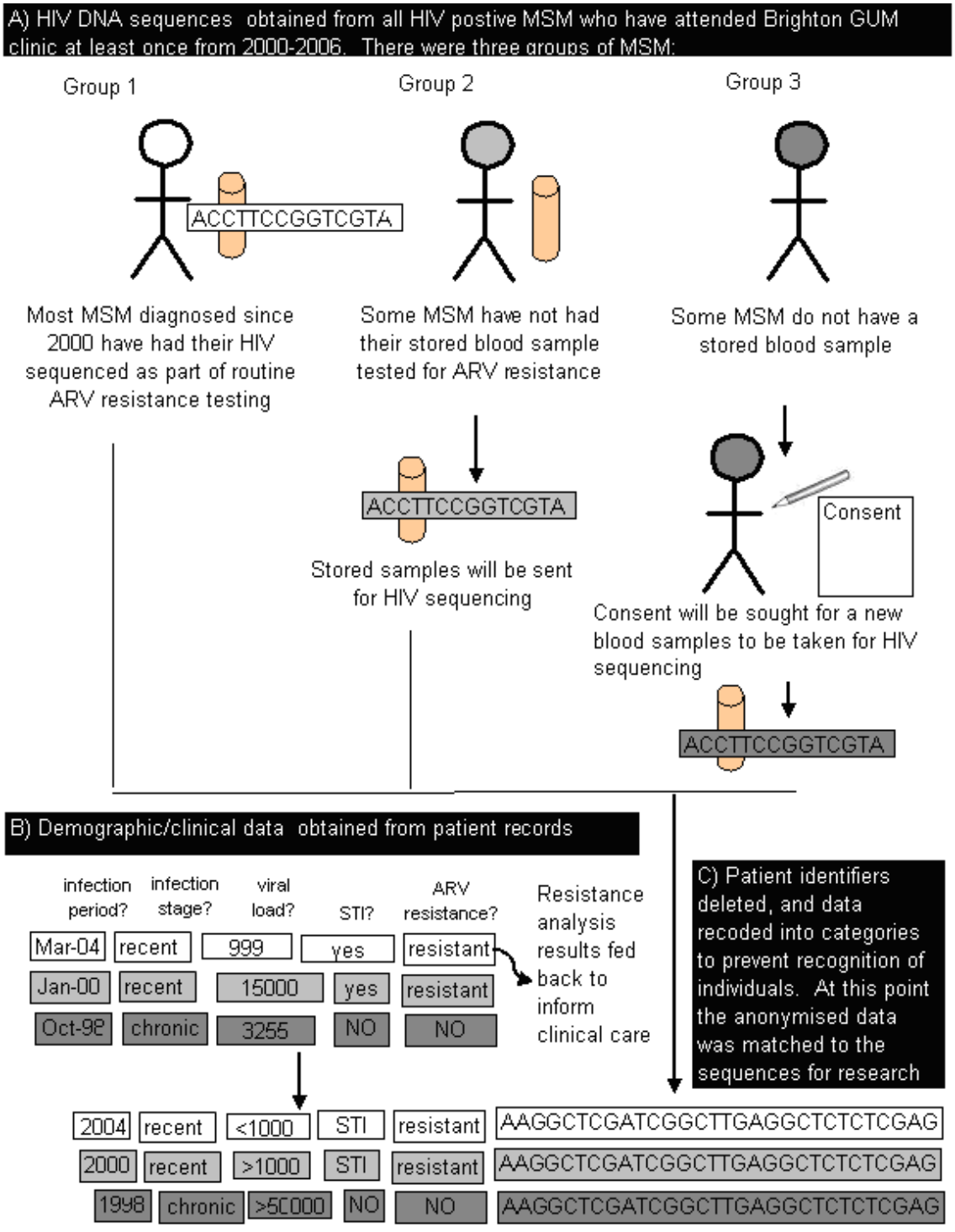
3.4.5 Anonymisation

Patient identifiers were used initially to link clinical records from Brighton HIV/STI clinic to longitudinal clinical and demographic data from UK CHIC study and to sequences stored at the UK HIV drug resistance database. The patient identifiers were then removed and replaced with anonymous ID labels by a third party prior to phylogenetic analysis.

3.4.6 Ethics

The project obtained LREC approval in November 2006 (06/Q1907/93). Patients from whom a fresh DNA sample was sought for the study had written consent obtained. The leaflet and consent letter are given in appendix B.

Figure 3.2: Flow diagram showing collation of Brighton data



3.4.7 Data set specifics

Between 2000 and 2006 1144 HIV-infected patients attended Brighton HIV/STI clinics with clinical and demographic data available. Complete sequences were linkable to 859 of the 1144 HIV-infected patients. Of these, 431 were newly diagnosed during the study period.

In total 159 of 859 HIV-infected patients linkable to sequences were found to be recently infected. Of these 21, were found to be recently infected through p24 antigen, three through western blot, 119 through STARHS and 16 through the date algorithm.

For each patient, data were also available on treatment history, viral loads, CD4 count and any AIDS defining illnesses. These data were updated for each quarter, for each patient, following diagnosis. Where data were not available for a specific quarter, data from the previous quarter were carried forward. Where a patient did not attend for 12 consecutive months, they were considered lost to follow-up and excluded from subsequent calendar quarters.

3.5. HIV sequences from MSM diagnosed in London and Manchester

An extract of 500 *pol* subtype B sequences from MSM diagnosed at a London and 500 from a Manchester clinic were selected at random from an existing dataset (Lewis, Hughes et al. 2008). This dataset has London Multicentre Research Ethics Committee approval (MREC/01/2/10). These data were selected to explore the consistency of the phylogenetic approach in chapter four. Consequently, no additional data were linked to the sequences.

3.6. Comparison of the datasets

The key characteristics of the datasets are summarized in **Table 3.1**. This section compares the attributes of the first three datasets: CASCADE; UA STI survey; and Brighton only. The fourth dataset is not included because no additional data were linked to the sequences.

While the three datasets are each comprised of HIV-infected patients linkable to infection stage information and *pol* sequences, there are considerable differences between them.

3.6.1 Geographical coverage

Each dataset has a different level of geographical coverage: European (CASCADE); national (UA STI); and local (Brighton). In terms of phylogenetic analyses, this may mean there is a higher probability of identifying possible transmission events in the Brighton dataset, since there is a greater chance that patients from the same sexual networks will be included within the population sample. Consequently, comparing the number of transmission events reconstructed between sequences within each dataset will not be meaningful.

Table 3.1: Key characteristics of the four datasets

Data source	Geographical coverage	Number of *pol* sequences	Years sampled	Sample characteristics	Additional data	Method use to ascertain recent infection	Estimation of infection date
CASCADE	11 European cohorts.	677 or 165*	1989-2004	Exclusively recently HIV-infected. Mainly MSM/IDUs.	Risk group, age-group, country	P24 antigen, western blot and interval of <90 days between last negative test and HIV diagnosis.	Midpoint of HIV negative and positive test is used. For laboratory evidence, sample date is taken.
UA survey	15 clinics in England Wales and Northern Ireland.	127	1999-2002	Exclusively recently HIV-infected MSM. A proportion are undiagnosed	Undiagnosed or newly diagnosed. VCT uptake. STI diagnosis, age-group, attendance quarter, world region of birth.	STARHS	Quarter of attendance is taken as a proxy.
Brighton HIV/STI clinic	One local HIV/STI clinic.	859	2000-2006	Both recently and chronically HIV-infected MSM attending Brighton HIV/STI clinic	Infection stage, STI, age-group, WRB, diagnosis quarter, treatment history.	P24 antigen, western blot, STARHS and interval of <90 days between last negative test and HIV diagnosis.	See section 6.2.1
Randomly selected sequences	MSM diagnosed in clinics in London /Manchester.	500 from London and 500 from Manchester.	N/A	Subtype B and from MSM	None.	N/A	N/A

* depending on the definition of recent HIV infection used (see section 3.2.6).

3.6.2 Demographic data

The datasets have different demographic variables available. All three datasets included age-group and exposure group. While CASCADE includes data on country of diagnosis, it does not contain data on world region of birth or ethnicity. The UA survey and the Brighton data set contain data on world region of birth, with the Brighton dataset also containing data on ethnicity.

3.6.3 Clinical data

The level of clinical data available also varies between the datasets. Only the diagnosis date and exposure group is available for CASCADE (all patients are treatment naïve). The UA survey and Brighton both contain data on diagnosis date, and STI diagnosis. The UA survey is likely to be more comprehensive in terms of the STI data since the majority of patients are likely to be screened for an STI. For Brighton, only evidence of an STI diagnosis is available, since patients may not be screened at every attendance. The UA survey is able to categorize HIV-infected patients into the undiagnosed and the newly diagnosed. It also contains information on which patients received VCT.

The Brighton data set contains comprehensive data for markers of disease progression and treatment (plasma viral load, CD4 count, AIDS defining illnesses, and treatment (treatment native, treated and treatment interruption)). Additionally, unlike the UA survey and CASCADE, these variables are updated for each patient, by calendar quarter throughout the study period.

3.6.4 Infection stage

The methods used to ascertain recent infection vary between the three datasets, and are shown in **Table 3.1**. Since each method for ascertaining recent infection is different, comparing "recently HIV-infected patients" between data sets may not be entirely valid.

The UA survey is the only study for which the marker used to identify recent infection is consistent within the dataset. Both CASCADE and Brighton use a mixture of techniques to identify recent infection. This has implications for the accuracy and consistency of ascertaining "infection dates" for individuals within datasets, and the extent to which overall comparisons can be made both between and within studies.

3.6.5 Sequences

The number of sequences available for each data set varies **(Table 3.1)**. In order that phylogenetic analysis is not biased by convergent evolution, it is important that the sequences are taken from the drug naïve. For CASCADE, only sequences from the drug naïve were provided. For the UA survey, only samples from patients who were undiagnosed, or newly diagnosed were included – consequently all sequences should be from the treatment naïve. However, there is a possibility that patients with a diagnosed HIV infection may have attended one of the participating clinics for STI screening, and not revealed their HIV status. For Brighton data, information was available on the date the sequence was obtained, and the date the patient initiated treatment. It is therefore possible to verify which sequences are from the treatment naïve.

Since sequences are not available for every patient, it is important to consider whether patients with complete *pol* sequences differ in any substantial way to patients for whom no HIV sequence data are available **(Table 3.2).** For the UA STI survey, the demographic make up between the 127 recently HIV-infected MSM with HIV *pol* sequences and the 1072 previously undiagnosed HIV-infected MSM without HIV *pol* sequences were similar: the median age-group for both groups was 25-34 years and the proportion of non-UK born MSM was 40% (51) and 38% (407) respectively.

Table 3.2: Differences between HIV-infected patients with and without HIV *pol* sequences, UA survey and Brighton

Attributes	**UA survey (previously undiagnosed)**		**Brighton**	
	Sequence available N= 127 (%)	Sequence not available N= 1072 (%)	Sequence available N= 859 (%)	Sequence not available N= 285 (%)
Median age/age-group	25-34	25-34	38	39
Non-UK born/non UK national	51 (40.2)	407 (38.0)	239 (27.8)	76 (26.7)
London clinic	109 (87.2)	836 (78.0)	-	-
STI	47 (27.6)	501 (36.3%)	66 (7.6%)	13 (4.6)
Proportion undiagnosed	70 (55.2)	639 (59.7)	-	-
Recently HIV-infected	-	-	159 (18.5)	30 (10.5)
Median plasma viral load copies/mL	-	-	41,180	11,425
Median CD4 count (cells/mm^3)	-	-	392	404

For the Brighton data, the demographic attributes between the MSM at diagnosis with (n=859) and without (n=285) HIV *pol* sequences were also similar: for instance median age was 38 vs. 39 respectively. However, the clinical markers at diagnosis varied between the two groups. A higher proportion of MSM with HIV *pol* sequences had an STI (7.6%) and a higher average plasma viral load at diagnosis (41,180 copies/mL). The equivalent

figures for MSM without HIV *pol* sequences were 4.5% and 11,425 copies/mL. The elevated plasma viral load at diagnosis is likely to reflect the higher proportion of this population who were recently HIV-infected at diagnosis: 18.5% vs. 10.5% among MSM with and without HIV *pol* sequences respectively. This is because the proportion of patients with a recent infection at diagnosis increased between 2000-2006 (section 6.3.4); a time when the proportion of newly diagnosed MSM who had samples taken for HIV drug resistance testing also increased.

3.7. Conclusion

Four data sources are used in this thesis. Of these, three contain clinical and demographic data linkable to HIV *pol* sequences and are either entirely constituted from individuals with "recent infection" or have data on infection stage. Despite the similarities, substantial differences remain in relation to geographic distribution, the number of sequences and the availability of data. The fourth dataset contains no other data than subtype B *pol* sequences. The specific attributes of each dataset require consideration during study design and interpretation, and when making comparisons within and between datasets.

4. Chapter Four: The consistency of phylogenetic reconstructions of HIV transmission events.

This chapter assesses the accuracy of HIV transmission reconstructions in relation to their consistency. It uses two anonymous datasets of *pol* sequences from men who have sex with men, one from London and the other from Manchester. For each data set an initial phylogenetic reconstruction was undertaken, and the number of possible transmission events recorded. The phylogenetic reconstruction was repeated for each dataset with the sample size and evolutionary models used to construct the tree varied. In each instance, the number of initial transmission events retained, and novel transmission events created, were identified and described.

4.1. Introduction

The assumption that phylogenetic reconstructions of HIV transmission events are accurate is critical to this thesis and the increasing number of phylogenetic analyses of HIV *pol* sequences conducted for public health purposes (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007; Lewis, Hughes et al. 2008). The suitability of phylogenetics for public health purposes has not been investigated thoroughly, and inaccurate reconstructions have been reported (section 1.6.1).

It is not possible to verify the transmission events identified through phylogenetic reconstructions through linkage to sexual histories on a large scale. However, it is possible to perform sensitivity analyses to assess the consistency of the phylogenetic approach. This chapter takes HIV *pol* sequences from two population samples: MSM diagnosed in London and in Manchester. It assesses the consistency of the phylogenetic clusters formed within each dataset and describes how clusters change under varied sample sizes and models of evolution.

4.2. Method

This chapter used two anonymised datasets: HIV *pol* sequences taken from 500 MSM diagnosed with HIV in London, and 500 MSM diagnosed with HIV in Manchester (section 3.5). Whilst the specific sequence sources are not known, the datasets were derived from the data utilized in Lewis *et al*. (Lewis, Hughes et al. 2008). Drug resistance mutation sites (Shafer, Rhee et al. 2007) were deleted prior to reconstruction. Initial phylogenetic trees were constructed for both the London and Manchester datasets using a neighbour joining tree with

gamma rate heterogeneity set at 0.5 and 500 replicates of bootstrapping. The two sequence alignments were each run through ModelTest to ascertain the sample heterogeneity.

Using the initial neighbour joining trees, possible transmission events were identified and categorised as robust, medium or weak. The definitions were as follows - two or more sequences that share a common node and:

- **Robust:** bootstrap support of =>99% and an average genetic distance of <0.015 nucleotide substitutions per site (this is the definition used throughout the thesis – section 2.2.11);
- **Medium:** bootstrap support of >95% and <99% and a genetic distance of >0.015 and <0.03 nucleotide substitutions per site;
- **Weak:** bootstrap support of >90% and <95% and a genetic distance of <0.05 and >0.03 nucleotide substitutions per site.

Where the bootstrap and genetic distance value would have placed the cluster into different categories (e.g. a bootstrap support of 97% and a genetic distance of 0.014), the "weaker" marker was used to categorise the cluster (e.g. in the example above, the cluster would be categorised as “medium”).

Phylogenetic reconstruction was repeated with each of the two datasets, each time varying the sample size or evolutionary models applied. The number of original transmission events identified in the initial tree was noted, and in each subsequent reconstruction, the number of original transmission events that were maintained in the subsequent analyses, and any novel transmission events added, were recorded.

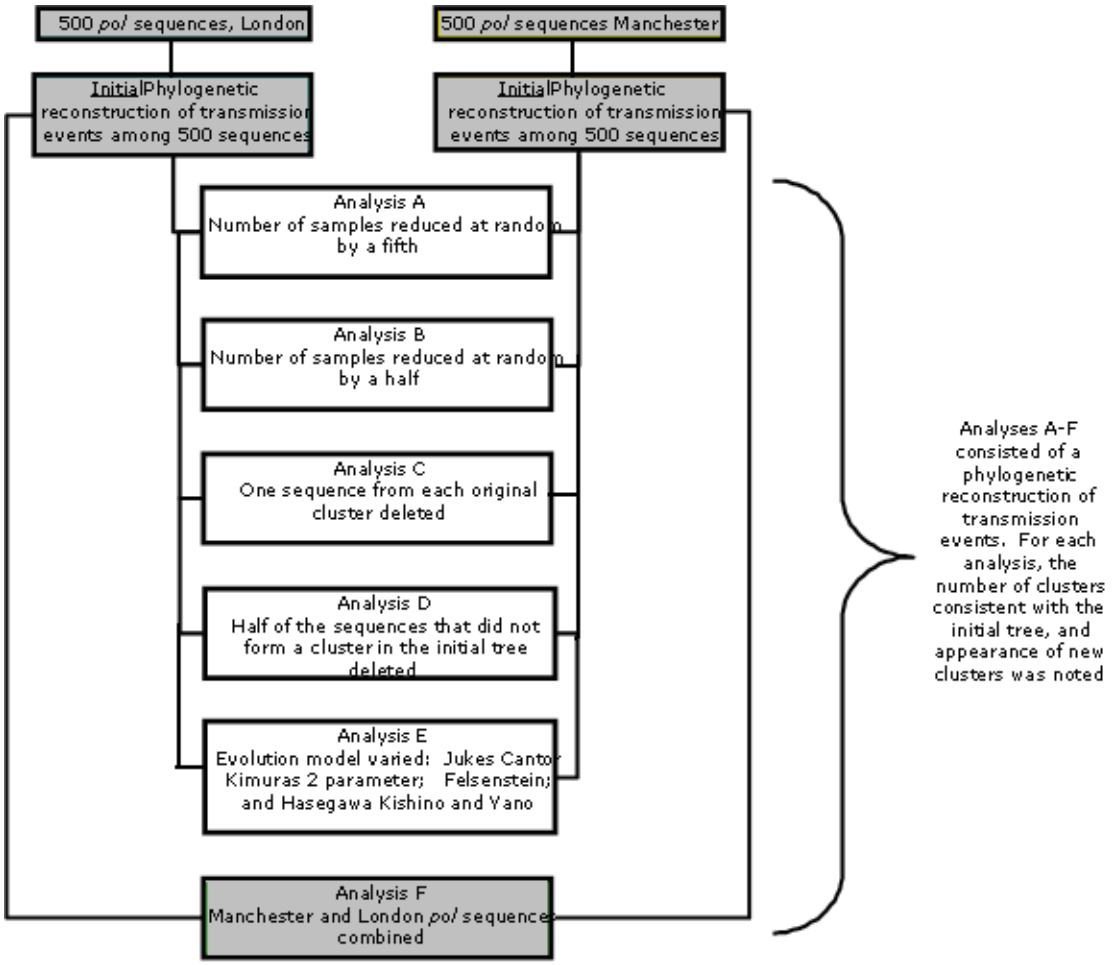
The variations were as follows (**see Figure 4.1**):

- **Analysis A:** Number of sequences included in each population sample reduced by a fifth at random (sequence labels were sorted numerically, and every fifth sequence deleted);
- **Analysis B:** Number of sequences included in each population reduced by half at random (sequence labels were sorted numerically and every other sequence was deleted);
- **Analysis C:** Number of sequences included in each population sample reduced non randomly (one sequence from each initial robust cluster deleted);
- **Analysis D:** Number of sequences included in each dataset reduced non randomly (half of the initial sequences that did not form a robust cluster deleted);
- **Analysis E**: Neighbour joining tree constructed under different models of evolution: General time reversible model (GTR); Jukes Cantor (JC)(Jukes 1969); Kimuras 2 parameter (K2P)(Kimura 1980); Felsenstein (F81)(Felsenstein 1981), and Hasegawa Kishino and Yano (HKY85)(Hasegawa, Kishino et al. 1985);
- **Analysis F:** Combining the sequences from Manchester and London.

The sample sizes were reduced in order to replicate the effect of partial representation of an HIV-infected population on phylogenetic reconstructions of transmission events within a population. The initial tree was used as a proxy for the entirety of sequences circulating in the population, and the reductions in sample size undertaken to mimic the effect of sampling within this population.

The random reduction was used to mimic the effect of random sampling. The non-random reduction (i.e. deleting one sequence from each cluster) was used to assess the potential effect of not including sequences that were part of actual transmission chains (i.e. when robust clusters are disrupted by the deletion of one sequence, does the remaining sequence "erroneously cluster" with another sequence?). The non-clustering sequences were deleted to see if the formation of the "true clusters" were dependent on the presence of other sequences in the sample, or whether the transmission events were absolute.

Figure 4.1: Flow diagram outlining phylogenetic analyses undertaken in chapter four



4.3. Results

4.3.1 Initial trees

London

An initial neighbour joining tree was constructed using the London dataset and subsequently run through ModelTest **(Figure 4.2**). ModelTest calculated a London kappa score 7.57 and gamma shape 0.76. Out of the 500 London sequences, 50 sequences formed 23 robust clusters (numbered L1-23), **(Figure 4.2)**. Twenty clusters consisted of sequence pairs, two of sequence triplets (L14 and L18) and one cluster made from four sequences (L12). The clusters are summarised in **Table 4.1**.

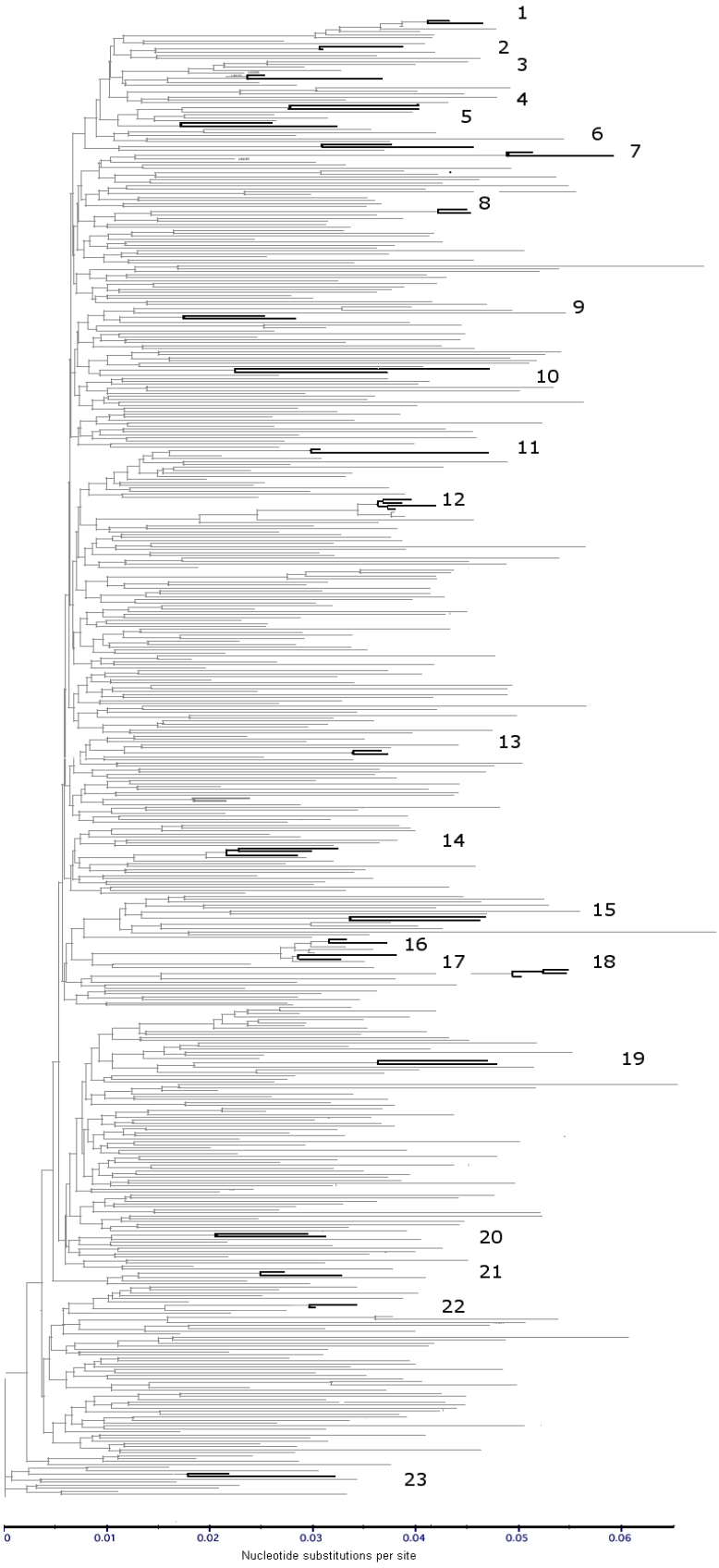
The criteria were relaxed to identify clusters that formed under the definition of a medium cluster (>95% and <99% bootstrap, and <0.03 and >0.015 nucleotide substitutions per site). This resulted in an additional five clusters (numbered L24-7 and L35). Three consisted of sequence pairs (L25, L26 and L35) and one triplet (L24) and one comprised eight sequences (L27).

Under the definition of a weak cluster (>90% and <95% bootstrap and between 0.05-0.3, nucleotide substitutions per site) an additional four clusters were added (numbered L28-L31). Three clusters consisting of sequence pairs (L28, L29 and L31), and one cluster consisting of five sequences (L30).

Table 4.1: Robust, medium and weak clusters ascertained through phylogenetic reconstruction, London

Cluster type	Cluster ID	Number of sequences per cluster
Robust clusters	L1	2
	L2	2
	L3	2
	L4	2
	L5	2
	L6	2
	L7	2
	L8	2
	L9	2
	L10	2
	L11	2
	L12	4
	L13	2
	L14	3
	L15	2
	L16	2
	L17	2
	L18	3
	L19	2
	L20	2
	L21	2
	L22	2
	L23	2
Total = 23		Total=50
Medium clusters	L24	3
	L25	2
	L26	2
	L35	2
	L27	8
Total=5		Total=17
Weaker clusters	L28	2
	L29	2
	L30	5
	L31	2
Total=4		Total=11

Figure 4.2: Phylogenetic reconstruction of HIV transmission events, initial tree, London



Clusters L1-23 highlighted in bold.

Figure 4.3: Phylogenetic reconstruction of HIV transmission events: initial tree, Manchester

Clusters M1-12 highlighted in bold.

Table 4.2: Robust, medium and weak clusters ascertained through phylogenetic reconstruction, Manchester

Cluster type	Cluster ID	Number of sequences per cluster
Robust	M1	2
	M2	2
	M3	2
	M4	2
	M5	2
	M6	2
	M7	2
	M8	2
	M9	2
	M10	2
	M11	2
	M12	2
Total=12		Total=24
Medium	M13	2
	M14	2
	M15	2
	M16	2
	M17	7
	M18	2
	M19	2
	M20	2
Total=8		Total=21
Weak	M21	2
	M22	2
	M23	2
	M24	2
Total=4		Total=8

Manchester

An initial neighbour joining tree was constructed for the Manchester dataset and run through ModelTest. ModelTest calculated the Manchester kappa score to be 7.18 and the gamma shape to be 0.35. Out of the 500 Manchester sequences, 24 sequences formed 12 robust clusters (numbered M1-M12) **Figure 4.3.** They all consisted of sequence pairs and are summarised in **Table 4.2.**

Under the definition of medium clusters, eight clusters were apparent (M13-M20). Seven clusters consisted of sequence pairs and one consisted of seven sequences (Cluster M17). Under the weak cluster definition a further four clusters were added (clusters M21-24). Cluster M22 comprised three sequences, with the remainder of clusters consisting of sequence pairs.

4.3.2 Reducing sample sizes

An overall summary of results from analyses A-F is presented in **Table 4.3.**

London

Analysis A

The number of sequences within the dataset was reduced by a fifth at random. Overall, 100 sequences were deleted, seven of which had been part of robust clusters in the initial tree (clusters L11, L12, L13, L15, L16, L20, and 21) **(Table 4.4**). Of the 400 remaining sequences, 30 formed 15 robust clusters, 14 of which were identical to the original tree **(Table 4.4**). Excluding the 7/23 robust clusters that were affected by the deletion, 87.5% (14/16) of the robust clusters found in the initial tree were also found in the analysis A **(Table 4.3**).

Clusters L1 and L18, robust clusters in the initial tree appeared as medium clusters in analysis A. Cluster 32 appeared as a novel robust cluster (i.e. it did not appear as a robust, medium or weak cluster in the initial analysis).

Analysis B

The number of sequences within the dataset was then randomly reduced further so that half of the original 500 sequences were deleted. A further 150

sequences were deleted, so that 13 of the robust clusters in the initial tree had a sequence within that cluster, deleted. Overall, of the 250 sequences deleted, 22 had formed 14 robust clusters in the initial tree (clusters L2, L5, L6, L8, L11, L12, L13, L15, L16, L19, L20, L21, L22, and L23) **(Table 4.4).** While one sequence making up cluster 18 was deleted, the initial robust cluster contained three sequences in the initial tree, consequently two of the original sequences remained.

Of the remaining 250 sequences, nine robust clusters were identified, comprised of eighteen sequence pairs. Of these, eight were identical to robust clusters found in the initial tree (clusters L3, L4, L7, L9, L10, L14, L17, and L18). Excluding 14/23 robust clusters affected by the deletions, 88.9% (8/9) of the robust clusters found in the initial tree were also found in analysis B **(Table 4.3).**

Robust cluster L32 was novel, but also had appeared in analysis A. One robust cluster present in the initial tree (L1) was apparent as a medium cluster in analysis B. L18, appeared as a robust cluster in this analysis, but was a medium cluster in analysis A.

Analysis C

One sequence from each of the initial clusters was deleted. In total, 24 sequences were deleted, one from each robust cluster, leaving 476 sequences remaining. Of the remaining 476 sequences, four robust clusters were formed, comprising 13 sequences **(Table 4.4**). Excluding 21/23 robust clusters affected by the deletions, 100% (2/2) of the robust clusters found in the initial tree were

also found in analysis C **(Table 4.3**) – clusters L12 and L18. Both of these clusters had more than two sequences within the cluster in the initial tree, and the remaining sequences had clustered. Cluster L27, a robust cluster in analysis C, was a medium cluster in the initial analysis. Cluster L33 was a novel robust cluster in analysis C.

Analysis D

Half of the sequences that did not form a cluster in the initial tree were deleted. In total 214 sequences were deleted, leaving 286 sequences. Twenty three robust clusters were found in this analysis. No robust clusters in the initial analysis were affected by deletions and 91% (21/23) of the robust clusters found in the initial tree was also found in analysis D **(Table 4.3).**

The remaining two robust clusters (L28 and L29) in analysis D were found as a weak and a medium cluster in the initial tree respectively. The two robust clusters in the initial tree not found in analysis D were L1 and L11. For analyses A-D L1 had not been present as a robust cluster, and L11 had been affected by deletions in analyses A-C. In all instances, where a sequence had been deleted that had been part of a robust cluster in the initial analysis, the remaining sequence did not form a cluster (robust, medium or weak) with any other sequence in subsequent analyses.

Manchester

Analysis A

The Manchester dataset was reduced by a fifth at random. Overall, 100 sequences were deleted, including seven sequences that had originally formed

a robust cluster in the initial tree (clusters M2, M3, M5, M6, M7, M8, and M12). Of the 400 remaining sequences, ten formed five robust clusters, four of which were identical to the initial tree **(Table 4.5**). Excluding the 7/12 robust clusters that were affected by the deletion, 80% (4/5) of the robust clusters found in the initial tree were also found in analysis A. Cluster M9, a robust cluster in the initial reconstruction did not appear as a robust cluster in the present analysis. Robust cluster M25 appeared as a novel cluster in this analysis **(Table 7.4).**

Analysis B

The number of sequences within the dataset was then randomly reduced further so that half of the original sample was deleted. In total 250 sequences were deleted, of which ten had formed a robust cluster in the initial tree (clusters M2-8, 10-12). Of the remaining 250 sequences, three robust clusters were retained, both consisting of sequence pairs. Excluding the 10/12 robust clusters that were affected by the deletion, 100% (2/2) of the robust clusters found in the initial tree were also found in the analysis B **(Table 4.3)**. One novel robust cluster was formed (cluster M25); the same cluster that appeared in analysis A.

Analysis C

One sequence from each of the initial robust clusters was deleted. In total 27 sequences were deleted, one from each cluster, leaving 473 sequences. Of the remaining 473 sequences, no robust cluster was formed.

Table 4.3: Overall consistency of robust clusters between analyses A-F, London and Manchester

London	Number of robust clusters in each analysis	Proportion consistent with initial tree x%(n/N)*	Number of robust clusters that appeared as medium/weak in initial tree	Number of novel clusters
Initial tree	23	-	-	-
Analysis A	15	87.5% (14/16)	0	1
Analysis B	9	88.9%(8/9)	0	1
Analysis C	4	100% (2/2)	1	1
Analysis D	23	91% (21/23)	2	0
Analysis E				
JC	23	95.7% (22/23)	0	1
K2P	23	95.7% (22/23)	0	1
F81	24	95.7% (22/23)	1	1
HKY85	21	87.0% (20/23)	0	1
Manchester				
Initial tree	12	-	-	-
Analysis A	5	80% (4/5)	0	1
Analysis B	3	100% (2/2)	0	1
Analysis C	0	N/A	0	0
Analysis D	13	100% (12/12)	1	0
Analysis E				
JC	12	91.7% (11/12)	1	0
K2P	11	83.3% (10/12)	1	0
F81	12	100% (12/12)	0	0
HKY85	13	91.7% (11/12)	2	0
Analysis F**	21	45.7% (16/35)	2	3

*Where n= number of robust clusters present in initial tree and in present analysis and N= total number of robust clusters in initial analysis. For analyses A-D, the denominator excludes the clusters that were disrupted by deletions.

** Both Manchester and London datasets combined. Total number of robust clusters in initial tree derived by adding London and Manchester robust clusters together.

Table 4.4: Stability of robust clusters under varying samples sizes, London

	Random deletions		Non random deletions	
Robust clusters	**Analysis A Fifth deleted**	**Analysis B Half deleted**	**Analysis C One sequence form each initial cluster deleted**	**Analysis D Half of non-clustering sequences deleted**
L1	No	No	N/A	No
L2	Yes	N/A	N/A	Yes
L3	Yes	Yes	N/A	Yes
L4	Yes	Yes	N/A	Yes
L5	Yes	N/A	N/A	Yes
L6	Yes	N/A	N/A	Yes
L7	Yes	Yes	N/A	Yes
L8	Yes	N/A	N/A	Yes
L9	Yes	Yes	N/A	Yes
L10	Yes	Yes	N/A	Yes
L11	N/A	N/A	N/A	No
L12	N/A	N/A	Yes	Yes
L13	N/A	N/A	N/A	Yes
L14	Yes	Yes	N/A	Yes
L15	N/A	N/A	N/A	Yes
L16	N/A	N/A	N/A	Yes
L17	Yes	Yes	N/A	Yes
L18	No	Yes	Yes	Yes
L19	Yes	N/A	N/A	Yes
L20	N/A	N/A	N/A	Yes
L21	N/A	N/A	N/A	Yes
L22	Yes	N/A	N/A	Yes
L23	Yes	N/A	N/A	Yes
L27*	N/A	N/A	Yes	No
L28*	N/A	N/A	N/A	Yes
L29*	N/A	N/A	N/A	Yes
L32*	Yes	Yes	No	No
L33*	No	No	Yes	No

N/A=sequences that constitute original clusters not available in present reconstruction
*Clusters present as robust clusters in analyses A-D were not present as robust clusters in initial tree, cluster L27 was present as a medium cluster, L28 and L29 were weaker clusters and clusters L32-3 are novel clusters

Table 4.5: Stability of robust clusters under varying samples sizes, Manchester

	Random deletions		Non random deletions	
Robust clusters	**Analysis A Fifth deleted**	**Analysis B Half deleted**	**Analysis C One sequence from each initial cluster deleted**	**Analysis D Half of non-clustering sequences deleted**
M1	Yes	N/A	N/A	Yes
M2	N/A	N/A	N/A	Yes
M3	N/A	N/A	N/A	Yes
M4	Yes	N/A	N/A	Yes
M5	N/A	N/A	N/A	Yes
M6	N/A	N/A	N/A	Yes
M7	N/A	N/A	N/A	Yes
M8	N/A	N/A	N/A	Yes
M9	No	Yes	N/A	Yes
M10	Yes	N/A	N/A	Yes
M11	Yes	Yes	N/A	Yes
M12	N/A	N/A	N/A	Yes
M17*	No	No	No	Yes
M25*	Yes	Yes	No	No

Notes:
N/A=sequences that constitute original cluster not available in present reconstruction.
*Clusters present as robust clusters in analyses A-D were not present as robust clusters in initial tree, cluster M17 was present as a medium cluster, and M25 appeared as a novel cluster in this reconstruction.

Analysis D

Half of the sequences that did not form a cluster in the initial tree were deleted. In total 223 sequences were deleted, leaving 277 sequences. With this dataset, 13 robust clusters were formed (clusters M1-12, and M17) from 27 sequences. No robust clusters in the initial tree were affected by the deletions of analysis D, with 100% (12/12) robust clusters in the initial tree also found in analysis D. The robust cluster M17, which was comprised of seven sequences, was identified as a medium cluster in the initial tree.

In all instances, where a sequence had been deleted that had been part of a robust cluster in the initial analysis, the remaining sequence did not form any other cluster (robust, medium or weak) in subsequent analyses.

4.3.3 Models of evolution

London – analysis E

Neighbour joining trees were constructed under four different models of evolution: JC; K2P; F81; and HKY85, for the London datasets **(Table 4.6**).

For the JC and K2P models gave similar results. All of the robust clusters in the initial tree were retained with the exception of cluster L1 – 95.7% (22/23). Cluster L34 appeared as a new robust cluster in the present reconstruction; this cluster was not apparent as a robust, medium or weak cluster in the initial analysis or analyses A-D.

Table 4.6: Stability of robust clusters under different models of evolution, London.

Cluster	JC	K2P	F81	HKY85
L1	No	No	No	No
L2	Yes	Yes	Yes	Yes
L3	Yes	Yes	Yes	Yes
L4	Yes	Yes	Yes	Yes
L5	Yes	Yes	Yes	Yes
L6	Yes	Yes	Yes	Yes
L7	Yes	Yes	Yes	Yes
L8	Yes	Yes	Yes	Yes
L9	Yes	Yes	Yes	Yes
L10	Yes	Yes	Yes	Yes
L11	Yes	Yes	Yes	Yes
L12	Yes	Yes	Yes	Yes
L13	Yes	Yes	Yes	Yes
L14	Yes	Yes	Yes	Yes
L15	Yes	Yes	Yes	Yes
L16	Yes	Yes	Yes	No
L17	Yes	Yes	Yes	Yes
L18	Yes	Yes	Yes	Yes
L19	Yes	Yes	Yes	Yes
L20	Yes	Yes	Yes	Yes
L21	Yes	Yes	Yes	Yes
L22	Yes	Yes	Yes	No
L23	Yes	Yes	Yes	Yes
L34*	Yes	Yes	Yes	Yes
L35*	No	No	Yes	No

*Clusters present as robust clusters in analysis E were not present as robust clusters in initial tree, M35 was a medium cluster in the initial analysis and M34 appeared as a novel cluster in this reconstruction.

Under the F81 model, the same observations were found as with the JC and K2P model giving a consistency of 95.7% (22/23) – **Table 4.3.** However, cluster L35, a medium cluster found in the initial tree, was found as a robust cluster in the present reconstruction. Under the HKY85 reconstruction, robust clusters L1, 16 and 22 found in the initial tree were not present in the current reconstruction giving a consistency of 87.0% (20/23). As with the JC, KP2 and F81 models, the novel, but robust cluster L34 appeared.

Manchester – analysis E

Neighbour joining trees were constructed under four different models of evolution: JC; K2P; F81; and HKY85 for the Manchester dataset **(Table 4.7).**

Table 4.7: Stability of robust clusters under different models of evolution, Manchester.

Cluster	JC	K2P	F81	HKY85
M1	Yes	Yes	Yes	Yes
M2	Yes	Yes	Yes	Yes
M3	Yes	Yes	Yes	Yes
M4	Yes	No	Yes	No
M5	No	No	Yes	Yes
M6	Yes	Yes	Yes	Yes
M7	Yes	Yes	Yes	No
M8	Yes	Yes	Yes	Yes
M9	Yes	Yes	Yes	Yes
M10	Yes	Yes	Yes	Yes
M11	Yes	Yes	Yes	Yes
M12	Yes	Yes	Yes	Yes
M18*	Yes	Yes	No	Yes
M20*	No	No	No	Yes

*Clusters present as robust clusters in analysis E were not present as robust clusters in initial tree, clusters 18 and 20 were present as medium clusters in the initial tree.

Under the JC model, all of the robust clusters in the initial tree were present, with the exception of cluster M5 – this gave a consistency of 91.7% (11/12) – **Table 4.3.** In addition, cluster M18, which appeared as a medium cluster in the initial tree, appeared as a robust cluster in present reconstruction.

Under the K2P reconstruction, all of the robust clusters in the initial tree were present with the exception of clusters M5 and M4 giving a consistency .of 83.3% (10/12). Cluster M18 that was apparent in the initial tree as a medium cluster was apparent as a robust cluster in the present analysis.

For the F81 analysis, all of the robust clusters apparent in the initial analysis were present in the current analysis – (100% 12/12), and no additional clusters appeared. For the HKY85 model, cluster M4 was not apparent in the current analysis, and clusters M18 and M20, which were medium clusters in the initial analysis, appeared as robust clusters in the present analysis.

4.3.4 Combining datasets

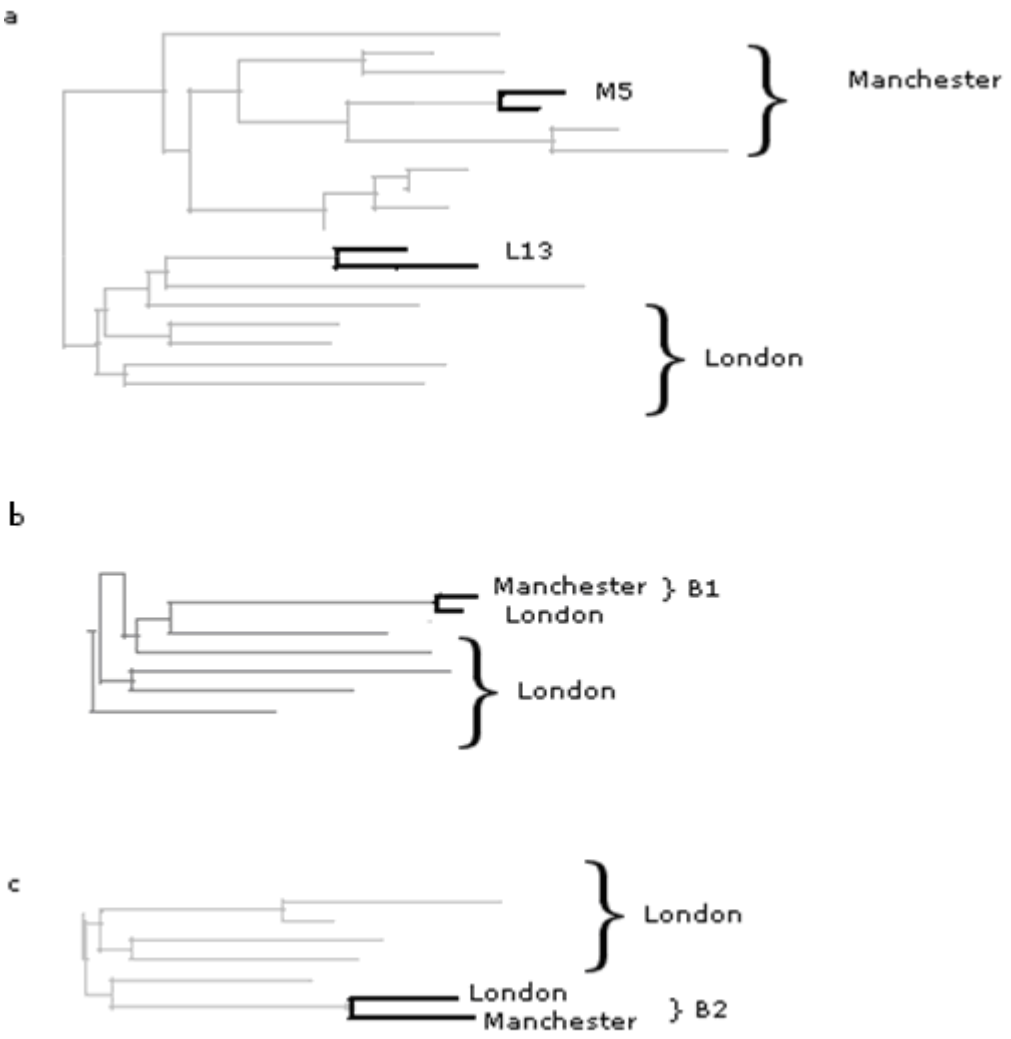
Analysis F

The 500 sequences from Manchester, and 500 from London were combined and a phylogenetic reconstruction undertaken using a neighbour joining methodology.

Of the 1000 sequences, 57 formed 23 robust clusters **(Figure 4.4a -** full tree not shown due to size)**. Table 4.8a** displays the robust clusters in the combined tree that were also robust clusters in the initial tree. Overall 35 robust clusters were found in the initial London and Manchester trees. Of these, less than half (20) were found as robust clusters in the combined tree **(Table 4.3).**

For London, 11 robust clusters were identical to those found in the London initial tree (L1, L2, L5, L7, L11, L12, L13, L14, L18, L21 and L22). The remaining 12 clusters that were robust in the initial London tree were apparent as medium or weak clusters in the combined tree. Of the 23 robust clusters in the initial London tree 47.8% (11/23) were identical London robust clusters in analysis F. Cluster L18 had one additional Manchester sequence added to it – it did not alter cluster the topology of L18.

Figure 4.4: Extracts from the phylogenetic reconstruction of HIV transmission events, combining sequences from London and Manchester



For the Manchester sequences, seven of the 12 robust clusters in the initial robust tree were also robust clusters in the combined tree (58.3% - M3, M4, M6, M8, M9, M10, M11, and M12). The remaining five clusters that were robust clusters in the initial Manchester tree (M1, M4, M5, M6, and M7) appeared as weaker or medium clusters in the combined tree. Cluster M11 had an additional London sequence added to it, which did not disrupt the topology of the cluster.

Table 7.8b shows robust clusters in analysis F that were not robust clusters in the initial tree. Two entirely new clusters appeared, clusters B1 and B2 **(Figure 4.4a and b**): these were each comprised of a London and Manchester sequence paired together. These sequences did not form any cluster (including weak) with any other sequences in the initial tree or analyses A-E. Clusters L24, L27 and L32 appeared as robust clusters in this analysis, but the former two were medium clusters in the initial analysis and L32 was novel.

Table 4.8: Number of robust clusters ascertained through phylogenetic analysis combining sequences from London and Manchester

a) Initial robust clusters maintained

Initial clusters	Present in combined tree?
L1	yes
L2	yes
L3	yes
L4	yes
L5	no
L6	yes
L7	yes
L8	yes
L9	no
L10	no
L11	no
L12	no
L13	yes
L14	no
L15	no
L16	no
L17	no
L18	no
L19*	yes
L20	no
L21	no
L22	no
L23	no
Total=23	Total=9
M1*	yes
M2	no
M3	no
M4	no
M5	yes
M6	no
M7	yes
M8	yes
M9	yes
M10	yes
M11	yes
M12	no
Total=12	total=7

b) Novel robust clusters

Novel clusters	Notes
L24	Present as medium cluster in initial tree
L27	Present as medium cluster in initial tree
L32	Not present in initial tree
B1	Two sequences, one from Manchester one from London
B2	Two sequences, one from Manchester and one from London

*Cluster L19 was maintained but incorporated a sequence from Manchester. Cluster M1 was maintained but incorporated a sequence from London.

4.4. Discussion

One thousand HIV *pol* sequences from two anonymous population samples were used: MSM diagnosed in London and in Manchester. The consistency of the phylogenetic clusters formed within each dataset was assessed under varied sample sizes and models of evolution. Overall, the consistency was moderately good with at least 80% of phylogenetic reconstructions of transmission events identified in the initial analysis also found in subsequent analyses. No more than three "novel" transmission events were found for each repeated analysis. Importantly, none of the robust clusters found in the initial analyses were disrupted by the sequences within forming clusters with other sequences in subsequent analyses.

4.4.1 Initial trees

Both datasets were run through ModelTest, which gave different gamma rate heterogeneity values for Manchester (0.35) and London (0.76). This suggests a higher proportion of the sites were identical among sequences in the London dataset, making this sample more homogeneous.

A higher proportion of clusters formed robust clusters in the London dataset compared with the Manchester dataset (23 versus 12). Overall, 10.4% (52/500) of sequences formed a cluster in London compared with 4.8% (24/500) in Manchester ($p<0.001$). This was expected, given the higher degree of homogeneity observed among the sequences from London.

If phylogenetic reconstructions are accurate representations of transmission events, and not formed by chance, it would be expected that relatively few "transmission events" would be added as parameters were weakened. The presence of additional weaker clusters would suggest that the creation of clusters was relative to the sample diversity, rather than reflecting actual transmission events. For London there were an additional five medium clusters and an additional four weak clusters. This is an additional nine compared to 23 robust clusters. For Manchester, an additional eight medium clusters and four weak clusters were added. This may indicate that the formation of clusters is affected by the relative genetic diversity of the sample. The greater occurrence of this phenomenon among the Manchester sample suggests weaker clusters are more likely to form when the population sample is more heterogeneous.

4.4.2 Reducing the sample size

The sample sizes were reduced in order to replicate what the effect of partial representation of an HIV-infected population might be for phylogenetic reconstructions of that population. The initial tree was used as a proxy for the entirety of circulating sequences in the population, and the reductions in sample size undertaken to mimic the effect of sampling within this population.

Overall, for both Manchester and London, there was moderately good consistency between the robust clusters found in the initial trees and among analyses A-D. Excluding the robust cluster affected by the deletions, in each analysis there was at least 80% consistency.

In each analysis, no more than two robust clusters appeared that had not been robust clusters in the initial tree. Approximately half of these were clusters that appeared as medium or weak clusters in the initial tree. The other half represented "novel clusters". It was reassuring to note that in all subsequent analyses, the specific robust clusters not apparent in the initial reconstruction were broadly consistent between analyses A-D.

Whilst the random reduction was used to mimic the effect of random sampling, the non-random reductions (i.e. deleting one sequence from each cluster) were used to assess the potential effect of not including sequences that were part of actual transmission chains. It is encouraging that in all instances, where one sequence had been deleted that had formed a robust cluster in the initial tree, the remaining sequence did not pair with any other sequences in subsequent analyses, even as a weak cluster. This indicates that the likelihood of phylogenetic reconstructions providing contradictory results or “false” matches is low. It also suggests that there are unlikely to be “false matches” as a consequence of not including both sequences from patients included in an actual transmission chain.

The non-clustering sequences were deleted to see if the formation of the “true clusters” were dependent on the presence of other sequences in the sample, or whether the transmission events were absolute. Analysis D indicated that the creation of robust clusters will occur if the sequences that comprise them are present; they are not substantially affected by the presence of other sequences (with which they do not form a cluster).

4.4.3 Models of evolution

For analysis E (changing the models of evolution) the results for both the Manchester and London datasets were broadly consistent both between the models of evolution and through comparing the subsequent analyses to the initial trees. At least 83% of the initial robust clusters were found in each analysis. However the slight discrepancies indicate that the models selected can affect results, and re-emphasizes that it is important to chose the model according to the parameters of the population sample rather than choosing one model for all reconstructions.

4.4.4 Combining the datasets

Analysis F (combining results) produced the least consistent results. Overall, 35 robust clusters were found in the initial London and Manchester trees. Of these, less than half (20) were found as robust clusters in the combined tree. There was a higher degree of consistency between the initial Manchester tree and the combined tree, compared with London. In addition, while “novel” clusters appeared in the combined tree between sequences from London, no novel clusters appeared between the Manchester sequences.

The reason why there was more consistency of clusters in the initial Manchester tree and analysis F may be due to differences between the London and Manchester populations. The gamma shape – a measure of sample heterogeneity - of London was 0.76 and of Manchester was 0.35. Therefore the Manchester sample contained more diverse sequences than the London sample. It may be more difficult to resolve robust clusters in a sample

characterized by greater homogeneity, since there are fewer informative sites to inform clusters.

For the novel clusters, once again, those that appeared were consistent with those found in analyses A-E, indicating that the formation of robust clusters is not a random process. Clusters B1 and B2 are not of concern since they both were formed between a London and a Manchester sequence. It is encouraging that the sequences involved in B1 and B2 did not form any sort of cluster with other sequences in analyses A-E.

4.4.5 Limitations to the approach

This work represents an exploratory attempt to examine the consistency of phylogenetic reconstructions of HIV transmission events for public health purposes. However, there are several limitations with the approach. Firstly, the analysis uses neighbour joining trees, which have been found to produce spurious results in previous analyses (Hillis D 1996). However, in order to have a population sample within which a sufficient number of clusters could be formed (even when a substantial proportion of sequences were deleted and for the sample size to be varied), a bigger sample size was needed. This precluded the construction of a maximum likelihood tree.

Secondly, despite the use of 1000 sequences, only a handful of robust clusters were available for comparisons between analyses (in some cases only two clusters per analysis). This means that the results cannot be meaningfully extrapolated to provide a margin of error for phylogenetic reconstructions.

Thirdly, the datasets from Manchester and London were anonymous. Nothing was known about the population from which the sequences were taken except that they were derived from MSM diagnosed in Manchester and London. Importantly, both London and Manchester are large cities and sexual mixing is not likely to be retained exclusively within each city. This means that there may be unusual aspects of the population samples that affected the results of this chapter that remain unknown. The specific attributes (e.g. the geographic/temporal sampling and sexual behaviour) associated with each population will have impacted upon the likelihood of finding transmission events within each dataset. It is possible that the results of analyses A-F may be affected by the make up of the specific populations. There was no attempt to compare the results with Lewis *et al*. (Lewis, Hughes et al. 2008) for two reasons. Firstly Lewis *et al*. use Bayesian methods, and secondly, the sequences were anonymised, so there was no way of ascertaining which sequences in the chapter were also used in the Lewis paper.

Fourthly, whilst the robust clusters observed in the initial trees are used as the "comparison point" for subsequent analyses, it is impossible to ascertain whether these represent actual transmission events. They were used as a standard against which consistency could be measured and not interpreted as the gold standard. Finally, this analysis was unable to assess another aspect crucial to the use of phylogenetic reconstructions of HIV transmission events; the direction of transmission.

4.5. Conclusion

This chapter represents an attempt to assess the consistency of phylogenetic reconstructions of HIV transmission events for public health purposes. The consistency of robust clusters between analyses was broadly good with at least 80% of robust clusters in each analysis sharing identical robust clusters when compared to the initial trees. However, the small numbers of clusters identified in the population samples means that results can only be exploratory.

Positively, the analysis indicates that the absence of one sequence from two patients involved in a transmission event is unlikely to lead to the remaining sequence erroneously forming a robust cluster with another sequence. Also, the identification of robust clusters does not seem substantially dependent on the presence of other sequences with which they do not share a cluster. Negatively, there are suggestions that phylogenetic reconstructions have insufficient “specificity” to identify robust transmission clusters. This is indicated by the appearance and disappearance of a limited number of “robust clusters” as sample sizes were varied.

Consistency analyses need to be repeated with other datasets using maximum likelihood methods to test whether these results are replicated. Further investigation is needed into the relationship between sample heterogeneity and consistency of robust clusters. Future analyses should work towards providing some sort of “margin of error” that phylogenetic reconstructions may reasonably be expected to produce. In the meantime, the need for caution in interpreting such reconstructions, including those presented in this thesis, remains.

5 Chapter Five: Phylogenetic reconstructions of HIV transmission from patients with recent HIV infection

This chapter describes phylogenetic reconstructions of HIV transmission events from patients with recent HIV infection at diagnosis to: 1) explore the transmission potential from this population and compare findings to those published in the literature; 2) critique the epidemiological definitions and assumptions used in phylogenetic reconstructions.

5.1 Introduction

This chapter explores the application of phylogenetic approaches to the analysis of HIV *pol* sequences from patients diagnosed during recent HIV infection. Phylogenetic analyses are conducted to explore the benefits of this approach, including examining the transmission potential from this group, and to compare the findings to those in the literature. The limitations of phylogenetic reconstructions of HIV transmission events between patients diagnosed with recent infection are also considered (see section 1.6.4), and tighter epidemiological definitions sought.

The first three datasets described in chapter three were used: both the Unlinked Anonymous STI survey (UA survey) and CASCADE are comprised of data drawn entirely from patients with recent HIV infection (at diagnosis or clinic attendance). For the Brighton dataset, only the 159 patients who were recently HIV-infected at diagnosis were included. Two analyses are presented here:

1. A phylogenetic reconstruction of HIV transmission events between patients with recent HIV infection at diagnosis was conducted using the UA survey. This was to explore the transmission potential of the recently HIV-infected population and compare results to those found in the literature;
2. Secondly, the CASCADE and Brighton datasets were used to critique the epidemiological definitions and assumptions used in such phylogenetic reconstructions.

5.2 Methods

5.2.1 Phylogenetic reconstructions of HIV transmissions from populations with recently acquired infection

Using the UA survey dataset (Catchpole, McGarrigle et al. 2000), a phylogenetic analysis was conducted to explore the occurrence of HIV transmission from recently HIV-infected individuals before diagnostic opportunity. The population characteristics of this dataset were described in section 3.3. HIV *pol* sequences were taken from patients found to have a previously undiagnosed HIV infection through UA testing and who were recently HIV-infected through the Serological Testing Algorithm for Recent HIV Seroconversion (STARHS) testing (Janssen, Satten et al. 1998). Patients were categorized as “newly diagnosed” or “undiagnosed” according to whether they received voluntary confidential HIV testing (VCT) during the attendance (patients HIV diagnosed before the clinic attendance were excluded). The newly diagnosed patients received VCT during the clinic attendance, but the undiagnosed did not receive VCT and remained undiagnosed when they left the clinic.

The phylogenetic approach used was described in section 2.2.11. Before phylogenetic analysis, the mutation sites associated with drug resistance (Shafer, Rhee et al. 2007) were deleted from all sequences. Protease and reverse transcriptase sequences were aligned across 998 nucleotides (Sequence Analyzer) and imported into the tree-building software PAUP. The sequence alignment was run through ModelTest and a heuristic search

conducted in PAUP for a maximum likelihood tree using the best fitting model (GTR+I+G) and its derived parameters (proportion of invariable sites=0.43 and gamma distribution=0.79) using the neighbour joining tree as the starting tree. A bootstrap analysis (500 replicates) was used to obtain statistical support for branching patterns. Genetic distances were calculated from the consensus tree for each terminal cluster. The transmission events were interpreted with regard to whether the observed transmission were from patients with undiagnosed or diagnosed HIV infection.

5.2.2 Moving towards more rigorous epidemiological definitions (CASCADE)

Two phylogenetic reconstructions of transmissions from patients who were recently HIV-infected at diagnosis are described in this section. The CASCADE data set comprises an HIV-infected cohort with precise definitions of recent HIV infection, and the Brighton cohort contains a subset of patients who were recently HIV-infected at diagnosis. Both datasets therefore not only enable the identification of transmission events, but also all of the transmission events to be dated crudely. Differences between the infection dates within the observed clusters of patients who were recently HIV-infected at diagnosis were calculated to determine whether transmission could have occurred during recent infection.

CASCADE

The dataset is described in section 3.2. For this analysis, a restricted definition of recent HIV infection was used. Patients were included if they had either an HIV positive test within 90 days of an HIV negative test or an HIV antibody

negative test with polymerase chain reaction positivity. The estimated infection date was taken as midpoint between the diagnosis date and last negative test for the former and the date the sample was taken for the latter.

The first plasma sample available after the estimated HIV infection date was obtained for genotypic resistance testing. Samples obtained more than 18 months after the estimated infection date or from drug experienced patients were excluded. Nucleotide sequence data spanned the protease gene and at least codons 41-236 of reverse transcriptase (RT). The sequencing methodology has been described previously (Masquelier, Bhaskaran et al. 2005). Drug resistance mutations were defined on the basis of the standardised list of mutations for use in epidemiological studies (Shafer, Rhee et al. 2007).

Before phylogenetic analysis, the sites associated with drug resistance mutations (Shafer, Rhee et al. 2007) were deleted from all sequences. Protease and RT sequences were aligned across 998 nucleotides and imported into the tree-building software PAUP. A neighbour joining (NJ tree) was created under the General Time Reversible (GTR) model with gamma rate heterogeneity set at 0.5. The resultant clusters were selected, together with 10 other sequences taken at random. This second sequence alignment was run through ModelTest and a heuristic search conducted in PAUP for a maximum likelihood tree using the best fitting model (GTR+I+G) and its derived parameters (proportion of invariable sites=0.43 and gamma distribution=0.79) using the NJ tree as the starting tree. A bootstrap analysis (500 replicates) was used to obtain statistical support for branching patterns. Genetic distances

were calculated from the consensus tree for each terminal cluster. A transmission “during recent infection” was defined as clusters containing sequences from patients with a difference of fewer than 180 days between estimated infection dates.

Brighton

The dataset was described in section 3.4. For this analysis, only patients who were recently HIV-infected at diagnosis were included. Methods to ascertain recent infection included: p24 antigen; western blot; STARHS; and an interval of six months between last HIV negative test date and date of HIV diagnosis. The calendar quarter of diagnosis was taken as the estimated infection date.

The phylogenetic approach used was described in section 2.2.11. The sequence alignment was run through ModelTest and a heuristic search conducted in PAUP for a maximum likelihood tree using the best fitting model (GTR+I+G) and its derived parameters (proportion of invariable sites=0.2070 and gamma distribution=0.64) using the neighbour joining tree as the starting tree. A bootstrap analysis (500 replicates) was used to obtain statistical support for branching patterns. Genetic distances were calculated from the consensus tree for each terminal cluster. The definition of a transmission generated by an individual during recent HIV infection was taken as a cluster containing sequences that had a difference of up to three calendar quarters between estimated infection dates of the sequences involved. This chapter is descriptive, but statistical tests of association were employed where appropriate.

5.3 Results

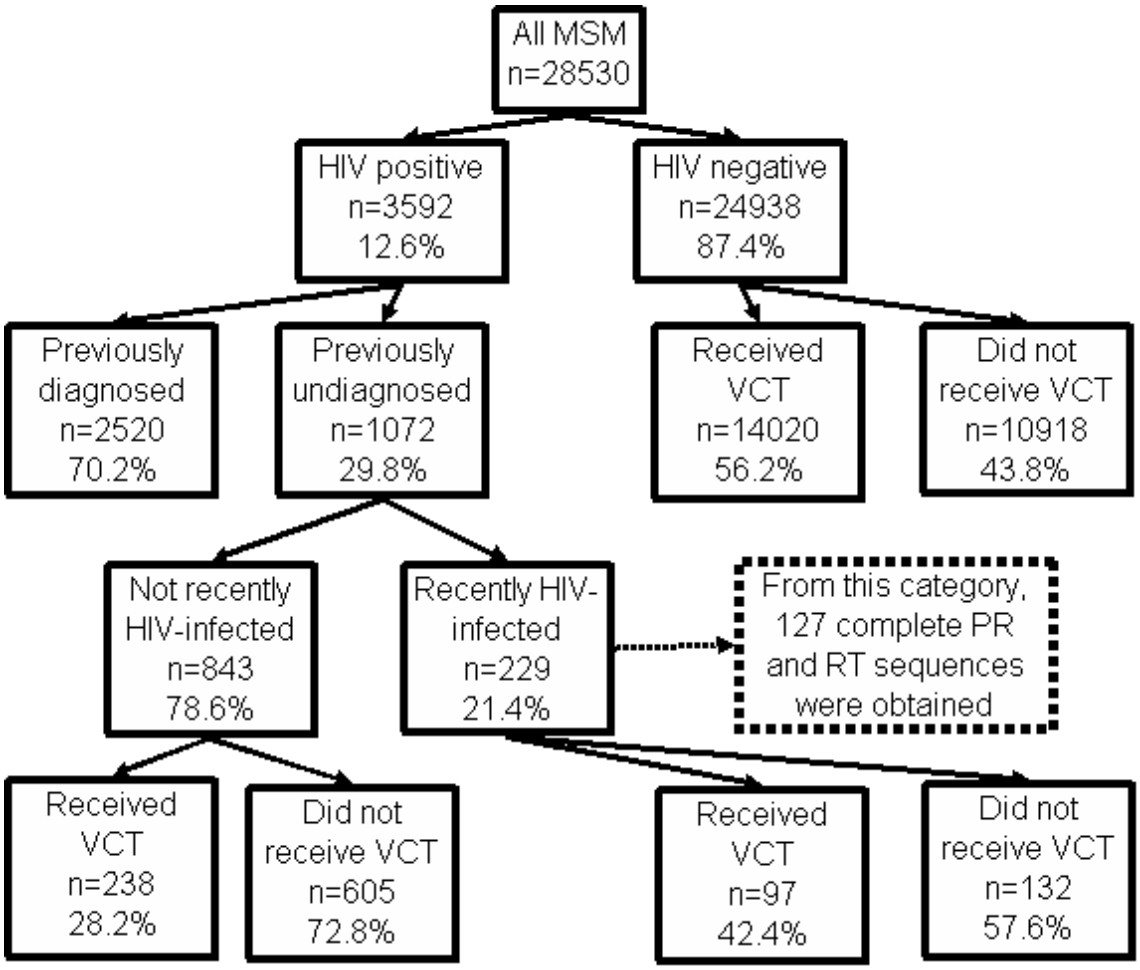
5.3.1 Phylogenetic reconstructions of HIV transmissions from populations with recently acquired infection (UA STI survey)

Between 1999 and 2002, 28530 UA tests were conducted on samples from MSM attending 15 STI clinics in England, Wales and Northern Ireland **(Figure 5.1).** Of these, 3592 samples were found to be HIV positive of which 1072 were derived from patients with a previously undiagnosed infection (newly diagnosed or undiagnosed HIV infection). Among this group, 21% (229) samples were derived from recently-infected MSM. Complete PR and RT sequences were obtained from 127 samples from recently HIV-infected MSM (the remaining 102 could not be amplified due to problems associated with using residual samples).

Overall, of the 229 samples from recently HIV-infected MSM, 86% (196) attended clinics in London, 56% (127) were UK-born and 15% (35) were born elsewhere in Europe. Of the 127 linkable to sequences, equivalent figures were 86% (109), 58% (74) and 15% (19) respectively.

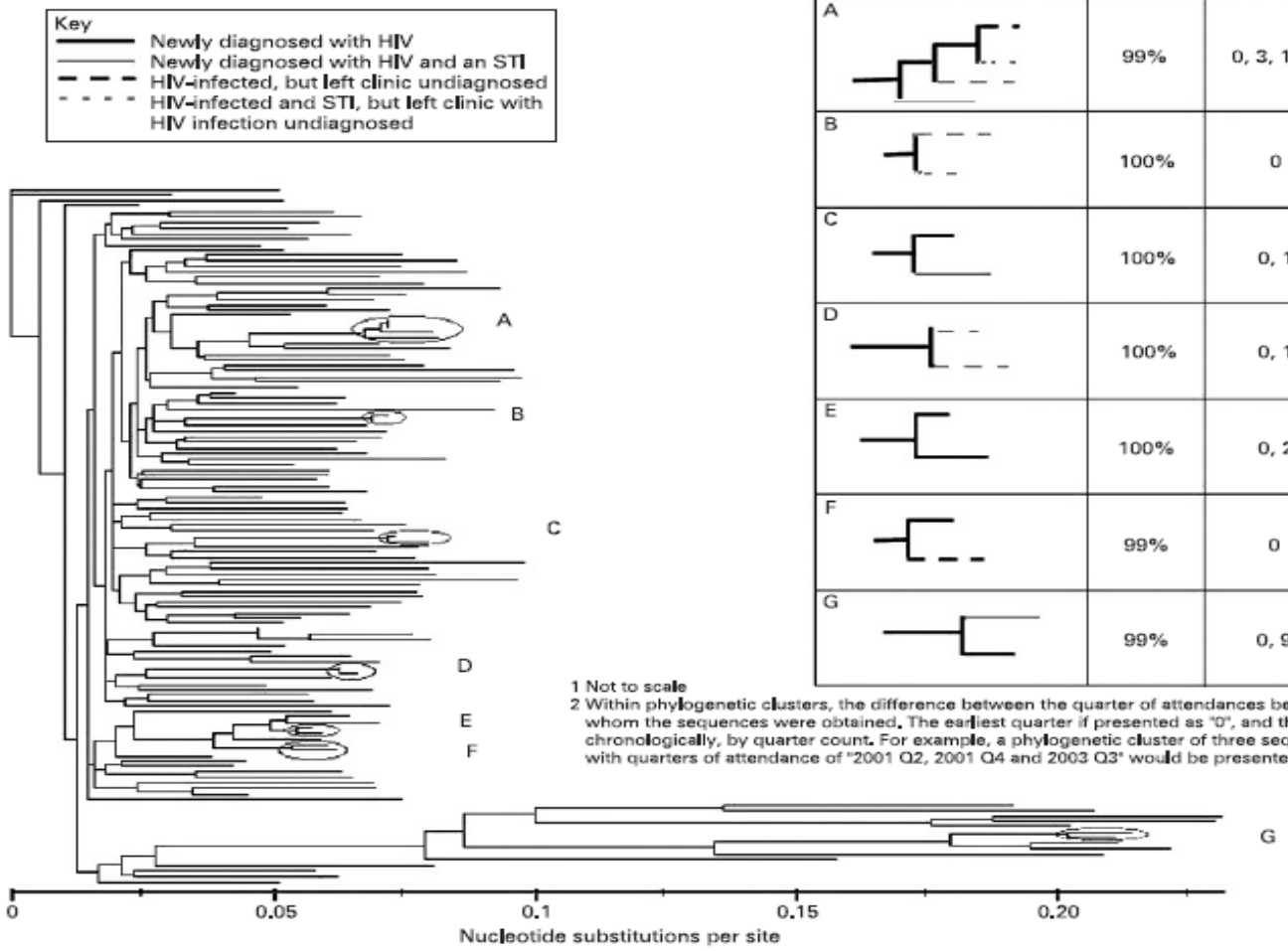
Of the 127 sequences from recently HIV-infected MSM, 16 (12.6%) fell into a cluster with at least one other sequence with a bootstrap support of at least 99% and a genetic distance under 0.015 nucleotide substitutions per site **(Figure 5.2)**. These formed seven clusters (A-G): cluster A contained four sequences, and clusters B-G each comprised sequence pairs. Cluster E comprised sequences from Wales; the remaining six were from London. Clusters C-G were from samples obtained within the same, or successive, calendar quarters.

Figure 5.1: Flow diagram of MSM attending sentinel STI clinics, UA survey: 1999-2002



Clusters B, C and F were comprised entirely from sequences from recently HIV-infected individuals who received voluntary confidential HIV tests (VCT), and consequently had their HIV infection diagnosed at that STI clinic attendance **(Figure 5.2)**. Clusters D and E comprised sequences from individuals who left the clinic remaining unaware of their HIV infection. Of the 16 sequences from MSM that clustered, 56.3% (9) were also diagnosed with an STI. This compares to 34.2% (38/111) from sequences from MSM that did not cluster (p=0.9).

Figure 5.2: Phylogenetic reconstruction transmission events among HIV-infected patients attending sentinel STI clinics during recent infection, UA survey: 1999-2002



Possible transmission events[1]	Bootstrap value	Difference in quarter of attendance[2]	Same world region of birth?	Same age group
A	99%	0, 3, 11, 1	Yes	Of the "four" patients, two share one age group, and two another
B	100%	0	Yes	Yes
C	100%	0, 1	Unknown	No
D	100%	0, 1	No	Yes
E	100%	0, 2	No	Yes
F	99%	0	No	Yes
G	99%	0, 9	No	No

1 Not to scale
2 Within phylogenetic clusters, the difference between the quarter of attendances between patients from whom the sequences were obtained. The earliest quarter if presented as "0", and the next chronologically, by quarter count. For example, a phylogenetic cluster of three sequences from patients with quarters of attendance of "2001 Q2, 2001 Q4 and 2003 Q3" would be presented "0, 2, 9"

5.3.2 Moving towards more rigorous epidemiological definitions

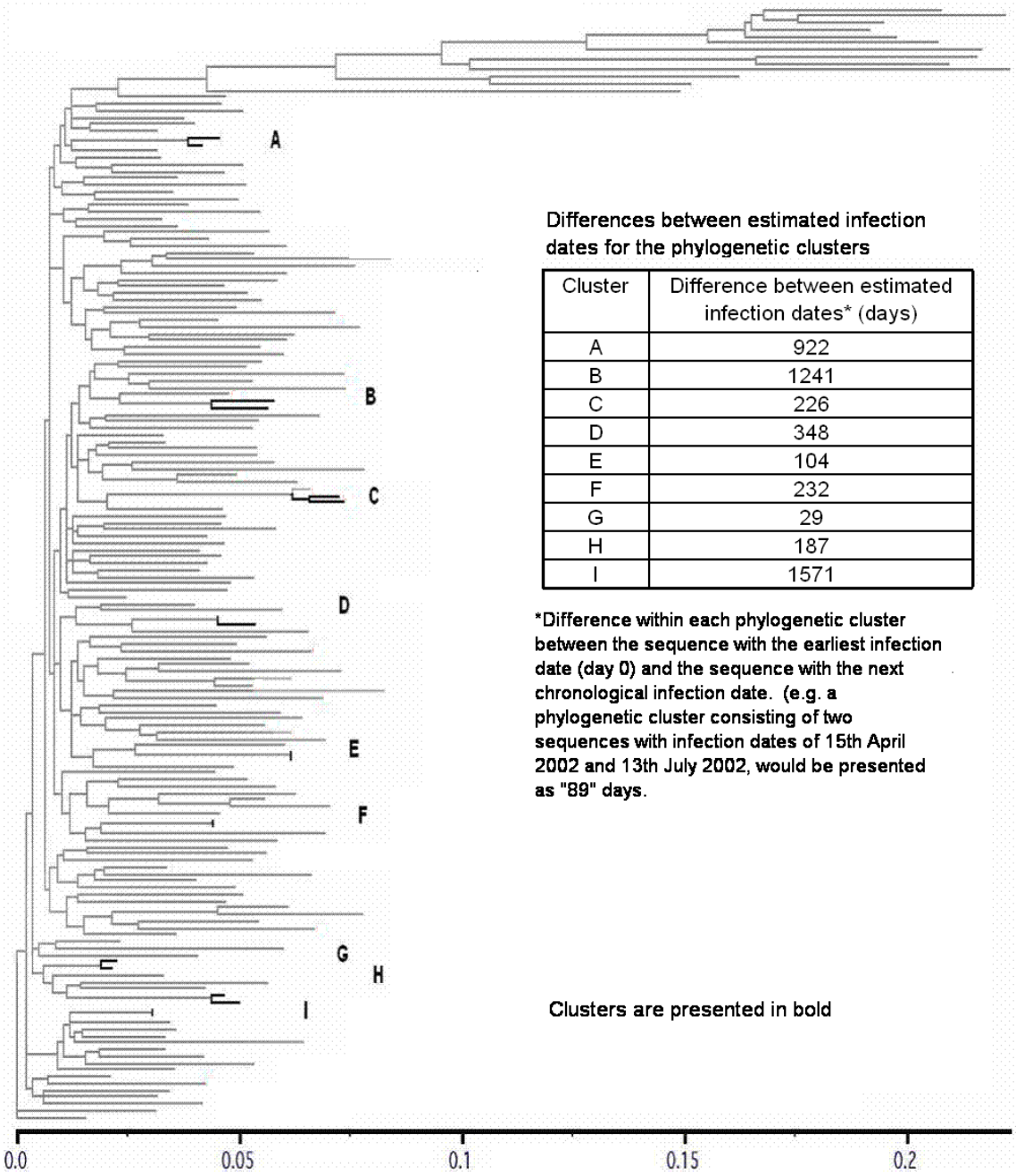
CASCADE

Of the 8993 dated HIV seroconverters from data pooled in December 2004, HIV sequences were submitted from 10 European cohorts. Using this definition, 165 sequences were full length and from drug naïve patients. Specifically, 131 (79%) had had a documented seronegative test within 90 days before their diagnosis (mean number of days between tests was 38 days) **(Figure 5.3**), and 34 (21%) were HIV antibody negative with PCR positivity.

The majority of sequences were from MSM (73%, 120/135). The sample proportion from each country varied annually: 23% (38) of the samples were from Italy, 23% (38) were from France, 20% (33) from Germany, and 17% (28) from the UK, with the remainder from Denmark, Spain and the Netherlands.

Of the 165 sequences, 18 (11%) formed a phylogenetic cluster with at least one other sequence (clusters A-I **Figure 5.3**). These formed nine phylogenetic clusters with two sequences in each cluster. All nine phylogenetic clusters were comprised of sequences derived from patients who had their samples obtained within the same country. Two of the nine clusters contained sequences that had a difference between infection dates of 104 and 29 days (clusters E and G respectively). For the remaining seven clusters, the average difference of infection dates within clusters was 675 days (187-1571 days).

Figure 5.3: Phylogenetic reconstruction of transmission events among European HIV-infected patients with recent HIV infection at diagnosis, CASCADE: 1989-2004



Differences between estimated infection dates for the phylogenetic clusters

Cluster	Difference between estimated infection dates* (days)
A	922
B	1241
C	226
D	348
E	104
F	232
G	29
H	187
I	1571

*Difference within each phylogenetic cluster between the sequence with the earliest infection date (day 0) and the sequence with the next chronological infection date. (e.g. a phylogenetic cluster consisting of two sequences with infection dates of 15th April 2002 and 13th July 2002, would be presented as "89" days.

Clusters are presented in bold

Brighton

Of the 859 HIV-infected patients with sequences available, 159 were derived from MSM who were recently HIV-infected at diagnosis. Of these 21 were

identified through p24 antigen; three were recognised through Western blot; 119 were identified through STARHS; and 16 had an interval of less than 183 days between last HIV negative test date and date of HIV diagnosis.

Where demographic data were available, patients with recent HIV infection at diagnosis were disproportionately white (98.1%, 152/155), born in the UK (83.0%, 127/153) or elsewhere in Europe (9.8%, 15/153). There was a fairly even distribution by age-group with approximately half under 35 years of age at diagnosis (48.4% 77/159). Around one fifth had an STI diagnosed at the same time as their HIV infection was diagnosed (21.0%, 34/159).

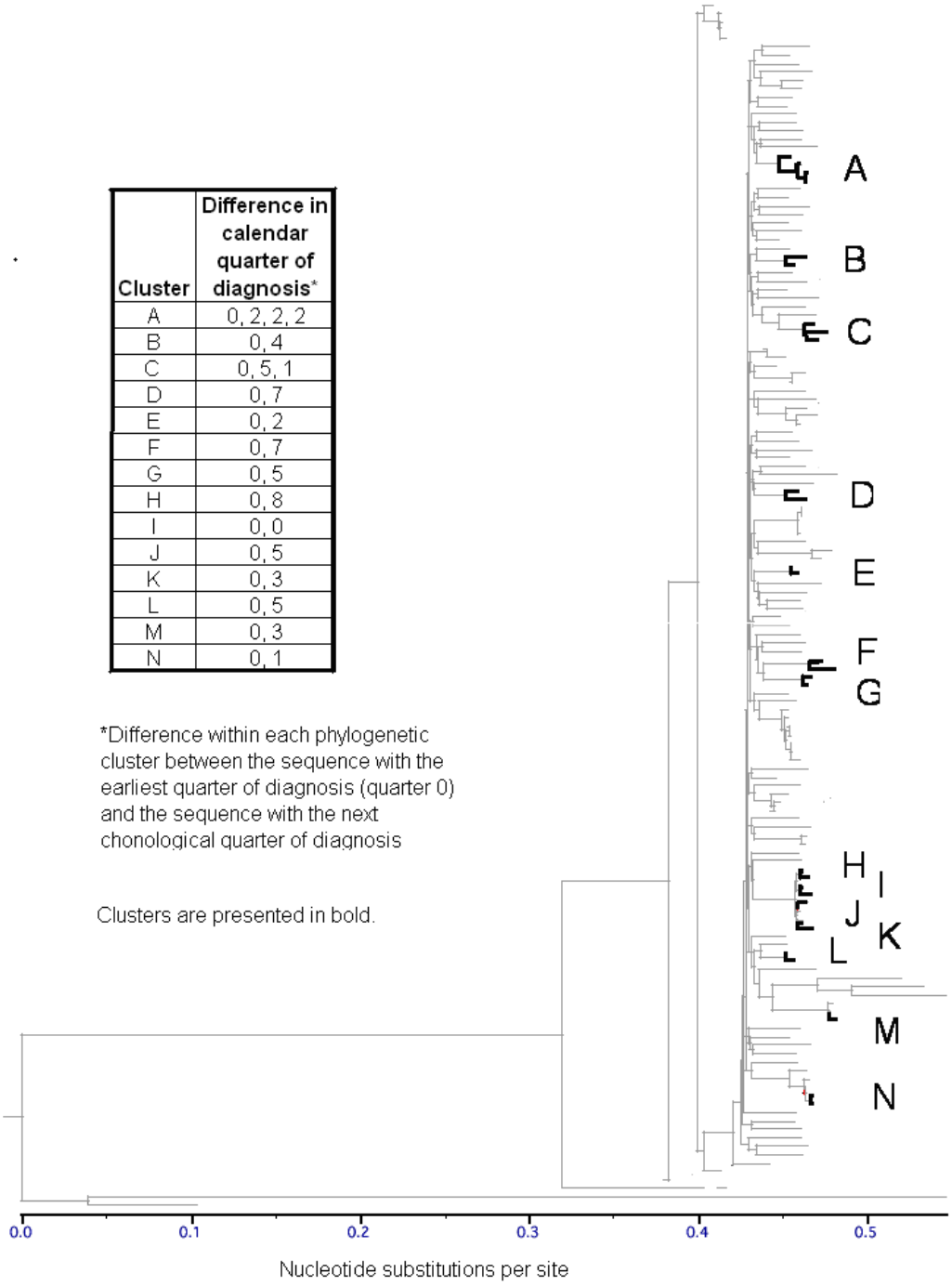
Of the 159 sequences from MSM, 31 (19.5%) sequences formed a phylogenetic cluster with at least one other sequence (Clusters A-N, **Figure 5.4**). These formed 14 clusters, with an average of 2.4 sequences per cluster (range 2-4). Of the 14 clusters, only five (A, C, E, I and N) contained at least two sequences that had a difference of under three calendar quarters of diagnosis. For the remaining nine clusters, the average number of calendar quarters between sequences was 7.8 (range 4-8 calendar quarters).

Figure 5.4: Phylogenetic reconstruction of transmission events among HIV-infected MSM with recent infection at diagnosis, Brighton: 2000-2006

Cluster	Difference in calendar quarter of diagnosis*
A	0, 2, 2, 2
B	0, 4
C	0, 5, 1
D	0, 7
E	0, 2
F	0, 7
G	0, 5
H	0, 8
I	0, 0
J	0, 5
K	0, 3
L	0, 5
M	0, 3
N	0, 1

*Difference within each phylogenetic cluster between the sequence with the earliest quarter of diagnosis (quarter 0) and the sequence with the next chonological quarter of diagnosis

Clusters are presented in bold.



5.4 Discussion

5.4.1 Phylogenetic reconstructions of HIV transmissions from populations with recently acquired infection

Phylogenetic analyses of HIV *pol* sequences from recently HIV-infected MSM attending sentinel STI clinics identified at least seven possible instances of HIV transmission before the first routine opportunity for diagnosis. Five clusters had closeness in the likely infection dates between the sequences involved. This demonstrates the potential for HIV transmission to occur rapidly from the recently HIV-infected, even among those who attend the clinic soon after infection (Brown, Murphy et al. 2009b).

The rate of clustering observed (12.6%) is lower than that observed in other studies of patients diagnosed during recent HIV infection: 34% (Pao, Fisher et al. 2005) and 50% (Brenner, Roger et al. 2007). This may be because Pao's and Brenner's populations were local whereas data from the UA survey was taken from 15 sites across England, Wales and Northern Ireland. Pao also found an association between clustering and STI diagnosis at HIV diagnosis. While data from Brighton presented in this chapter may have also indicated a similar association, this was not statistically significant at the 5% level (p=0.08).

The UA survey methodology allows patients to be sampled (within the same clinic) up to four times annually, but not within the same calendar quarter. Therefore, theoretically, the transmission clusters identified could have been

formed between the same individual, sampled multiple times. Furthermore, the UA survey is necessarily designed to prevent the identification of individuals. However, five of the possible transmissions (C-G) involved patient pairs who had a different birth regions or age-group **(Figure 5.2**) making it unlikely that the sequences came from the same person. Similarly, there must have been at least two individuals included in cluster A – but not necessarily four. The patients in cluster B both attended the same clinic within the same calendar quarter, meaning they should be different patients according to the UA algorithm.

Whilst the differences in calendar quarter of diagnosis between recently HIV-infected patients were relatively tight, the factors potentially associated with transmission (e.g. STIs) cannot be definitively associated with transmission events at around the time of transmission. For the UA survey, the definition used to identify recent infection (STARHS) and the time interval used to date diagnosis (calendar quarter) are too broad to allow an accurate comparison of infection dates. Therefore it can not be ascertained whether the transmissions identified between patients diagnosed with recent HIV infection actually occurred during recent infection.

The phylogenetic analysis illustrates the potential for HIV-infected patients who were recently HIV-infected at diagnosis to rapidly generate new infections. Even within a small dataset taken from 15 regions of England, Wales and Northern Ireland, three transmissions (B, C and F) were observed among those

that received VCT soon after their infection. This suggests that VCT alone may not have a sufficient impact on the transmission potential of recently HIV-infected MSM. Alternative strategies include: more rigorous partner notification; post exposure prophylaxis among recently–exposed MSM; the encouragement of more frequent regular testing; and the education of MSM and health care providers about the transmission risk and symptoms of seroconversion illness.

5.4.2 Moving towards more rigorous epidemiological definitions

Two phylogenetic reconstructions were described using populations of MSM diagnosed during recent infection. Both datasets contained MSM with well estimated infection dates. For each cluster observed, the difference between the estimated infection dates was calculated in order to ascertain whether the possible transmission event could have been generated during recent infection. For CASCADE, nine transmissions may have occurred between individuals in this sample. Of the possible transmission events observed, only two could have been generated during recent HIV infection (Brown 2009a). For Brighton, 14 possible transmissions may have occurred. Of these, only five could have been generated during recent infection.

This analysis demonstrates the need for caution in the design and interpretation of phylogenetic analyses that link HIV *pol* sequences to data relating to the patient's infection stage. In addition to the identification of transmission events between patients recently HIV-infected at diagnosis, the time intervals between

infection dates between patients involved in a possible transmissions event have been explored. This comparison revealed only a small proportion of possible transmissions could have been generated during recent infection. Given the time intervals between the infection dates, the remainder are more likely to have been transmitted from individuals during chronic infection. Previous analyses that use more liberal definitions of recent HIV infection and do not take the transient nature of this infection stage into account may be overestimating the extent that the recently HIV-infected population is generating new transmission events.

A criticism of phylogenetic reconstructions between patients with recent infection is that the datasets do not represent patients with chronic infection. However, this work demonstrates that sequences involved in phylogenetic reconstructions should not be permanently categorized as coming from the recently and chronically HIV-infected according to the patients' infection stage at diagnosis. This is because just as phylogenetic reconstructions of possible transmission events between patients diagnosed as recently HIV-infected do not necessarily equate to recent to recent transmissions, transmission events identified between patients diagnosed with chronic infection do not necessarily equate to transmissions during chronic infection. Without information on the approximate date of the infection for the majority of sequences included in the analysis, and without dating the transmission events in some way, the infection stage of the transmitter, at the time of transmission, cannot be ascertained. Instead, provided that the study population has been well selected in terms of

coverage of a local HIV-infected population, and contains a substantial sample of sequences with well estimated infection dates, both the recently and chronically infected will be in represented in the dataset.

It is recommended that future analyses should comprise sequences from patients from a reasonably complete and closed HIV-infected population and contain a large proportion of patients with well-estimated infection dates. Methods should then allow each patient's infection stage to change from recent to chronic infection to reflect the natural progression of HIV infection. Secondly, it is suggested that transmission events involving patients with recent infection could be used to approximate the "transmission date" i.e. transmissions involving a recently HIV-infected individual are likely to have occurred around that patient's diagnosis. This "dating" will allow researchers to ascertain whether the patient most likely to have generated the recent infection was themselves recently or chronically infected at the transmission date (provided they also have an estimated infection date, or a diagnosis date of at least six months earlier).

This analysis is not an attempt to measure the extent that the recently HIV-infected are generating new transmissions. The data were selected because they contained well estimated infection dates. This allowed the temporal relationship between sequences involved in a possible transmission event to be explored. However, the size and inconsistent geographic and demographic population make up of the CASCADE prevents the wider extrapolation of the

results and the calculation of the rate of transmission from the recently HIV-infected population. While the Brighton dataset is more representative of a local HIV population, the limitations of using exclusive recently HIV-infected cohorts limits the interpretation of this analysis.

There are further problems associated with datasets that are exclusively comprised from patients who were recently HIV-infected at diagnosis. While it may be possible to account for the transient nature of recent infection through calculating infection dates for each patient and allowing them to progress from recent to chronic infection, there may be an intrinsic difference between patients who were diagnosed during recent infection and those diagnosed during chronic infection in terms of the HIV testing and sexual behaviours. Those who present for HIV testing shortly after a recent risk exposure may have different levels of risk behaviour. Patients who are diagnosed during recent infection may be more likely to have symptoms of seroconversion illness.

Patients only remain recently HIV-infected for a brief period of time compared to the overall length of their infection. Consequently, studies that focus upon transmission from the recently HIV-infected population ignore individuals living with chronic HIV infection; the majority of the HIV-infected population. Within the chronically HIV-infected population, transmission risk will vary with viral load/treatment status, and STI acquisition. Indeed, plasma viral load has been demonstrated to be the most important risk factor in increasing the likelihood of transmission. It is essential that future phylogenetic reconstructions

incorporate other variables of importance, and also essential that variables that constitute transient values (e.g. viral load, STI presence, CD4 counts) are treated as such. For instance, they will only be informative for phylogenetic reconstructions provided that transmission, and the presence of the values, can be dated in some way.

5.6 Conclusion

This chapter has demonstrated the utilities of phylogenetic reconstructions of HIV transmission events between patients who were diagnosed with recent HIV infection. It has identified that this population may have an important role in generating new transmissions. It also suggested that such approaches are likely to overestimate the extent this population generates new infections, unless precise definitions of recent infection are used, and some technique of "dating" transmission events is employed.

Work undertaken in this chapter did not attempt to determine the extent that those with recent HIV infection are driving transmission, but suggested that such attempts require three conditions. Firstly, the sample population needs to include patients who were recently, and chronically HIV-infected at diagnosis (and be broadly representative of an HIV-infected population with respect to geography and time). Secondly, the sample population must include a substantial proportion of patients with well estimated infection dates and recognize that such patients are only recently HIV-infected for a short duration.

Thirdly, it requires a mechanism that allows transmission events to be dated, so that the infection stage of the transmitter can be ascertained, at around the time of transmission. Only then can transmissions from recently HIV-infected patients be measured with any precision. The chapter also suggests analyses need not and should not be restricted investigating transmissions from the recently HIV-infected; the approach can be extended to include other relevant risk factors for HIV transmission.

6 Chapter Six: The source of new HIV infections in a local HIV-infected population

In this chapter, a local UK population of diagnosed HIV-infected UK MSM is described, with particular emphasis on its transmission potential. A phylogenetic reconstruction of HIV transmission events is conducted. An approach that links transmission events with the specific factors associated with the transmitters, at around the time of transmission, is used to identify the factors that are important in generating new transmissions.

6.1 Introduction

To date, the majority of phylogenetic reconstructions of HIV transmission have focussed on the recently HIV-infected population (Pao, Fisher et al. 2005), (Yerly, Kaiser et al. 1999; Brenner, Roger et al. 2007). The chronically HIV-infected population is usually either absent or the infection stage poorly defined (Brenner, Roger et al. 2007) (see section 1.5.1). Generally, patients have been categorised as being either recently or chronically HIV-infected according to their infection stage at diagnosis, with a failure to recognise the transient nature of infection stage (see section 1.6.4). This failure may have overestimated the transmissions events generated by the recently HIV-infected population (chapter five), (Brown 2009a).

This chapter aims to describe the HIV-infected population in Brighton with particular emphasis on its transmission potential (e.g. plasma viral load, treatment and STI diagnoses). A phylogenetic reconstruction of HIV transmission events was undertaken among this population. An approach was used that can link "dated" transmission events to specific risk factors associated with the transmitters at the time of transmission. Finally, the groups that are important in generating HIV transmission events were ascertained.

6.2 Methods

6.2.1 The categorization of patients by risk factor

This section describes the diagnosed HIV-infected population of Brighton with reference to its transmission potential. The demographic and clinical details of the Brighton dataset are described in section 3.4. In brief, during the study period (2000-2006, comprising 28 calendar quarters), HIV *pol* sequences were obtained from HIV-infected men who have sex with men (MSM) attending Brighton HIV clinic. Sequences were linked to the following data **(Figure 3.2**):

- recent/chronic infection;
- treatment (untreated/treated/treatment interruption);
- plasma viral load;
- CD4 count;
- STI diagnosis;
- age-group;
- ethnicity;
- world region of birth.

The study period was stratified into a series of three month calendar quarters and the characteristics of patients under follow-up were summarised at the start of each quarter. For each patient, data were collected for the first calendar quarter of attendance during the study period (this was the quarter of diagnosis if diagnosis occurred during the study period), and updated for each consecutive calendar quarter throughout the study period. This enabled patients' variables (where relevant) to change over time. It was not possible to

know whether patients attended during every calendar quarter during the study period. Where patients did not attend during a calendar quarter, the data from the previous quarter were carried forward. Patients who did not attend for 12 consecutive months were considered lost to follow-up and excluded from the study period from their last quarter of attendance.

For each patient ascertained as recently HIV-infected at diagnosis, an earliest and latest infection date was calculated according to which marker was used to identify the recent infection **(Table 6.1).** An estimated infection date was taken as the mid point between the earliest and the latest infection date for each patient.

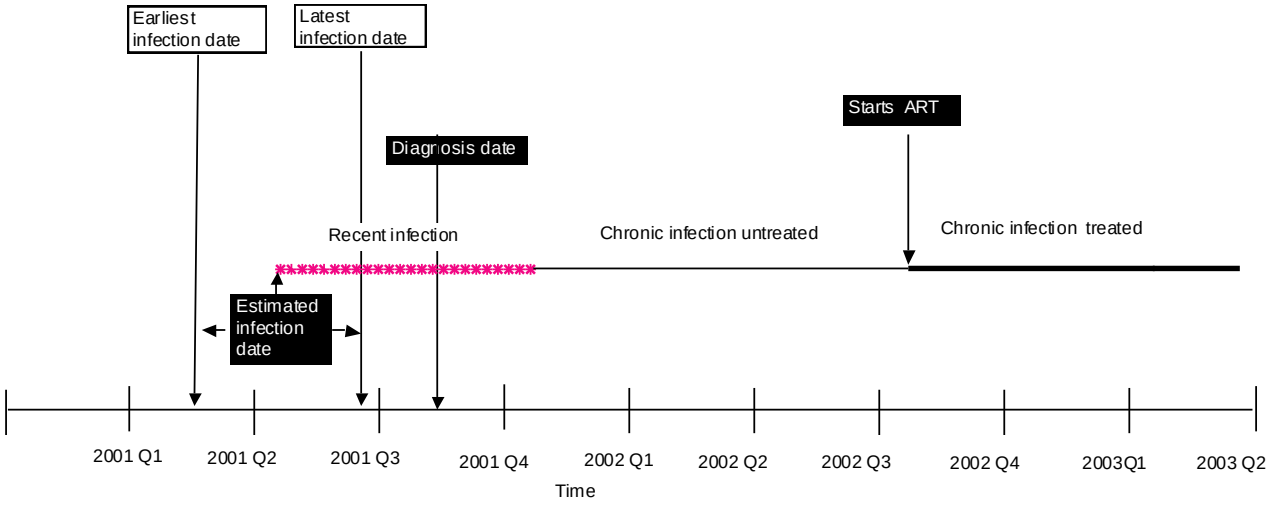
Table 6.1: Estimated infection dates for patients identified as recently HIV-infected at diagnosis: Brighton

Algorithm	Earliest infection date	Latest infection date
Evolving antibody response/p24 antigen	Diagnosis date -30 days	Date diagnosis -1 day
Western blot	Diagnosis date -60 days	Date diagnosis -30 days
STARHS	Diagnosis date -183 days	Date diagnosis -60 days
HIV negative <6 months before diagnosis	Date last negative +1 day	Date diagnosis -30 day

To ensure consistency, each patient was categorized as recently HIV-infected for exactly six months following their estimated infection date. However in order to mask the identity of patients, it was necessary to present data by calendar quarters. While each patient remained recently HIV-infected for exactly six months, the number of calendar quarters during which they remained recently

HIV-infected depended on where their estimated infection date fell in relation to the calendar quarter. For instance, a patient whose infection date was at the end of March, would be categorised as recently HIV-infected for three calendar quarters, whereas a patient infected at the beginning of April would remain recently HIV-infected for only two quarters **(Figure 6.1).**

Figure 6.1: Flow diagram showing how infection category was ascertained and updated over time for each patient: Brighton



Combining the estimated infection date information with the clinical data, patients were subdivided into infection category as follows:

- recently HIV-infected
- chronically HIV-infected and untreated with ARV
- chronically HIV-infected and treated with ARV
- chronically HIV-infected currently interrupting ARV treatment

Each patient had their infection category updated for each subsequent calendar quarter following diagnosis (or their first attendance in the study period if they

were diagnosed before the study period), according to the data available (**Figure 6.1**).

6.2.2 Describing the HIV-infected population and its transmission potential

Due to the difficulties of describing a diagnosed HIV-infected population in relation to its distribution by infection category (see section 2.3.5) and risk factors for transmission (see section 2.4.7), three approaches were developed. All three approaches included every patient who attended during the study period, but the calendar quarter of the study period selected varied between the different groups. These are outlined in **Table 6.2.**

6.2.3 Phylogenetic reconstruction of HIV transmission events in Brighton

A phylogenetic reconstruction of transmission events among all MSM in Brighton is described in this section. It compares the risk factors from patients whose sequences formed a cluster against those whose sequences did not form a cluster.

Phylogenetic methods

Amino acid positions associated with ARV resistance mutations (Shafer, Rhee et al. 2007) were deleted prior to phylogenetic analysis. The sequences were aligned across 998 nucleotides using Sequence Analyzer. A neighbour joining tree was constructed with gamma rate heterogeneity set at 0.5 and 500 bootstrap replicates.

Table 6.2: Description and attributes of the three approaches used to analyse the Brighton HIV-infected population

Approach	Description	Method	Attributes	Representing?
1	Only included the patients' first attendance within the study period.	The patients' calendar quarter of diagnosis was selected for those diagnosed during the study period. For those diagnosed before the study period, the first quarter of the study period was selected.	859 patients with one calendar quarter included.	Previous phylogenetic analyses that categorise according to data obtained at diagnosis
2	Only included data from one calendar quarter for each patient for the entire study period, but the calendar quarter was selected at random.	Patients were allocated an anonymous number which was sorted in ascending order. Every calendar quarter of the study period was presented chronologically (i.e. quarter 1, 2, 3) for each patient. For the first patient, the first quarter was selected, for the second quarter was selected etc. Once the first 28 patients had a calendar quarter selected (there were 28 calendar quarters during the study period) the process started again with patient 29 having the first calendar quarter selected. Where patients did not have the relevant calendar quarter (for instance if information was only available for quarters 1-4 for a patient, but quarter 16 was due to be selected), the patient was skipped. The process was repeated until all 859 patients had a calendar quarter selected.	859 patients with one calendar quarter included.	A "snap-shot" of the HIV-infected population attending Brighton clinic between 2000-6.
3	Included every patient, for every calendar quarter (including and) following diagnoses	Each patient could be included up to 28 times within this dataset. Patients were not included if they were lost to follow-up (no attendance for 12 months). It is not know whether patients attended every quarter. Where data were unavailable, data from the previous quarter were carried forward.	Number of calendar quarters of each patient following diagnosis or the first study period, until end of study period, or lost to follow-up.	A representation of the cumulative HIV-infected population of Brighton between 2000-6.

The large sample size precluded maximum likelihood methods (see section 2.2.11). Consequently, sequences that fell into a robust cluster were selected, along with a random selection of 50 sequences that did not form any sort of cluster. Together, these sequences underwent a second neighbour joining tree reconstruction with gamma rate heterogeneity set at 0.5. The alignment was then run through ModelTest. A maximum likelihood tree was constructed using the selected model (K81uf+I+G) and its estimated parameters (the proportion of invariable sites set as 0.4477 and the heterogeneity set at 0.6946). The same cluster definitions were used; only those clusters that were identical between the neighbour joining and maximum likelihood trees were retained.

Factors associated with HIV transmission

The number of possible transmission events were calculated and described. The attributes associated with patients involved in a possible transmission event were ascertained: infection category; plasma viral load; STI diagnosis; WRB; and age-group. For each patient, the variables selected were those allocated at the patient's first attendance.

6.2.4 A revised method: where do new infections come from?

Unlike previous reconstructions (see section 1.5.1), this method adopted a strategy that: a) identified the most likely sources of transmission of patients diagnosed with a recent HIV infection; b) used the date of the recent infection to approximate the date of transmission; c) ascertained the attributes (e.g. infection category, plasma viral load, treatment status) of the most likely transmission source, at around the time of transmission.

Identifying the transmitters with estimated transmission dates

The phylogenetic reconstruction described in section 6.2.3 was used for this analysis. However, in this instance, only the transmission sources of patients diagnosed during recent HIV infection were sought. This was in order to date the transmission events approximately; transmissions to patients with recent infections, by definition, must have occurred shortly before the diagnosis. Such transmitters are referred to as "transmission sources with estimated transmission dates" throughout the remainder of this thesis. The method is graphically summarized in **Figure 6.2.**

All patients within the dataset, regardless of infection category were considered potential transmission sources and a window of transmission was considered from their earliest appearance in the study period to the end of the study period (or when they were lost to follow-up). In order to be considered as a transmission source with an estimated transmission date, candidate sequences had to form a robust cluster with a patient who was diagnosed during recent infection. Where there was more than one candidate, the candidate with the shortest genetic distance was selected. Where genetic distances between two candidates were identical, the candidate with the lowest number of nucleotide differences to the sequence from the recently HIV-infected patient was selected. Where this was equal, transitions were preferred over transversions. A distinction was made between where the clusters formed between the recent HIV-infection and its transmission source was unambiguous (the only sequence to form a cluster with the sequence from the recently HIV-infected patient) and

transmitters that formed a cluster with the sequence from the recently HIV-infected sequence, along with other sequences.

Candidate transmission sources with estimated transmission dates were identified for as many patients who were recently HIV-infected at diagnosis as possible. Candidate transmitters were excluded if they had been diagnosed during chronic infection and their quarter of diagnosis occurred *after* diagnosis quarter of the recently HIV-infection patient, with which it was linked. This is because, in these instances, the direction of transmission could not be ascertained (e.g. it could not be established whether the recent infection generated the chronic infection or vice versa).

Where a transmission event was identified between two patients both diagnosed as recently HIV-infected during the same calendar quarter, the patient identified as recently HIV-infected using the marker with the longest window period was selected as the transmission source. The recent infection with the shortest window period was excluded as a transmission source. Where the same marker of recent infection had been used for both patients, the pair was excluded from analysis since it was impossible to determine the direction of transmission between the two. Candidate transmitters were only included in the analysis provided they were also present in the dataset during the calendar quarter in which the recently HIV-infected patient of interest was diagnosed (i.e. the period of transmission). Transmission sources that "transmitted" after they were lost to follow up were excluded.

Once the subset of transmission sources with estimated transmission dates were identified, the calendar quarter of transmission was taken to be the calendar quarter of diagnosis of the recent HIV-infection (which the candidate transmission source had generated). Clinical data (infection category, plasma viral load, CD4 count and STI diagnosis) was then collated for the transmission source, during the calendar quarter of transmission. In this way, the risk factor data were obtained at the approximate time of the transmission event.

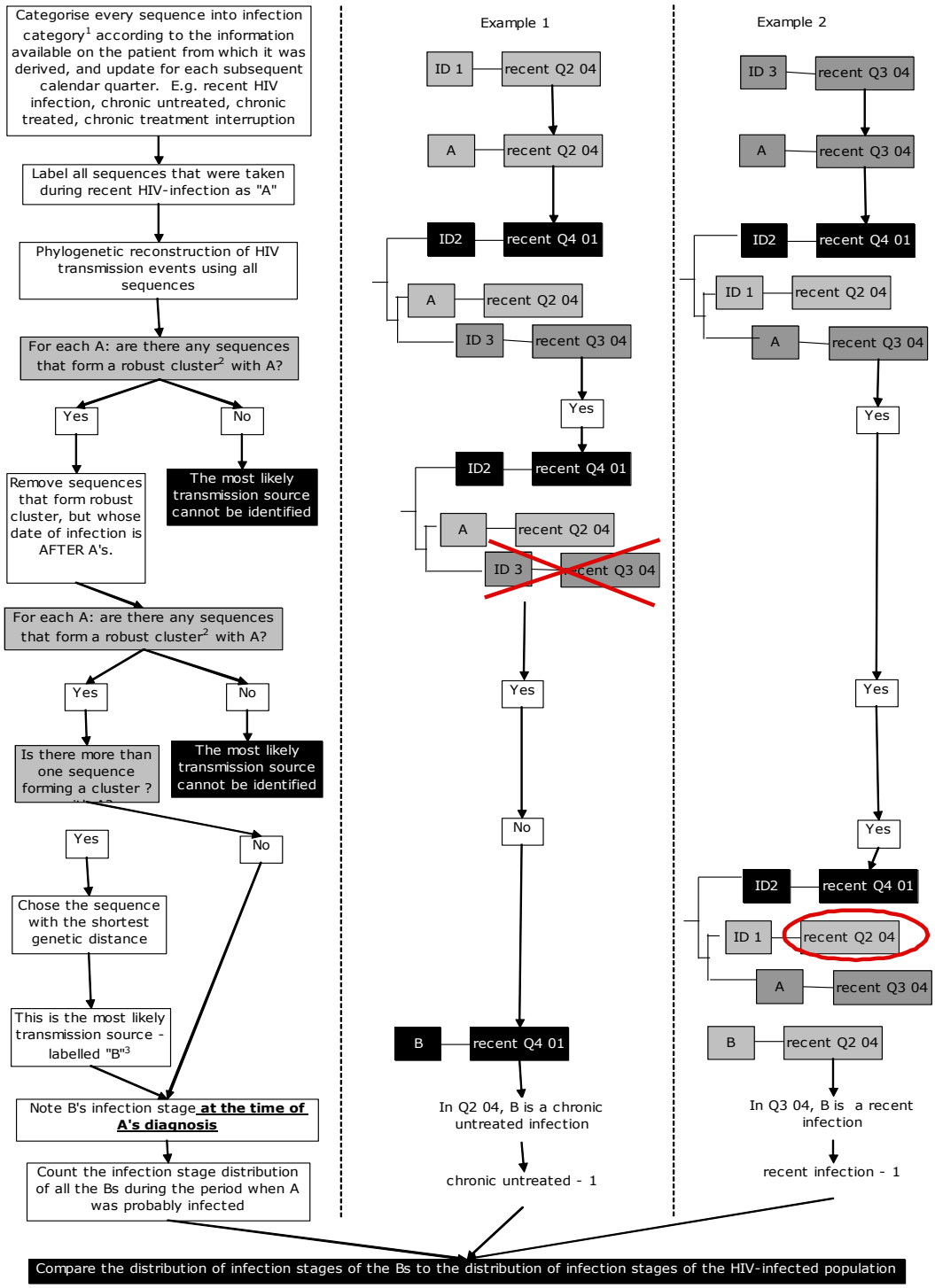
Descriptive analysis

Once the subset of sources with estimated transmission dates were ascertained and collated, the distribution of this groups' infection categories, plasma viral load, treatment status and STI diagnoses were described. This was compared to the distribution of the same variables of the concurrent Brighton population. Statistical tests of association were employed where appropriate.

Statistical analyses

The rate of transmission was calculated for each risk factor, per 100 person years of follow-up (PYFU). The PYFU were calculated through summing every calendar quarter of the study period for each patient from their first attendance to the end of the study period (or the patient's last attendance, where the patients were lost to follow-up).

Figure 6.2: Flow diagram of method used to identify transmission source of patients diagnosed during recent HIV infection: Brighton



1 Infection stage is updated for each calendar quarter. E.g. an individuals with recent HIV infection in quarter one 2003 will be a chronic infection in quarter three 2003.
2 Robust cluster: >99% bootstrap and genetic distance <0.015 nucleotide substitutions per site.
3 A's may form robust clusters with other As. Each A is considered individually, for instance, one A may have another A considered its most likely transmission source. This second A, in turn, is examined separately to find its most likely transmission source.
4 E.g. B may have been a recent infection at the time B's sample was sequenced - however, B may have "transmitted", generating A after recent HIV infection. There the infection stage of B, at the time of As infection, is taken.

Factors independently associated with transmission were identified using univariable and multivariable Poisson regression models (SAS version 9.1). Every calendar quarter of the study period was included separately in analyses with a covariate to indicate whether or not transmission occurred during that period. Sensitivity analyses were conducted to assess the impact of the assumptions used to allocate infection categories, and the impact of missing data. This involved repeating the multivariable analyses, whilst adjusting relevant variables (e.g. altering the estimated infection date to the earliest, and then the latest infection date).

Case studies of transmitters

For each of the transmission sources with estimated transmission dates, a graphical representation is provided for the time period they were present within the dataset. The possible transmission event was illustrated within the wider context of their evolving viral loads, STI diagnoses and infection categories.

6.3 Results

6.3.1 Transmission potential of the HIV-infected population

Population size and characteristics

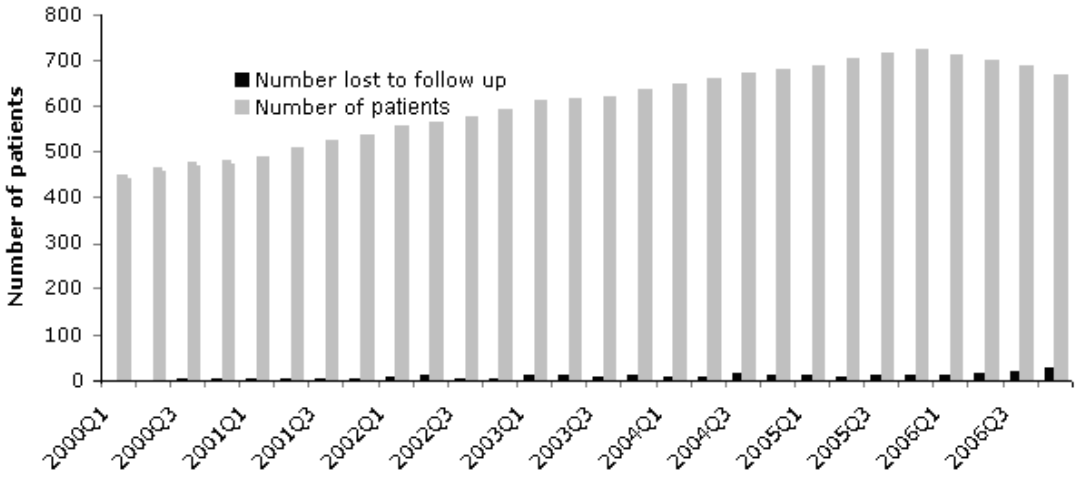
Of 1144 HIV-infected patients attending the Brighton HIV clinic during 2000-2006, 75.1% (859) patients were linkable to complete protease (PR) and reverse transcriptase (RT) sequences. Only these patients are considered in this analysis. The description of the demographic data and the difference between the patients with and without sequences is provided in section 3.6.5.

Data completeness

Of the 859 patients included in the phylogenetic analysis, 428 were diagnosed before the year 2000 and 431 were diagnosed during the study period (2000-6). The total number of patients, with sequences available, represented in each calendar quarter of the study period is presented in **Figure 6.3**. The number of patients within the dataset rose annually until the first quarter of 2006 when numbers began to decline. Overall, there were 6176 patient years of follow-up. Between 2000-2006, the average number of patients represented was 606.7 (range 451-724) for each calendar quarter. On average, patients had data available for 21.2 consecutive calendar quarters (range: 3-28, standard deviation 8.8 calendar quarters) and eight patients were lost to follow-up every calendar quarter.

The proportion of patients with missing demographic data was calculated for each field for each patient (world region of birth, ethnicity, and age-group) and also for each calendar quarter the study period for clinical data (plasma viral load, CD4 count and STI diagnosis). The proportion of patients with missing demographic data was 10.1% (87/859) for world region of birth, 2.6% (22/859) for ethnicity and 0% (0/859) for age-group. The proportion of fields with missing data was 9.9% (1685/16988) for plasma viral load and 9.7% (1641/16988) for CD4 counts. The records that had missing values for viral load were more likely to have missing CD4 values.

Figure 6.3: Number of diagnosed HIV-infected patients with complete *pol* sequences represented, and number lost to follow up, by calendar quarter, Brighton: 2000-2006



Infection category distribution

Of the 859 patients, 159 (18.5%) were found to be recently HIV-infected at diagnosis. Of these, 104 were ascertained through STARHS, 20 through the p24 antigen, two from western blot and 21 through the dates algorithm (**Table 6.1**). Three approaches were devised to understand the attributes of the Brighton population – these are outlined in **Table 6.2**. The distribution of infection category, using all three approaches, is given in **Figure 6.4.**

Approach one

Approach one described the population through taking the first calendar quarter of attendance. Patients were further subdivided into those diagnosed before, and during, the study period. Of the patients diagnosed before the study period, 2.2% ((11/498) 95% confidence intervals (CI) 0-14.95) were recently HIV-infected at their first attendance during the study period (none were HIV-diagnosed during this calendar quarter). In contrast, 34.3% (148/431, 95%CI

30.1-38.9) of patients diagnosed during the study period were recently HIV-infected at their first clinic attendance (which was their calendar quarter of diagnosis) Overall, during the first calendar quarter of attendance during the study period, 18.5% (159/859, 95%CI 16.1-21.3) were recently HIV-infected, 41.0% (352/859, 95%CI 37.7-44.3) were chronically HIV-infected and not ARV treated, 32.0% (275/859, 95%CI 29.0-35.2) were chronically HIV-infected and ARV treated, and 8.5% (73/859, 95%CI 6.8-10.6) were chronically HIV-infected and currently interrupting ART.

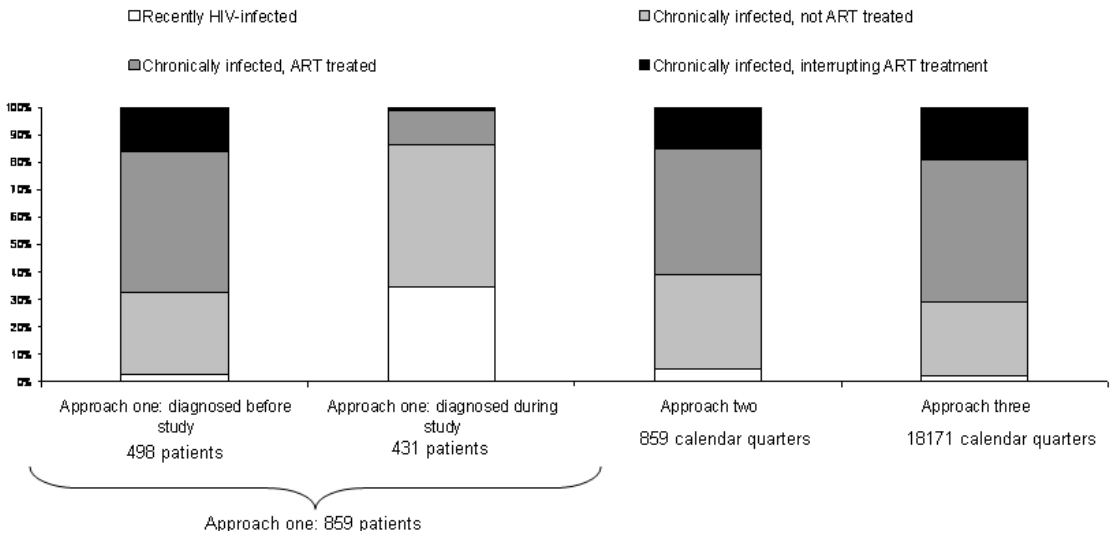
Approach two

Approach two included every patient who attended during the study period but for each, a random calendar quarter of the study period was selected **(Table 6.2)**. Arguably, this approach shows a "snap-shot" of the population between 2000-6. Using this approach, 4.7% (40/859, 95%CI 3.4-6.3) of patients were recently HIV-infected, 34.3% (295/859, 95%CI 31.2-37.6) were chronically infected and untreated, 46.1% (396/859, 95%CI 42.9-49.4) were chronically HIV-infected and treated and a further 14.9% (128/859, 95%CI 12.7-17.4) were currently interrupting treatment.

Approach three

This approach included every calendar quarter of the study period for every patient (from their first quarter of attendance to the end of the study period (or last quarter of attendance if lost to follow-up)).

Figure 6.4: Infection category distribution among diagnosed HIV-infected MSM, Brighton: 2000-2006



Overall, 2.1% (355/16988, 95%CI 1.9-2.3) of calendar quarters during the study periods were linked to patients experiencing recent HIV infection. A further 27.6% (4695/16988, 95%CI 27.0-28.3) of quarters were linked those who were chronically infected and untreated, and 55.4% (9411/16988, 95%CI 55.4-56.2) of quarters were linked to the chronically HIV-infected and treated. Finally, 14.9% (2531/16988, 95%CI 14.4-15.4) of quarters were linked to those currently interrupting treatment.

The distribution of infection category over time is presented, using data from approach three in **Figure 6.5.** The proportion of calendar quarters linked to the patients who were recently HIV-infected increased from 1.49% (28/1875) in 2000 to 2.8% (80/2834) in 2005 (p=0.001) but then dropped to 0.64% (18/2774) in 2006. The proportion of calendar quarter linked to patients who were currently ARV treated increased from 50.1% (956/1875) in 2000 to 63.4% (1760/2774) in 2006 (p<0.0001).

Plasma viral load by infection category

This section uses approach three to describe the data. Plasma viral load by infection category, using every patient and every quarter they were represented in the study period, is presented in **Figure 6.6.** There was a strong association between viral load and infection category (Yates' chi square, corrected for continuity p<0.0001). Those not currently on treatment (the recently and chronically infected, and those interrupting treatment) had significantly higher viral loads than those currently on treatment. Specifically, the proportion of calendar quarters linked to patients with viral loads over 10,000 copies/mL was 69.3% (95%CI 64.0-74.09, 221/319) among the recently HIV-infected, 78.1% (95%CI 76.8-79.4, 2957/3784) among the chronically-infected and untreated patients, and 47.7% (95%CI, 45.7-49.8, 1077/2257) among patients interrupting treatment. In contrast the proportion of calendar quarters linked to patients with viral load over 10,000 copies/mL was 3.5% (95% CI 3.2-3.9, 316/8943) among the treated population, Chi square, (p<0.0001).

The proportion of calendar quarters linked to patients with viral loads under 50 copies/mL was 5.64% (95%CI 3.6-8.74, 18/319) among the recently HIV-infected, 0.71% (95%CI 0.49-1.03, 27/3784), among the chronically HIV-infected who were untreated and 31.10% (95%CI 29.2-33.0, 702/2257) among calendar quarters linked to those currently interrupting treatment. In contrast, the proportion of calendar quarters linked to patients with viral loads of under 50 copies/mL was 83.1% (95%CI 82.3-83.8) among the treated population (Chi squared, p<0.001).

Figure 6.5: Distribution of patient infection category for each calendar quarter of the study period (using approach three), Brighton: 2000-2006

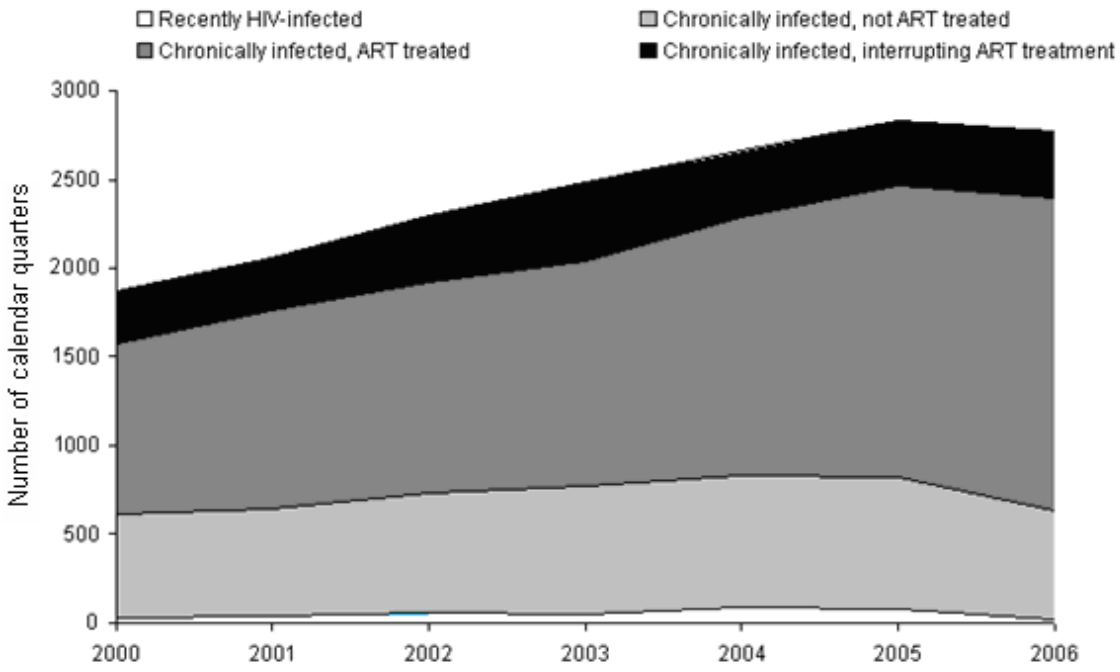
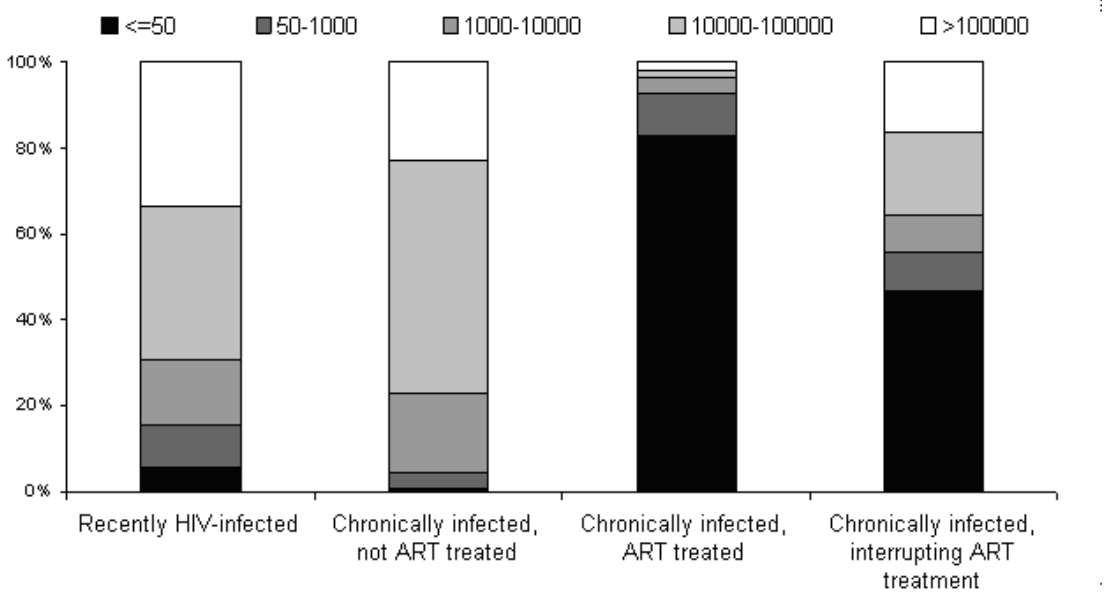


Figure 6.6: Distribution of patient viral load, by infection category, using data from every calendar quarter of the study period (using approach three) Brighton: 2000-2006

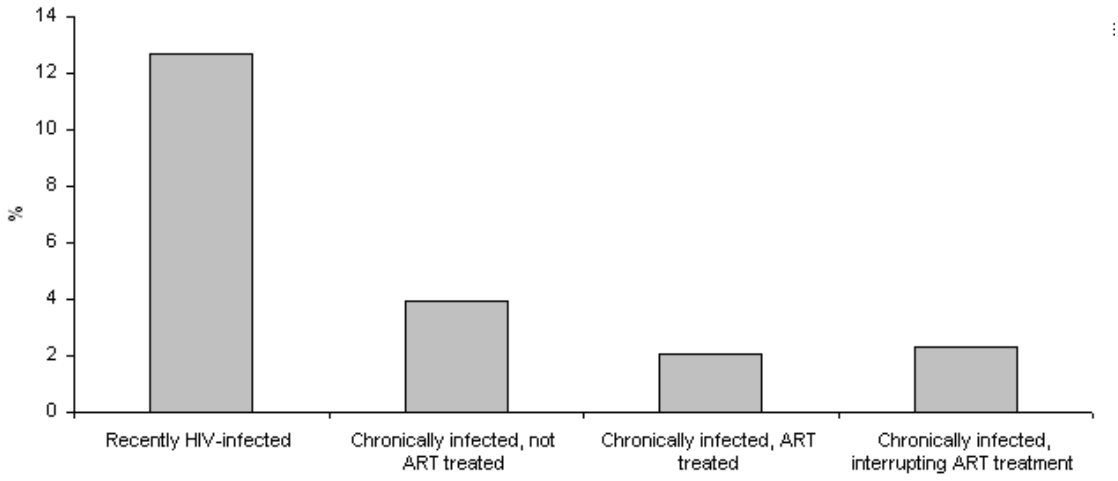


STI diagnosis by infection category

Overall, 39.5% (339/859) patients had an STI diagnosis at some point during the study period. Excluding patients who had STIs diagnosed at the same quarter as their HIV diagnosis, this proportion reduced to 35.5% (305/859). Of those who were diagnosed with an STI 64.6% (219/339) had one diagnosis, 23.0% (78/339) were diagnosed twice, 8.8% (30/339) were diagnosed three times, 3.2% (11/339) four times and 0.29% (1/339) five times. Data were only available on whether patients had an STI diagnosis, and not on the frequency of STI screening.

Using approach three, the proportion of calendar quarters linked to patients who had STI diagnoses was 12.7% (45/355, 95%CI 9.6-16.6) among attendances from recently HIV-infected population **(Figure 6.7)**, 3.9% (185/4695) from the untreated, 2.1% (193/9407) from the treated, and 2.3% (58/2531) from those currently interrupting treatment.

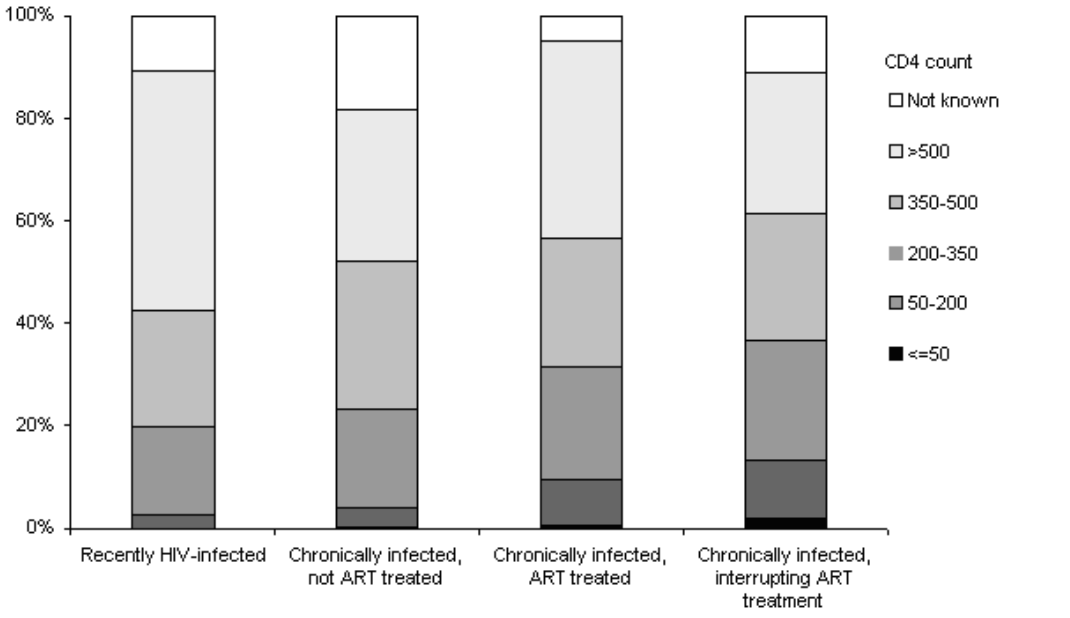
Figure 6.7: Proportion of calendar quarters linked to diagnosed HIV-infected MSM with an STI diagnosis, by infection category, (using approach three) Brighton: 2000-2006



CD4 by infection category

The distribution of patient CD4 count by infection category is provided in **Figure 6.8**. Using approach three, no calendar quarters (0/355) were linked to patients with CD4 counts less than 50 cells/mm^3 among the recently infected. However, the proportion with CD4 counts under 50 cells/mm^3 was 0.37% (14/3835) from calendar quarters linked to the chronically HIV-infected and untreated population, 0.61% (55/8947) from the currently treated population and 2.14% (48/2248) among the population currently interrupting treatment.

Figure 6.8: Distribution of patient CD4 count, by infection category, using data from every calendar quarter of the study period (using approach three) Brighton: 2000-2006



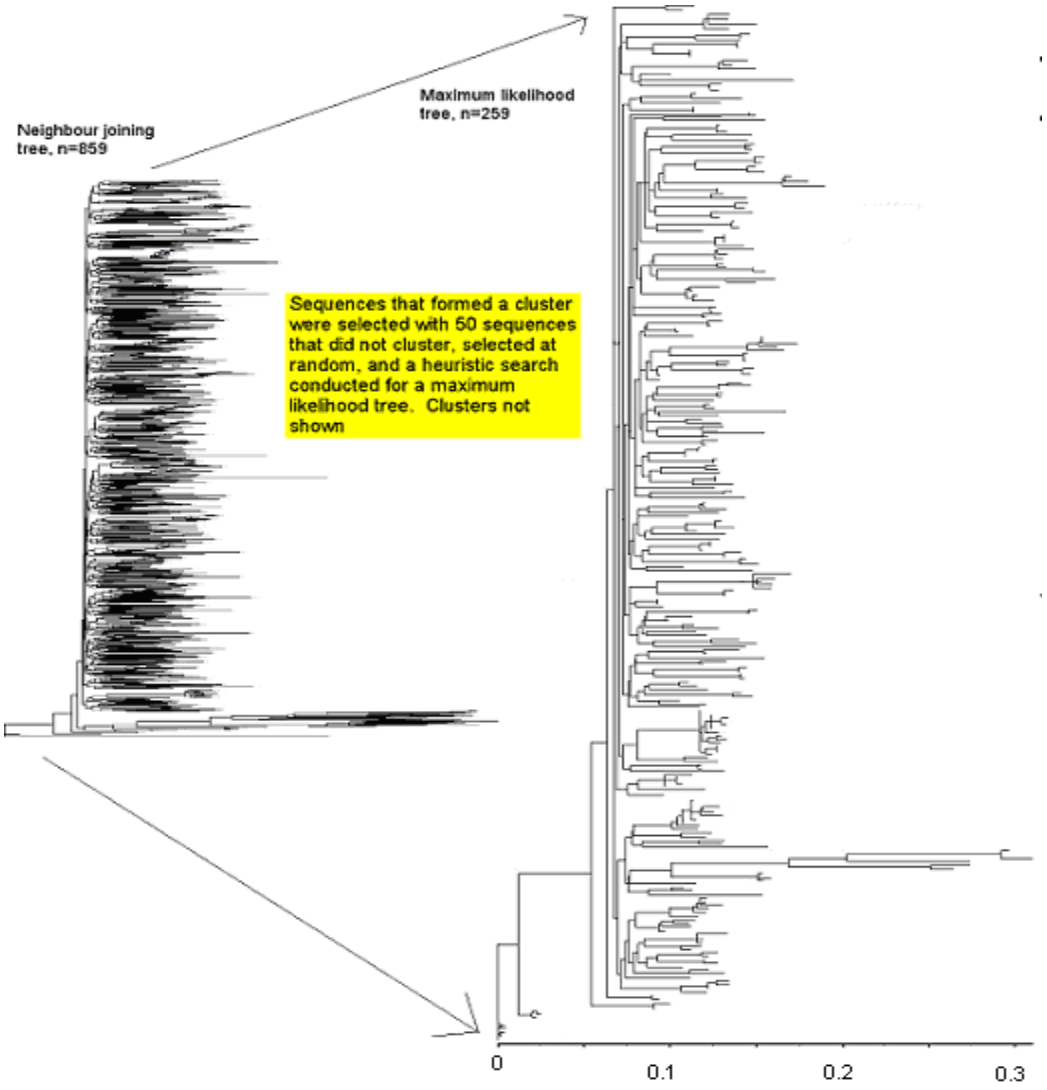
6.3.2 Phylogenetic reconstruction of HIV transmissions in Brighton

Phylogenetic analysis

Using all sequences obtained from Brighton, the first neighbour joining tree found 24.3% (209/859) sequences fell into a robust cluster with another sequence **(Figure 6.9**). These formed 78 clusters (mean size=2.67 sequences per cluster, range 2-10). The 209 clustering sequences were selected, along with 50 sequences that did not cluster, chosen at random, to create a maximum likelihood tree.

The maximum likelihood reconstruction retained 61.7% (129/209) of the original clustering sequences from the neighbour joining tree. Of the original clusters, 19.2% (15/78) were not retained in the maximum likelihood tree. These clusters were disproportionately those containing larger numbers of sequences per cluster. The mean number of sequences per cluster was 2.08 (range 2-3) for the retained clusters, and 3.43 (range 2-9) for the clusters lost. Eight novel clusters, involving 19 sequences, were formed in the maximum likelihood tree that had not clustered in the neighbour joining tree. Overall 62 clusters were consistent between the neighbour joining tree and the maximum likelihood tree: only these were considered to be robust clusters for the remainder of this chapter. Within clusters, the genetic distances altered slightly for some sequences within clusters, between the two trees, but still met the definition of robust cluster.

Figure 6.9: Phylogenetic reconstruction of transmission events among diagnosed HIV-infected MSM, Brighton: 2000-2006



Attributes of clustering sequences

Overall, 15% (129/859) of sequences fell into a robust cluster with another sequence. Of the 159 patients who were recently HIV-infected at the first attendance during the study period, 29.6% (47/159) fell into a cluster with any other sequence and 11.3% (18/159) fell into a cluster with one or more sequences from a patient diagnosed with recent infection in the same or consecutive calendar quarter. In comparison, 11.7% (82/700) of sequences from chronically HIV-infected patients clustered with any other sequence.

Characteristics associated with sequences that did cluster, and sequences that did not cluster, are provided in **Table 6.3.** Sequences that clustered were more likely to be from patients who were recently HIV-infected at their first attendance during the study period compared to those who were not recently HIV-infected at their first attendance: 36.4%, (47/129) vs. 15.3% (112/730) p<0.001). Similarly, sequences that formed a robust cluster were less likely to have been from a patient treated at some point compared to those that did not cluster: 22.5% (29/129) vs. 43.7% (319/730), p=0.028. An STI diagnosis at the first attendance was associated with an increased risk of clustering: of those that clustered, 16.3% (21/129) had an STI at first attendance, compared to 6.2% (45/730) of those who did not have an STI at their first attendance (p<0.001). Equivalent figures for those who *ever* had an STI diagnosis during the study period were 56.6% (73/129) and 36.9% (270/730) (p=0.0002).

Table 6.3: Characteristics of the *pol* sequences that did, and did not, cluster, Brighton: 2000-2006

Risk factor	Sequences that clustered (N=129)		Sequences that did not cluster (N=730)		Chi square
	%	n	%	n	
Recent HIV infection	36.4	47	15.3	112	p=0.028
Ever treated	22.5	29	43.7	319	p<0.0001
Age-group*					
<35	58.1	75	37.0	270	p<0.0001
>35	41.9	54	63.0	460	p<0.0001
World region of birth					
UK	71.3	92	72.5	529	p=0.8
Sub-Saharan Africa	4.7	6	3.3	24	p=0.4
Rest of Europe	12.4	16	7.7	56	p=0.07
STI at diagnosis	16.3	21	6.2	45	p<0.001
STI during study period	56.6	73	36.9	270	p<0.0002

6.3.3 Revised method: where do new infections come from?

Transmitters with estimated transmission dates

Of the 159 sequences that were derived from patients who were recently HIV-infected at first attendance during the study period, 47 fell into a robust cluster and could thereby be allocated a candidate most likely transmitter. Of the 47 candidate transmitters identified, 42 were confirmed the most likely transmission source of the patients recently HIV-infected at diagnosis. This is because the transmission sources were diagnosed *before* the recent HIV-infected patient with whom they were associated (or same calendar quarter if they had a chronic infection), indicating the direction of transmission (see section 6.2.4). Of the 42 transmission sources, 41 had data available during the calendar quarter in which the transmission took place (i.e. the calendar quarter of diagnosis of the recent infection generated) **(Figures 6.10** and **6.11)**.

Two potential transmission sources were excluded since it was impossible to select the most likely transmitter in these instances. This is because the sequences from the two candidate transmitters were identical (although one sequence was shorter than the other, indicating that this was not a cutting and pasting error). This left 39 most likely transmission sources. Of the 39 transmitters, 27 were unambiguously paired with the recent infection and 12 formed a cluster with the recent infection alongside additional sequences.

Figure 6.10: Phylogenetic reconstruction of HIV transmission events among diagnosed HIV-infected MSM attending Brighton clinic, with most likely transmission sources highlighted, Brighton: 2000-2006

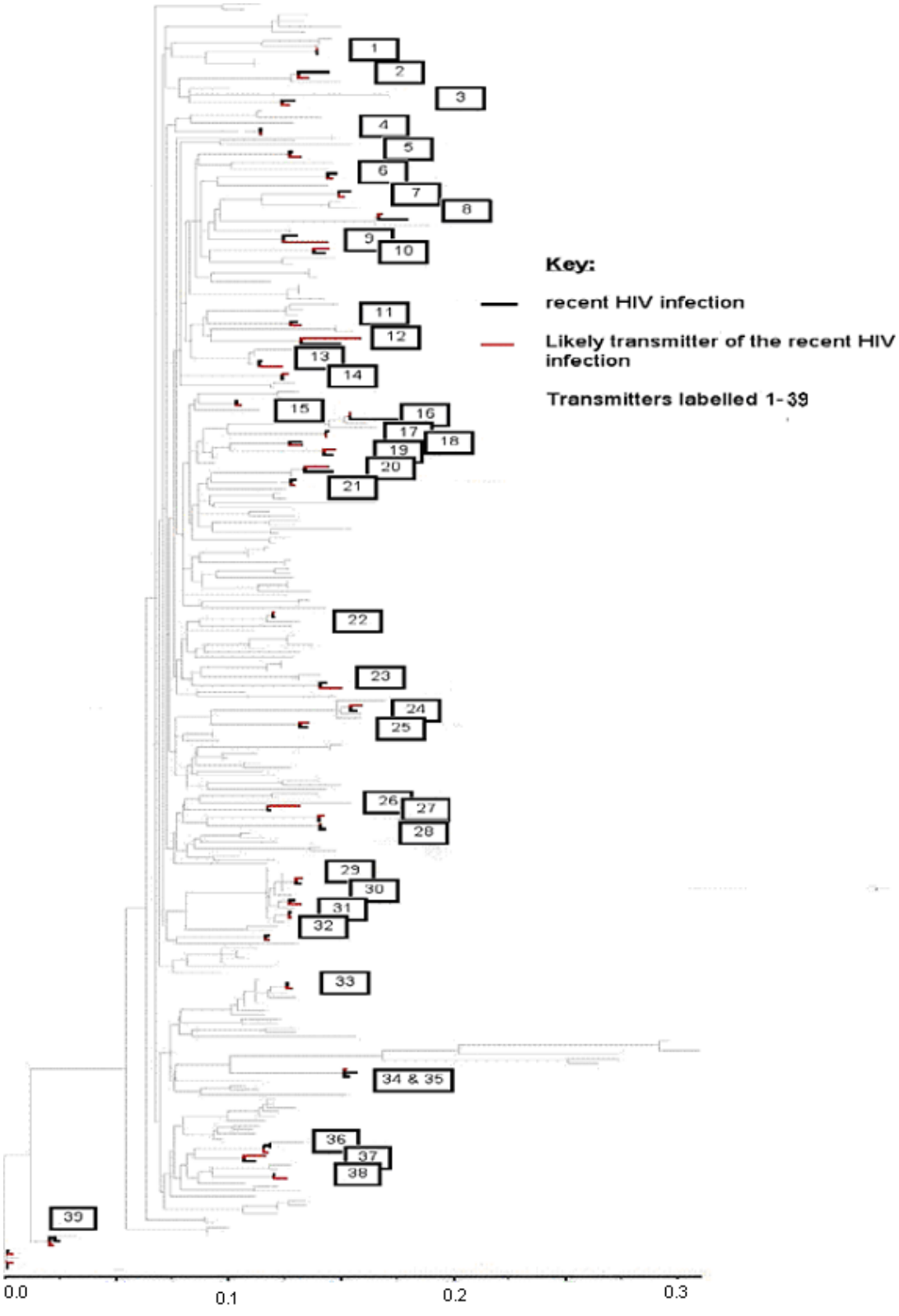
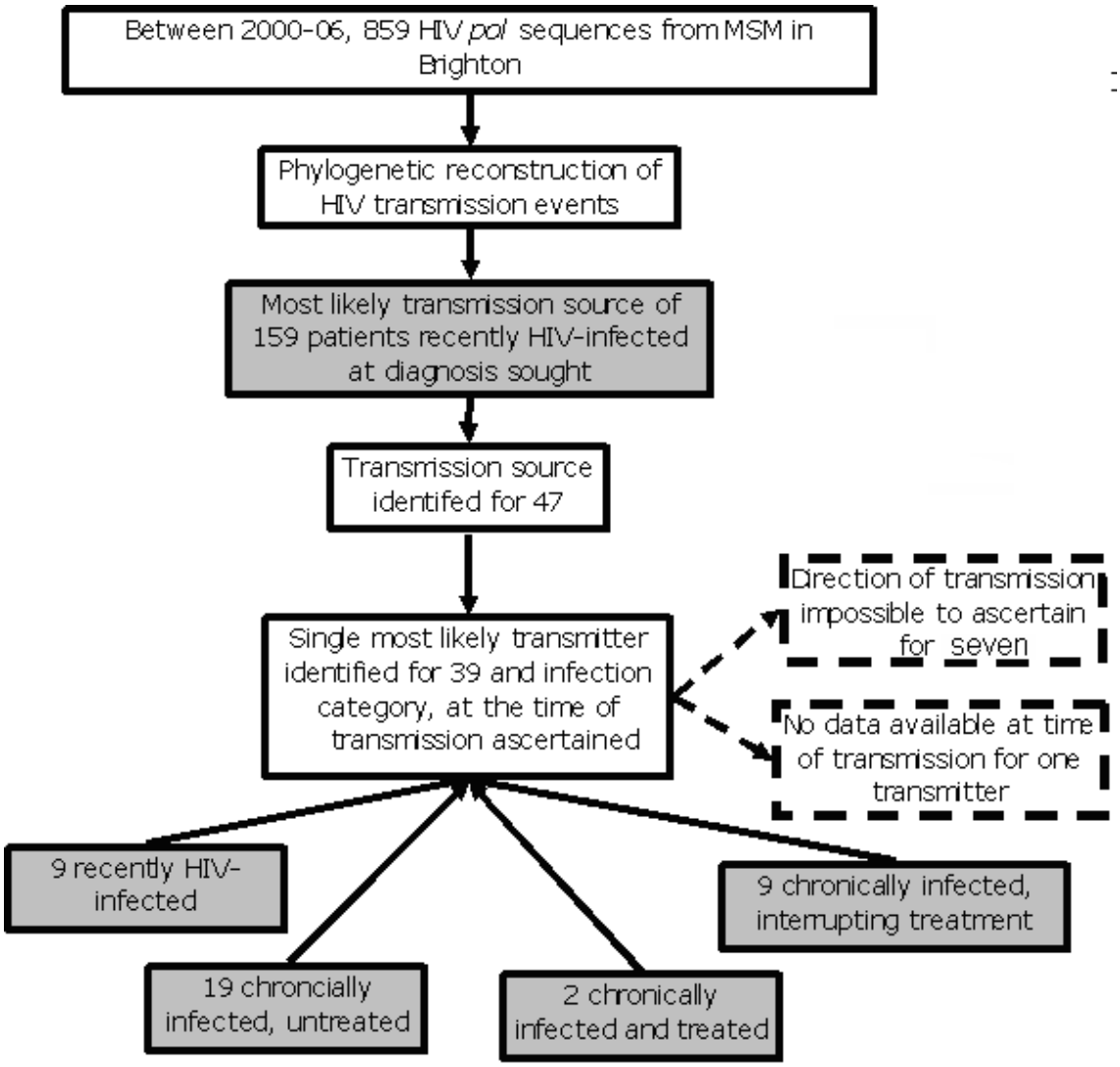


Figure 6.11: Flow diagram showing ascertainment of transmission sources with estimated transmission dates, Brighton: 2000-2006



The 39 most likely transmission sources were predominantly white (37/39), and born in the UK (29/39) with the remainder born elsewhere in Europe and in sub-Saharan Africa. The majority of transmitters were aged under 35 years (27/39) with the transmissions disproportionately occurring during or after 2005 (16/39).

Descriptive analysis

Overall, of the 39 transmitters with estimated transmission date, 24.4% (9/39) were recently infected at the time of transmission and 46.3% (19/39) were chronically HIV-infected but untreated. The remainder comprised 4.8% (2/39) who were currently on treatment, and 22.0% (9/39) who were undergoing treatment interruption. The distribution of the infection categories of the transmissions was compared the distribution of the infection categories of concurrent diagnosed HIV-infection population (described through approach three) (**Figure 6.12).**

At the estimated time of transmission, 39 of the most likely transmission sources, 2.7% (1/39) had plasma viral loads under 50 copies/mL, with an additional 2.7% (1/39) with viral loads between 50-1000 copies/mL. A further 7.7% (3/ 39) had viral loads from 1001-10,000 copies/mL, and 38.5% (15/39) had viral loads from 10,001-100,000 copies/mL. An additional 38.5% (15/39) had viral loads over 100,000 copies/mL. **Figure 6.13** shows the distribution of viral load for the transmitters compared to concurrent Brighton population, described through approach three.

Figure 6.12: Comparison of the distribution of infection category between the transmission sources with estimated transmission dates and the concurrent Brighton population (using data from every calendar quarter of the study period - approach three) Brighton: 2000-2006

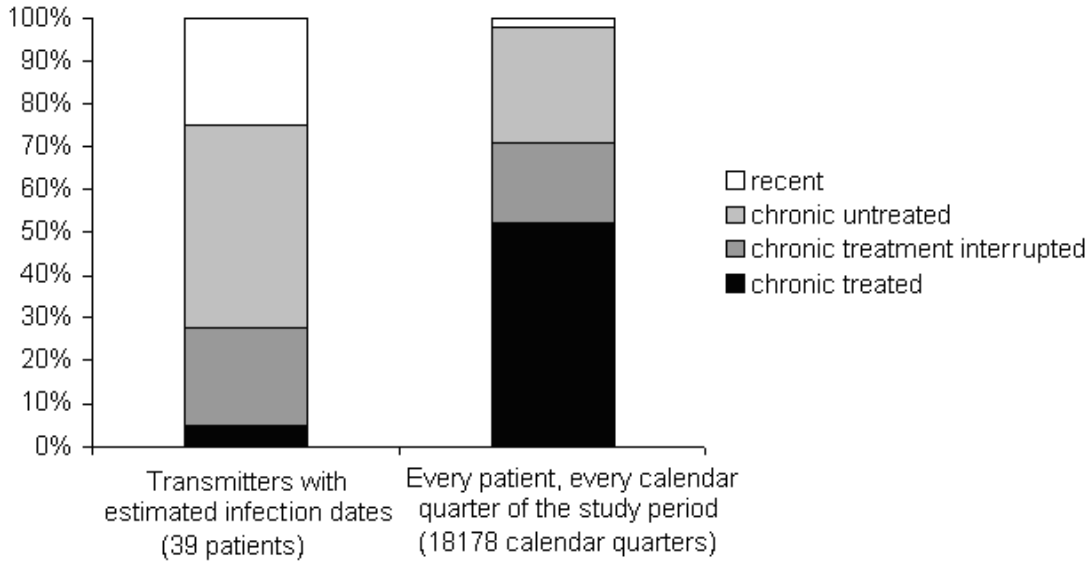
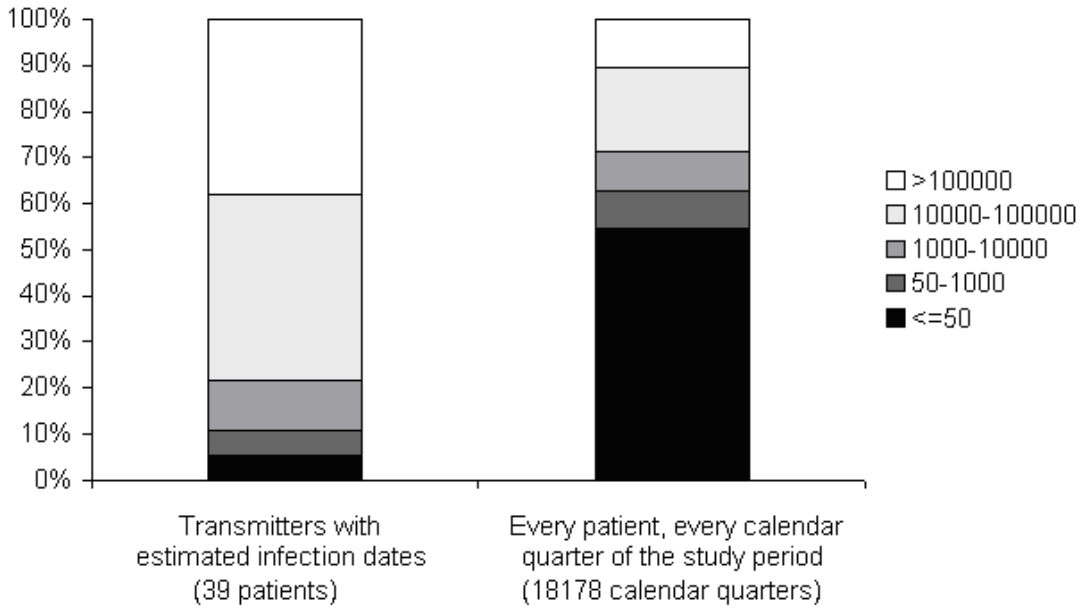


Figure 6.13: Comparison of the viral load distribution between the transmission sources with estimated transmission dates and the concurrent Brighton population (using data from every calendar quarter of the study period - approach three) Brighton: 2000-2006



At the estimated time of transmission, 0% (0/39) of transmission sources had a CD4 count less than 50 cells/mm^3, 5.1% (2/39) had CD4 counts from 50-200 cells/mm^3, 23.1% (9/39) had CD4 counts from 200-350 cells/mm^3 and 28.2% (11/39) had CD4 cell counts from 350-500 cells/mm^3 and 41.0% (16/39) over 500 cells/mm^3. **Figure 6.14** shows the distribution of the CD4 counts of the transmitters compared to the concurrent Brighton population

Overall 25.6% (10/39) of the transmission sources had an STI diagnosis around the estimated time of transmission. **Figure 6.15** shows the distribution of STI diagnoses among the transmitters compared to the concurrent Brighton population. Overall, 69.2% (27/39) of patients had an STI diagnosis at some point during the study period, excluding two who only had STI diagnoses during the same calendar quarter in which they were diagnosed.

Figure 6.14: Comparison of the CD4 count distribution between the transmission sources with estimated transmission dates and the concurrent Brighton population (using data from every calendar quarter of the study period - approach three) Brighton: 2000-2006

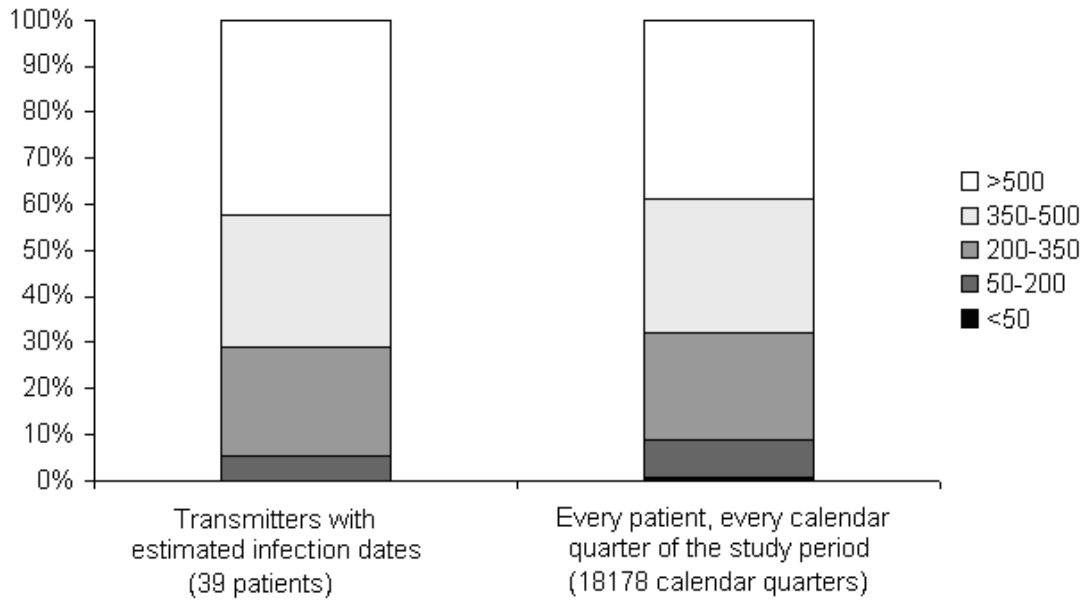
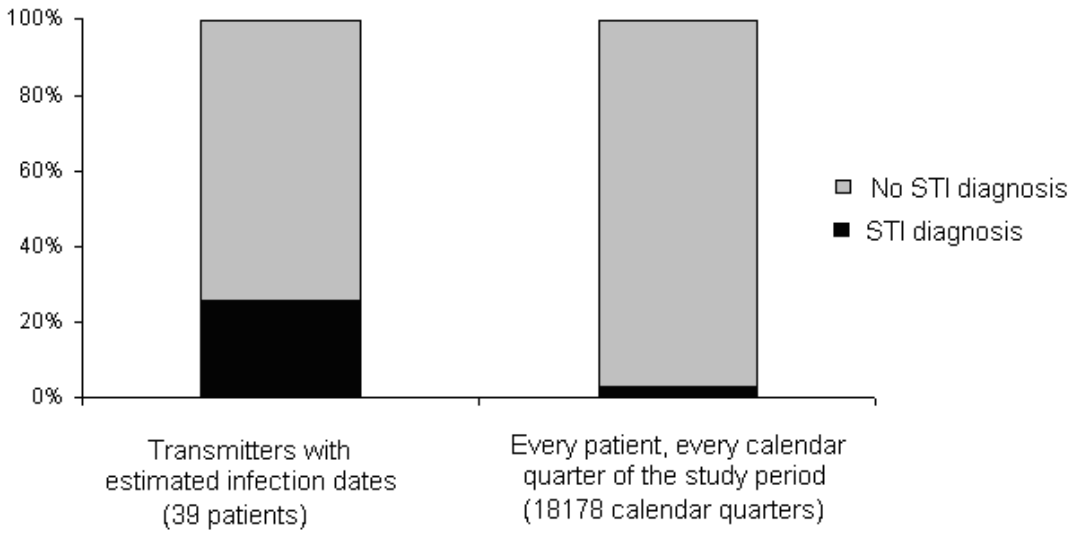


Figure 6.15: Comparison of the distribution of STI diagnoses between the transmission sources with estimated transmission dates and the concurrent Brighton population (using data from every calendar quarter of the study period - approach three) Brighton: 2000-2006



Among the transmission sources, nine had their estimated transmission dates within the same calendar quarter in which they themselves were diagnosed with HIV. This suggests they were likely to be undiagnosed at the time of transmission.

Statistical analysis

In order to account for the biases inherent within the descriptive analyses (e.g. the shorter length of time that the recently HIV-infected population have to generate infections compared to chronically infected population) the rate of observable HIV transmission per person years of follow-up was ascertained for each risk factor **(Table 6.4**). Overall, 39 transmissions occurred from the most likely transmission sources during 6176 years of patient follow-up. This generated a transmission rate of 0.63 per 100 person years of follow-up

(PYFU). Transmission rates of over one per 100 years of follow up were found among the under 35s, those with viral loads over 10,000 copies/mL, the recently HIV-infected and the chronic untreated population.

Univariable and multivariable analyses

Rate ratios of transmission were calculated using Poisson regression models. The results from the univariable analysis are shown in **Table 6.5**, and the results from the multi-variate analysis are shown in **Table 6.6.**

The univariable analysis indicated that the rate ratio of transmission increased 2.38 per log10 of viral load increase. STI diagnosis during the same period as transmission had a rate ratio of 12.53. The rate ratio of transmission was higher among younger ages ($p<0.0001$). AIDS was associated with lower transmission rates; there was no association between CD4 cell count and HIV transmission rate.

Within the multivariable analysis, overall, the factors associated with transmission were: younger age; high viral load; recent HIV infection; and STI diagnosis at around the same period of transmission. Treatment was associated with lower transmission risk in a univariable model, rate ratio 0.04 (0.01-0.19; p=0.0), but the effect was less profound in the multivariable model rate ratio 0.24 (0.05-1.24; p=0.11).

Table 6.4: Number of transmissions, person-years of follow-up (PYFU), transmission rates and 95% confidence intervals generated by transmission sources with estimated transmission dates, Brighton: 2000-2006

Factor		Transmissions	PYFU	Rate (/100PYFU)	95%CI
Overall		39	6167	0.63	0.46-0.86
Age-group	<35	27	1502	1.80	1.24-2.61
	35-44	9	2592	0.35	0.18-0.66
	>45	3	2082	0.14	0.05-0.42
CD4	<200	2	582	0.34	0.09-1.24
	201-350	9	1269	0.71	0.37-1.34
	351-500	11	1564	0.70	0.39-1.25
	>500	16	2242	0.71	0.44-1.15
	Not known	1	521	0.19	0.03-1.08
Viral load	<50	1	3176	0.03	0.01-0.18
	50-1000	1	482	0.21	0.04-1.17
	1001-10,000	3	427	0.70	0.24-2.04
	10,001-100,000	15	941	1.59	0.97-2.61
	>100,000	15	611	2.45	1.49-4.0
	Not known	4	541	0.74	0.29-1.89
Infection category	Recently HIV-infected	9	194	4.64	2.46-8.58
	Chronic infection, untreated	19	1485	1.28	0.8-2.0
	Chronic infection, treated	2	3556	0.06	0.02-0.21
	Chronic infection, treatment interruption	9	941	0.96	0.51-1.81
AIDS		38	6016	0.63	0.46-0.86
		1	160	0.63	0.11-3.46
STI diagnosis	No	29	4972	0.58	0.4-0.83
	Yes	10	1204	0.83	0.45-1.52
Year	2000-1	5	1403	0.36	0.015-0.84
	2002-3	12	1663	0.72	0.41-1.26
	>2004	22	3050	0.72	0.48-1.11

Table 6.5: Univariable analysis of transmission rate ratio using poisson regression model, generated by transmission sources with estimated transmission dates, Brighton: 2000-2006

Factor		Rate ratio	95% CI	p-value
Age-group	<35	5.37	2.53-11.37	0.0001
	35-44	1		
	>45	0.41	0.11-1.53	0.19
	Per five years older	0.51	0.41-0.65	0.0001
CD4	<200	0.49	0.10-2.25	0.36
	201-350	1		
	351-500	0.99	0.41-2.39	0.99
	>500	1.01	0.44-2.28	0.99
	Not known	0.81	0.22-3.00	0.75
	Per 50 cells per mm^3 higher	1	0.94-1.07	0.93
Viral load	<50	0.07	0.01-0.37	
	50-1000	0.44	0.08-2.42	0.002
	1001-10,000	1		0.35
	10,001-100,000	1.7	0.24-2.04	
	>100,000	2.44	0.56-5.13	0.34
	Not known	0.79	0.8-7.42	0.12
	Per log 10 higher	2.38	1.82-3.11	0.74
Infection category	Recently HIV-infected	4.03	1.88-8.68	0.0001
	Chronic infection, untreated	1		0.0004
	Chronic infection, treated	0.04	0.01-0.19	0.0001
	Chronic infection, treatment interruption	0.75	0.34-1.65	0.47
AIDS	No	1		
	Yes	0.11	0.01-0.77	0.03
STI diagnosis	No	1		
	Yes	12.53	6.13-25.64	0.0001

Table 6.6: Multivariable analysis of transmission rate ratio using poisson regression model generated by transmission sources with estimated transmission dates, Brighton

Factor		Rate ratio	95% CI	p-value
Age (per five years older)		0.68	0.54-0.86	0.001
Viral load (per log higher)		1.64	1.16-2.31	0.005
Infection category	Recent	3.06	1.32-7.08	0.009
	Recent, untreated	1		
	Chronic treated	0.26	0.05-1.38	0.11
	Chronic, treatment interruption	1.66	0.71-3.86	0.24
STI diagnosis	No	1		
	Yes	6.07	2.83-12.99	0.0001

Sensitivity analyses

Sensitivity analyses were conducted to assess the impact of the algorithms used to estimate infection date, and the impact of missing data on the results. This involved repeating the multivariable analyses four times, each time adjusting a specific variable under investigation.

a) Excluding patients who had missing CD4/viral load data at the start of each calendar quarter.

Under this variation, records that had missing CD4/viral load data at the start of each calendar quarter were excluded. The number of transmitters with estimated infection dates dropped from 39 transmitters over 6167 PYFU to 27 transmitters over 5463 PYFU. The association between increased transmission rate ratio with viral load and infection category did not remain significant **(Table 6.7).**

Table 6.7: Multivariable analysis of transmission rate ratio from transmission sources with estimated transmission dates, using poisson regression model, excluding patients who had missing CD4/viral load data at the beginning of each calendar period, Brighton: 2000-2006

Factor		Rate ratio	95% CI	p-value
Age (per five years older)		0.70	0.53-0.91	0.007
Viral load (per log higher)		1.4	0.97-2.04	0.08
Infection category	Recent	2.13	0.48-9.40	0.32
	Recent, untreated	1		
	Chronic untreated	0.18	0.03-0.99	0.05
	Chronic, treatment interruption	1.37	0.55-3.37	0.50
STI	No	1		
	Yes	3.88	1.32-11.36	0.01

b) *Taking data from the quarter end, rather than the quarter start, and excluding patients with missing data*

Under this variation, data related to the patients was taken from the end of each calendar quarter (rather than the beginning), and patients with missing data were excluded. Three transmission sources were lost. The relationship between increased rate ratio of transmission risk, and viral load remained **(Table 6.8)**. However, the association with infection category became non-significant.

Table 6.8: Multivariable analysis of transmission rate ratio from transmission sources with estimated transmission dates, using poisson regression model, taking treatment data from the end of the quarter and excluding patients who had missing data, Brighton: 2000-2006

Factor		Rate ratio	95% CI	p-value
Age (per five years older)		0.68	0.54-0.86	0.001
Viral load (per log higher)		2.26	1.56-3.26	0.0001
Infection category	Recent	1.31	0.38-4.51	0.67
	Recent, untreated	1		
	Chronic untreated	0.54	0.14-2.06	0.36
	Chronic, treatment interruption	1.14	0.49-2.63	0.76
STI	No	1		
	Yes	6.58	3.08-14.07	0.0001

c) *Reclassifying recent HIV infection based on earliest infection date*

Each patient had an estimated infection date calculated as the mid-point of their earliest and latest infection date, based on the markers used to ascertain recent HIV infection (see section 6.2.1). For this variation, the estimated infection date was taken as the earliest date.

When patients were categorized as being recently HIV-infected based on the earliest possible date of infection then only two of the transmissions occurred from the "recently HIV-infected" **(Table 6.9**). The association between increased rate ratio of transmission and recent HIV infection stage did not retain its significance. The relationship with viral load was retained.

Table 6.9: Multivariable analysis of transmission rate ratio from transmission sources with estimated transmission dates, using poisson regression model, using the earliest possible infection date, Brighton: 2000-2006

Factor		Rate ratio	95% CI	p-value
Age (per five years older)		0.68	0.54-0.86	0.001
Viral load (per log higher)		2.26	1.56-3.26	0.0001
Infection category	Recent	1.31	0.38-4.51	0.67
	Recent, untreated	1		
	Chronic untreated	0.54	0.14-2.06	0.36
	Chronic, treatment interruption	1.14	0.49-2.63	0.76
STI	No	1		
	Yes	6.58	3.08-14.07	0.0001

d) *Reclassifying recent infection based on the latest infection date*

When patients were categorized as being recently HIV-infected based on the latest possible date of infection, a further nine transmissions were classified as occurring during "recent infection". The association between increased transmission rate ratio and infection category and viral load, was retained **(Table 6.10**).

Table 6.10: Multivariable analysis of transmission rate ratio from transmission sources with estimated transmission dates, using poisson regression model, using the latest possible infection date, Brighton: 2000-2006

Factor		**Rate ratio**	**95% CI**	**p-value**
Age (per five years older)		0.67	0.53-0.85	0.0008
Viral load (per log higher)		1.69	1.20-2.39	0.003
Infection category	Recent	3.39	1.48-7.73	0.004
	Recent, untreated	1		
	Chronic untreated	0.28	0.05-1.44	0.13
	Chronic, treatment interruption	1.64	0.71-3.78	0.25
STI	No	1		
	Yes	5.86	2.77-12.40	0.0001

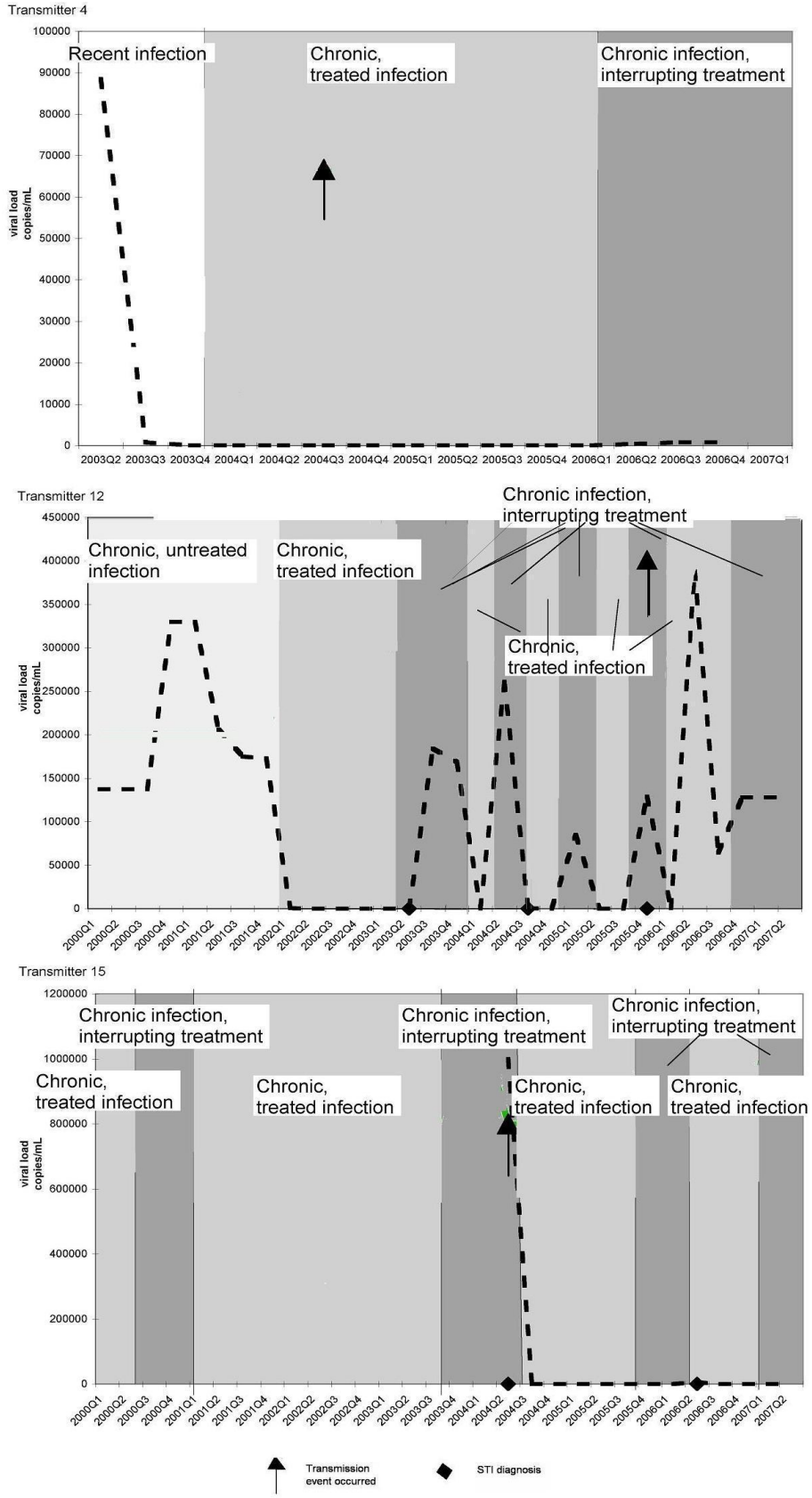
Case studies of transmitters

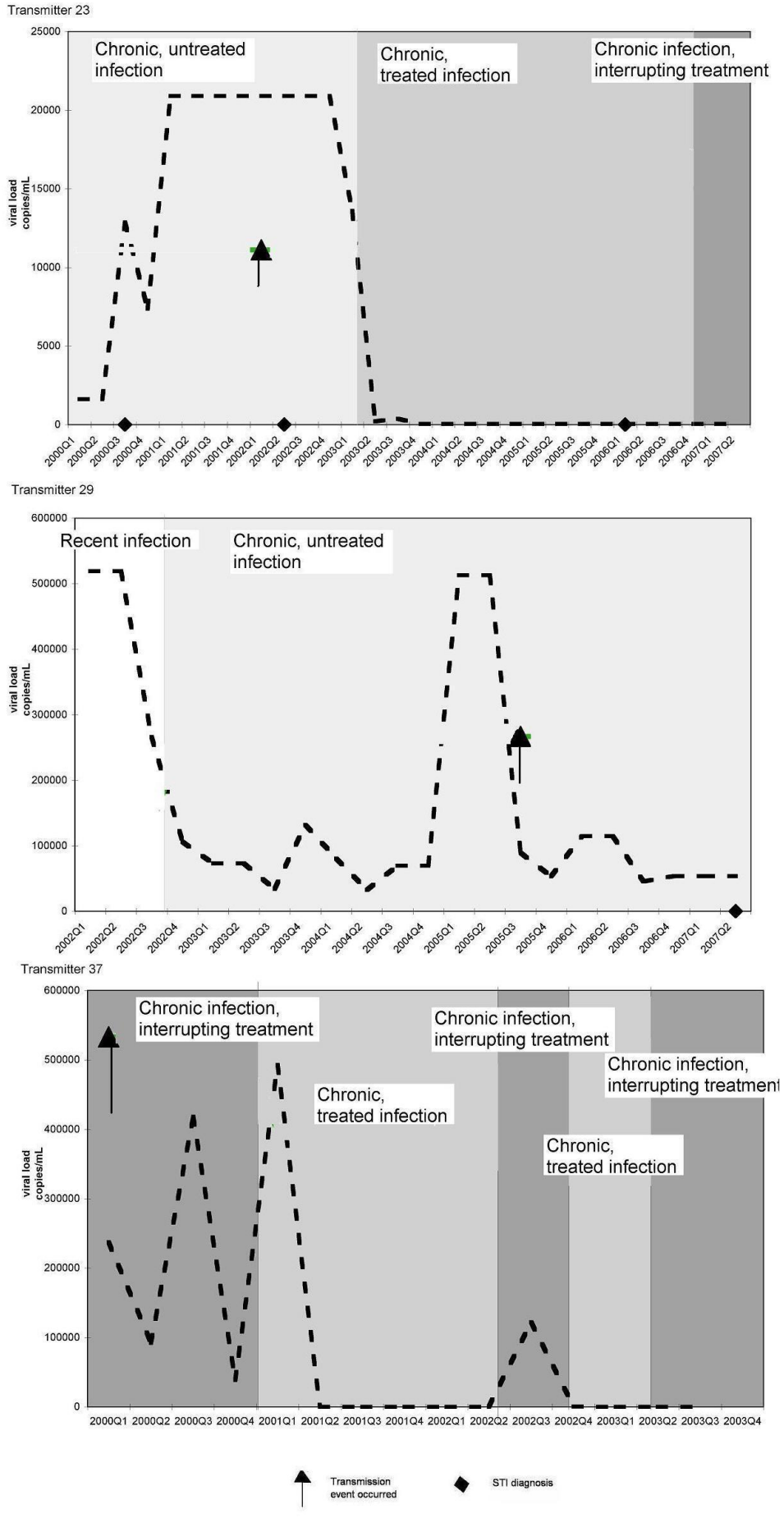
The life histories of most likely transmission sources are presented graphically in **Figure 6.16**. Selected graphs are presented to illustrate relevant points: all of the graphs are presented in appendix C. For each graph, the transmission event is marked as a black arrow; the viral load is shown on the y axis. The background indicates the infection category. STI diagnoses are marked with a black diamond.

The life history graphs demonstrate the potential of plasma viral load to fluctuate within an individual (transmitter 37 and 38) and its relationship to treatment (transmitter 12, 23, 30). Frequently, possible transmission events coincided with a surge of viral load (transmitter 9, 29, and 33) and/or treatment interruption (transmitter 12, 15 and 29). Overall, 55% (15/27) of transmitters had an STI diagnosis, which coincided with the calendar period of estimated transmission (or an adjacent calendar quarter).

One of the two transmitters, who were estimated to have generated infection whilst being on treatment, was interrupting treatment on the calendar quarter before the transmission (transmitter 15). There was one instance out of 39 where there was no “explaining factor” as to why HIV transmission may have occurred (patient was treated, with viral load <50 copies/mL, no STI – transmitter 4).

Figure 6.16: Life histories of a selection of transmission sources with estimated transmission dates, Brighton: 2000-2006




Transmitter 23
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
Transmitter 29
Recent infection
Chronic, untreated infection
Transmitter 37
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic, treated infection
Chronic, treated infection
Transmission event occurred
STI diagnosis

6.4 Discussion

6.4.1 The HIV-infected population and its transmission potential

The Brighton dataset provides a rare opportunity to study localized HIV transmission that can be linked to demographic and clinical data. Over a six year period, patients attending the clinic were subdivided into infection categories, and their clinical data updated for each subsequent calendar quarter. The distribution of infection category was estimated for the Brighton population and analysed with specific regard to its transmission potential. Using a method that improved upon methods of previous analyses, the transmission sources of new HIV infections were sought.

Data completeness

Data were analysed to examine the length of time patients were represented during the study period, and the drop out rate. The size of the dataset increased over time since new patients were diagnosed each year, and the proportion of patients lost to follow-up was relatively low. Consequently, the majority of patients were represented through the entire study period.

Overall, the demographic data had high levels of completeness. Over 90% of viral load and CD4 count data were complete for each patient per calendar quarter. Since the rate of STI screening was unavailable, the STI diagnosis data presented is likely to be an underestimate of the STI prevalence in this population.

Infection category

It is important to ascertain the relative size of the HIV-infected Brighton population by infection category in order to determine whether any particular patient category is disproportionately driving new infections. This is because if all else is equal, the chance that a new infection will be generated by a particular infection category will be directly proportional to the relative size of that infection category. For instance, if the proportion of the HIV-infected population who were recently HIV-infected was 5%, it would be expected that 5% of infections generated to be driven by patients who were recently HIV-infected themselves. If the proportion of new infections generated by those with recent HIV infection was significantly higher than 5% this would suggest that the transmissions generated from this group were disproportionate.

The method for ascertaining the most accurate representation of the Brighton HIV-infected population is therefore crucial. The Brighton dataset was characterized using three approaches to assess which provided the best representation of the size and transmission potential of the local HIV-infected population. Approach one included data collected at the patients' first clinical attendance only; approach two included only one calendar quarter for each patient for the entire study period, with the calendar quarter selected at random; and approach three included every patient, for every calendar quarter during the study period. Each patient could be included up to 28 times within this dataset.

Approach one resulted in one in five patients categorized as recently HIV-infected. This is substantially higher than estimated using approach two (5%) and three (2%). This is because the first calendar quarter of attendance in the study period was synonymous with HIV diagnosis for approximately half the study population. Using diagnosis data to ascertain the proportion of recent HIV infections will lead to an overestimate: patients may have been prompted to attend for VCT (voluntary confidential HIV testing) following a recent risk exposure (Calzavara, Burchell et al. 2002). Approach one demonstrates that using data collected during a patient's diagnosis is likely to overestimate the size of the recently HIV-infected population.

Approach two provided a snap-shot of the Brighton population between 2000-2006 and goes some way to reduce the bias arising from the over inclusion of patients with recent HIV infection – the proportion was measured at 5%. However, bias still remains; this approach draws from a population, half of whom are more likely to be included at earlier stages of their HIV infection. This population, on average, will also have fewer calendar quarters available for selection, increasing the chance that the quarter selected is one in which the patient was recently HIV-infected.

In contrast to approaches one and two, approach three allowed the natural course of HIV infection to progress for each patient over the course of the study period. Approach three showed that the proportion of the HIV-infected population who were recently HIV-infected at any one time is likely to be relatively low, at around 2%. For this reason, approach three was used as the

best representation of the size and transmission potential of the Brighton HIV-infected population, and was consequently used as the comparator for the descriptive analysis.

However, for approach three, it was not known if patients attended for every calendar quarter. Infection category data was updated for each calendar quarter according to section 6.2.1. Where other data were missing for a calendar quarter, the data from the previous quarter was carried forward. Where patients had not attended for 12 consecutive months, they were considered lost to follow-up, and excluded from subsequent quarters of the study period. This method was preferable to only including calendar quarters for each patient, where patients were known to have attended. This is because attendances during the study period are unlikely to be distributed equally between infection category (e.g. patients who are ARV treated may be more likely to attend more frequently than the untreated population). However the potential for each patient to transmit their infection onwards needed to be constant throughout the study period, regardless of clinic attendance.

Approach three is not with out its limitations; like approach one and two, it will exclude certain groups of patients. All approaches exclude patients resident in Brighton who attend other providers for HIV care. Approximately 10% of HIV-infected MSM patients resident in Brighton and Hove city PCT attended sites other than the Brighton clinic for their treatment and care (Personal communication - SOPHID). All approaches exclude patients with undiagnosed HIV infection who are disproportionately likely to be recently HIV-infected and

have higher risk behaviour due to being unaware of their infection. Additionally, approaches two and three exclude those who did not attend the clinic following diagnosis (who may constitute a more vulnerable group that may also have higher infectivity and risk behaviour).

Approach three also contains a major bias within the dataset. Patients who were diagnosed with HIV infection towards the end of the study period will be represented in fewer calendar quarters compared to those diagnosed before the study period. This will have affected those recently HIV-infected at diagnosis disproportionately. It is for this reason that transmission rates per PYFU were calculated for the statistical analysis.

Plasma viral load by infection category

Patients who were not being treated (including the recently HIV-infected population) had higher plasma viral loads compared to the treated population. This is expected since treatment is designed to suppress viral replication (Ledergerber, Egger et al. 1999). The viral load levels of the patients undergoing treatment interruption was slightly lower than that of the untreated patients, perhaps because they have recently been on treatment and/or they are doing well clinically. The high proportion of patients currently being treated who have undetectable viral loads (83% less than <50 copies/mL) is striking. This work supports the evidence that treatment is associated with a reduced patient infectivity.

STI by infection category

Nearly 40% of HIV-infected patients were also diagnosed with an STI at some point during the study period. This reduced only slightly to 35% when the STI diagnoses made during the same calendar quarter as HIV diagnoses were removed. This demonstrates continuing sexual risk among a population aware of their HIV infection.

CD4 count by infection category

CD4 counts were highest among those interrupting treatment and those on treatment.

6.4.2 Phylogenetic reconstruction of HIV transmissions in Brighton

A phylogenetic reconstruction was undertaken among the Brighton population, and the factors associated with clustering were ascertained. Overall, 15.3% of sequences fell into a robust cluster. The rate of clustering between patients who were diagnosed during recent HIV infection was 11.3%. Sequences that clustered were significantly more likely to be from MSM who were recently HIV-infected at diagnosis, have an STI diagnosis, and not ARV treated at first attendance, compared to those that did not form a cluster.

The observed rate of clustering is somewhat lower than that found in the literature. In a similar population, (containing only Brighton patients who were recently HIV-infected at diagnosis, and taken from earlier years), Pao calculated the rate of clustering between patients recently HIV-infected at diagnosis, to be almost double, at 29.6% (Pao, Fisher et al. 2005). Similarly

Brenner found that 50% of sequences clustered from patients who were recently HIV-infected at diagnosis (Brenner, Roger et al. 2007).

The rate of clustering may be lower in the current analysis because the majority of sequences contained within the dataset were not recently HIV-infected at diagnosis. Pao and Brenner both (albeit initially for the latter) restricted phylogenetic reconstructions to include only patients diagnosed during recent infection. Partner notification exercises, and an increased likelihood that patients diagnosed during recent HIV infection within the same calendar period will be drawn from the same transmission network, raises the chance that these populations contain a higher proportion of patients who infected each other. In contrast, sequences from patients who were chronically HIV-infected at diagnosis will have a broader range of infection dates, and will be more likely to be drawn from separate transmission networks.

The present analysis is an improvement upon on the analyses conducted in chapter five, because they are not restricted to populations who were recently HIV-infected at diagnosis. As discussed in section 4.4.2, such populations may be different from populations who were not diagnosed during recent infection, which may be related to infectivity. However, two major limitations remain. Firstly, the risk factors apparently associated with clustering may not have been present at the time of transmission. Secondly, the current population is likely to disproportionately contain patients diagnosed with recent HIV infection. These limitations are discussed in turn.

As with recent infection, the observation of transmission events between sequences from MSM who had a specific risk factor at diagnosis, does not necessarily mean that that risk factor was present at the time of transmission (this is applicable to infection category, viral load, CD4 count, and STI diagnosis). For instance, the comparison of infection dates in the clusters that formed between patients diagnosed with recent HIV infection in this analysis, as with chapter four, found that only about half could have been generated during recent infection. However, an association between clustering with infection category and STI does seem to exist, perhaps indicative that patients with risk factors at diagnosis may be at high risk of transmission throughout the course of their infection. This requires further examination.

This analysis also revealed inconsistencies between phylogenetic reconstructions. There were inconsistencies between the original neighbour joining tree and the maximum likelihood tree. Nearly one in five clusters originally ascertained in the neighbour tree were lost in the maximum likelihood reconstruction. The neighbour joining clusters that did not appear in the maximum likelihood tree contained a higher proportion of sequences per cluster; clusters that constituted sequence pairs were more likely to be retained. The occurrence of novel clusters in the maximum likelihood tree requires further investigation.

Though the datasets contain the same patients, there was a difference between the clustering rate found between patients who were recently HIV-infected at diagnosis in this analysis (11.3% - when the recently and chronically HIV-

infected at diagnosis were included) compared to that found in chapter four (20.8% - when only the 159 patients diagnosed during recent infection were included). This may indicate that reconstructions may have been affected by the presence of other sequences (section 4.4.2).

6.4.3 A revised methodology: where do new infections come from?

A phylogenetic reconstruction was conducted using a dataset from a well defined population to ascertain the origin of new infections. The reconstruction included developments from previous analyses. The richness of the dataset allowed a wide range of risk factors potentially associated with transmission to be explored for patients throughout the study period (compared to studies that only categorise patients as “recently” and “chronically” HIV-infected at diagnosis). Additionally, a method was employed that allowed the attributes of candidate transmitters to be ascertained at around the time of transmission.

A quarter of transmission sources had an STI diagnosis during the calendar quarter, or the previous calendar quarter. Interestingly, the transmitters with estimated transmission dates were more likely to ever have an STI diagnosis during the study period, compared to the rest of the HIV-infected population. This suggests that the transmitters with estimated transmission dates may have riskier sexual behaviours compared with the rest of the population.

Through multi-variate analysis, the factors associated with an increased rate ratio of transmission risk were: recent infection; high viral load; and STI diagnosis. Almost one in four observed transmissions were likely to have been generated by patients with recent HIV infection: in contrast, this group

constituted 2% of the concurrent Brighton population (using approach three). Those with viral loads over 10,000 copies/mL contributed 70% of observed transmissions. Importantly 69.2% of transmission sources originated from patients with CD4 counts over 350 mm^3. This is *above* the level at which discussions about ARV should commence between patient and clinician. Finally, despite comprising more than half of the HIV-infected Brighton population, only two instances were recorded of transmission occurring from the treated population.

The analysis suggests that the transmission rate from the recently HIV-infected is elevated, but not as high as previous calculations (Brenner, Roger et al. 2007). Unlike other studies, it has also highlighted that the transmission rates from the chronically HIV-infected is also important, and varies considerably within this population. Specifically, the untreated population (containing an uncertain, but elevated proportion of the recently HIV-infected individuals) was responsible for generating approximately half of the transmissions identified. Furthermore, the use of treatment was significantly associated with a reduced risk of transmission. Around 20% of transmissions originated from those currently interrupting treatment. Those with the low CD4 counts (<200 copies/mm^3) were least likely to generate transmission. This could be because such individuals are more likely to be treated (and/or more likely to have symptomatic illness), which may impact on their sexual activity.

The sensitivity analysis examined the consistency of the statistical significance of the findings as the assumptions and definitions of treatment and estimated

infection dates were varied. Throughout each repetition, the presence of an STI was consistently associated with an increased transmission risk. When patients with missing viral load/CD4 data at the start of a calendar quarter were excluded, the association with infection stage and viral load did not retain its significance. This is likely to be because the exclusion disproportionately affected patients with recent HIV infection; such patients may not have viral load/CD4 count recorded until a few weeks after diagnosis. Varying the definition of recent infection from the mid-point to earliest and latest possible date of infection substantially impacted on the association between infection category and transmission risk. This indicates the importance of allowing infection category to follow the natural course of infection in the analysis.

Case studies of transmitters

The case studies provide some validation of the methods of section 6.4.3. The fluctuating viral loads observed within individuals confirm the importance of finding a method to date the transmission event in relation to viral load: viral load at diagnosis will not reflect viral load at later stages of infection. The congruence of transmission events with surges of viral load, lack of treatment and STI diagnoses provide some reassurance of the accuracy of this analysis in identifying the right individual as the transmitter and the approximate time of transmission. The instances where transmission occurred shortly after viral load surges/STIs/treatment changes may be indicative of the relative inaccuracy of the technique in approximating the transmission date. The case studies put the findings into context, but should not be over-interpreted.

In one of the two instances where transmission occurred despite the patient being on treatment, the case studies suggest that the patient had been

interrupting treatment very recently. For the second patient, it may be that transmission occurred despite the patient having an undetectable plasma viral load. Alternative explanations include: the data may not have captured surges of viral load at the relevant transmission time; the phylogenetic reconstruction was incorrect, and did not identify actual transmission source.

Limitations

The analysis has several limitations related to: the representativeness of the population of transmitter sources with estimated transmission dates; and assumptions about sexual mixing.

As with all phylogenetic reconstructions, the results may not necessarily represent actual transmission events (section 1.6.1 and 1.6.2 and chapter four). Even in cases where the clusters remained consistent between the neighbour joining tree and the maximum likelihood tree, small differences in genetic distances were observed within clusters. This is relevant because it was at this level that decisions were made on which candidate sequence was the most likely transmitter. The overall method relies exclusively upon data from specific transmission pairings. If the specific pairings are incorrect, then the results will be incorrect. While the life histories demonstrate a congruence of transmission dates with events that are likely to increase the chance of transmission, the accuracy of phylogenetic approaches in reconstructing specific transmission events requires further assessment. Finally, the analysis does not consider chains of transmissions; just transmission pairs.

The markers used to ascertain the estimated infection dates varied. Consequently using the diagnosis date of the recently HIV-infected patients in

combination with the marker used to identify recent infection to “date” the transmission events may not be entirely accurate. The “calendar quarter of transmission” may have been up to six months after the actual transmission event, particularly where the serological testing algorithm for recent HIV seroconversion (STARHS) was used. Additionally, extent that STARHS can reasonably reconstruct infection dates is uncertain due to variation between individuals (Reed 2004). However, no adjustments were made in these instances. Future work needs to include intervals of uncertainty or calculate what scale of adjustments need to be made in such analyses.

The data were summarized into three monthly intervals to anonymise the data; attendance dates were not available. For viral load, and treatment data, the data were carried forward from the previous calendar quarter unless that data had changed. Consequently the accuracy of data around the time of the estimated transmission event is dependent on the patient attending the clinic regularly and having the relevant fields updated. The exclusion of patients lost to follow-up (no attendance within 12 months) from candidate transmission sources will limit any error. However, in order to assess the timeliness of data linkable to the 39 transmitters at the time of transmission, the occurrence of “movement” in viral load during the period of transmission, or the quarter directly preceding or following it, was taken as evidence that such measures were timely.

In all but seven of 39 transmission sources, there was movement in viral load in the period of transmission or the calendar quarters consecutive with it. For the remaining seven transmission sources: transmitters 4, 18 and 24 all had a gap

of under six months between the quarter a new viral load were recorded, each side of the transmission quarter. For transmitters 16, 23 and 30, there was a gap of 12 months in total spanning the interval in which the transmission event occurred. For each of these, there was no recorded change in treatment during the period, suggesting that viral load was unlikely to change substantially. Transmitter 2 did not have a viral load recorded until treatment was started, but was untreated at the time of transmission, suggesting viral load was detectable at this time.

The multivariable analysis assumed each person was at risk for the whole three month period of the calendar quarter when calculating person years. This assumption is probably reasonable since even if a patient was not diagnosed until half way through the period, they could still have transmitted in the weeks leading up to the diagnosis.

The methods do not consider the undiagnosed HIV-infected population. This group is likely to have an important role in transmission since they are: unaware of their infection; more likely to be recently HIV-infected; and have riskier behaviour, compared to the rest of the HIV-infected population. It is noted that for 39 transmitters, nine transmitted during the period close to their own diagnosis, suggesting they were undiagnosed at the time of transmission. Additionally, the fact that only 39 transmission sources of 159 recent infections were identified suggests that the undiagnosed population is likely to have a substantial role in generating transmission events.

The transmission sources had higher rates of STI diagnoses compared to the concurrent Brighton population. This suggests that they may be a group with higher risk sexual behaviour. This indicates that behavioural, in addition to biological factors drive transmission. However, behaviour markers were not measured or considered in the analysis.

Further limitations relate to assumptions made about the Brighton population. The methodology assumes that the Brighton population is closed, with random sexual mixing patterns. However patients attending care in Brighton may meet sexual partners elsewhere in the UK, and abroad. It is also known that sexual risk behaviours are affected by diagnosis status (Dodds, Mercey et al. 2004) and perceived infectivity is related to treatment (Stephenson, Imrie et al. 2003).

6.5 Conclusion

The Brighton dataset has provided a rare opportunity to understand the sources of new infections in a localised population. It allowed the development of a method capable of creating phylogenetic reconstructions of transmission events, but also of dating these events. Consequently, this enabled the identification of risk factors for transmitters, at around the time of transmission.

The factors associated with transmission are: recent HIV infection; elevated viral load, and presence of an STI. Additionally, this chapter provided estimates of the size of the population in relation to transmission risk; the association between infection category and transmission; that despite making up 2% of the HIV-infected population, the recently HIV-infected generate a quarter of new infections; that 70% new infections originate from the untreated population; 70%

of new infection originate from those with CD4 counts >350 mm^3; and that despite making up more than half of the HIV-infected population, few transmissions occur from the treated population.

This work has important public health implications. Current BHIVA guidelines (BHIVA 2008) (that suggest discussions to start ARV should commence when CD4 counts reach 200-350 cells/mm^3) are unlikely to impact upon transmissions between MSM. The analysis suggests that treating HIV-infected patients on a large scale, and reducing phases of treatment interruption, could potentially reduce transmission.

Additionally, with a quarter of observed transmissions originating from the recently HIV-infected, the continued drive to increase VCT may be limited in its ability to prevent HIV transmission. Alternative strategies include promoting: frequent HIV testing among the at-risk population; awareness of seroconversion illness and its association with higher infectivity; and universal access to fourth generation HIV tests. The need to prevent and treat STIs continues.

The analysis is not without its limitations. Future work needs to: better account for the undiagnosed HIV-infected population; ensure algorithms that ascertain recent infection are consistently applied; and consider the implications and importance of rapid transmission chains.

7 Chapter Seven: The prevalence, source and onward transmission of HIV drug resistance

This chapter describes the prevalence of transmitted HIV drug resistance among patients who were recently HIV-infected at diagnosis. The distribution of drug resistance mutations are compared between datasets and infection categories. The transmission potential of the drug resistant population is compared between the drug naïve and those who have been treated. Phylogenetic reconstructions of possible transmission of resistant strains are undertaken. Finally, the sources of transmitted drug resistance are explored using a phylogenetic approach.

7.1 Introduction

It is the higher prevalence of TDR among the recently HIV-infected (Devereux, Youle et al. 1999) coupled with their elevated viral load (Wawer, Gray et al. 2005) that suggests that this population has an important role in generating TDR (section 1.5.3). However, this may not be the only factor in determining TDR. Patients with diagnosed HIV infection continue to have unprotected sex (Elford and Hart 2005) and a proportion of them will have drug resistant viruses. Chapter six indicated that the recently HIV-infected, the chronically HIV-infected untreated population and those interrupting therapy had potentially important roles in generating onward transmission. Additionally, the relative fitness of mutations may affect which mutations are transmitted onwards (both from men who have sex with men (MSM) with transmitted and acquired drug resistant mutations) (section 1.5.3 and 2.5.2).

This chapter aims to measure the level of TDR within the available datasets, and describe the distribution and transmission of specific mutations between patients recently HIV-infected at diagnosis. Using the whole Brighton dataset, it also aims to ascertain the transmission sources of patients with TDR who were recently HIV-infected at diagnosis, and consequently ascertain their attributes at around the time of transmission.

7.2 Methods

7.2.1 Prevalence of HIV drug resistance

Transmitted drug resistance

The prevalence of TDR was ascertained for patients recently HIV-infected at diagnosis using the CASCADE, UA survey and Brighton datasets. The definition of drug resistance mutations was taken from the list of mutations judged suitable for surveillance purposes (Shafer, Rhee et al. 2007).

Differences between transmitted and acquired drug resistance

For Brighton, the prevalence of drug resistance, and the specific mutations involved were compared between the recently HIV-infected, the chronically HIV-infected and the treated population. Statistical tests of association were employed where appropriate.

Distribution of specific mutations

Specific mutations were identified from sequences obtained from patients recently HIV-infected at diagnosis and compared between datasets. Using the Brighton dataset, all patients (regardless of infection stage at diagnosis) with drug resistant mutations were categorized as: TDR (recently infected), TDR (chronically infected) or acquired on the basis of infection stage, and clinical and treatment history.

7.2.2 Phylogenetic reconstructions of HIV transmissions of TDR from the recently HIV-infected

The phylogenetic reconstructions described in this chapter are identical to those in previous chapters, but this analysis maps the drug resistance mutations on to

sequences included in the reconstructions. Two phylogenetic reconstructions were undertaken from the Brighton dataset: first only including those sequences from patients who were recently HIV-infected at diagnosis (section 5.3.2) and second including the whole dataset (section 6.3.2).

For each dataset, robust clusters (section 2.2.11) that contained drug resistance mutations were identified, and compared to clusters formed between wild-type sequences. Comparisons were made between the mutations occurring among sequences that were involved in a possible transmission event and mutations found among non-clustering sequences. The differences between estimated infection dates were compared within clusters in order to ascertain whether the transmission could have taken place during recent infection.

7.2.3 Attributes and transmission potential of patients with transmitted and acquired drug resistance

Using the Brighton dataset, patients with viruses containing drug resistance mutations were placed in to three categories on the basis of infection stage and ARV history. The categories include: those with TDR and recently HIV-infected at diagnosis; those with TDR and chronically HIV-infected at diagnosis; and those with acquired drug resistance. The plasma viral load was compared between the three groups and also to the population with wild-type viruses. This was to gauge whether the transmission potential of the patients with drug resistance mutations was different to that among those with wild-type sequences.

7.2.4 Where does TDR come from?

Using all of the Brighton dataset, a phylogenetic reconstruction was undertaken of all possible transmission events. The most likely transmission source for each recently-infected patient with TDR was identified using the same methodology described in section 6.2.4. The transmitter was categorised as having acquired resistance or TDR (recently or non-recently acquired) at the time of the likely transmission.

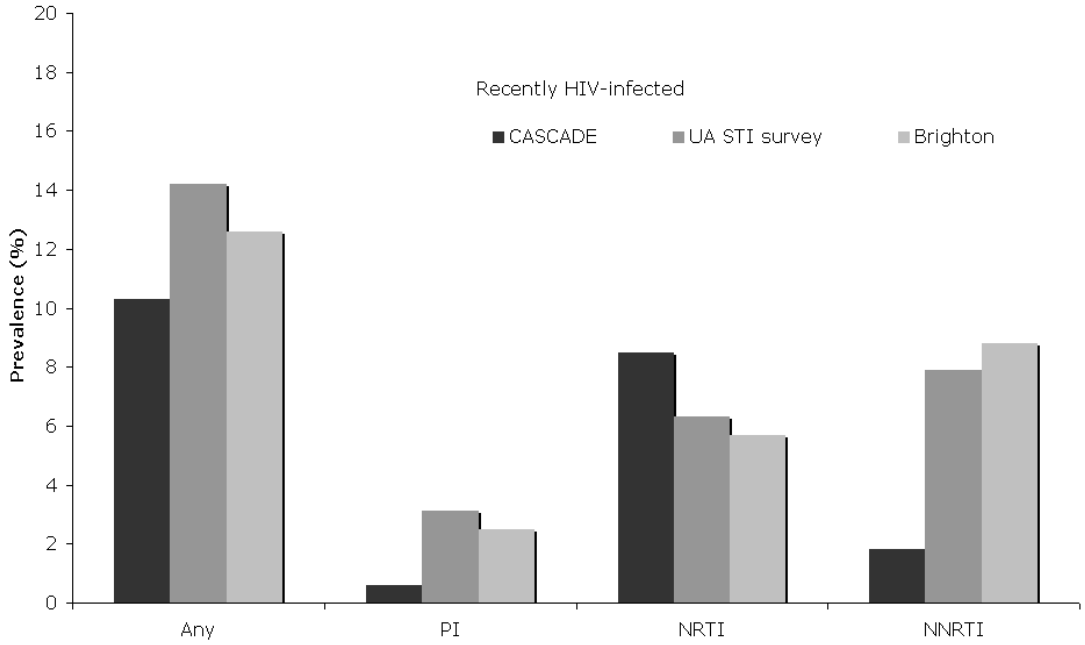
7.3 Results

7.3.1 Prevalence of HIV drug resistance

Transmitted drug resistance

Prevalence of TDR was ascertained using populations that were recently infected at HIV diagnosis only. Overall, 10.9% (18/165) of recently HIV-infected CASCADE patients sampled were infected with a drug resistant HIV variant **(Figure 7.1)**. Fourteen patients had a virus considered to be resistant to one or more of NRTIs, three carried resistance mutations to NNRTIs and one had resistance to PIs**.** One patient was infected with viruses resistant to both NRTI and NNRTI mutations with the remainder being resistant to one class of drug resistance only (14 to NRTIs, two to NNRTIs and one to PIs).

Figure 7.1: Prevalence of TDR among MSM diagnosed during recent infection, CASCADE, UA survey and Brighton: multiple years



Overall, within the Unlinked Anonymous (UA) Sexually Transmitted Infection (STI) survey, 14.2% (18/127, 95% CI 9.3-21.6%) of recently HIV-infected MSM sampled were infected with a drug resistant strain. Eight patients had a virus considered resistant to NRTIs, nine to NNRTIs, and four to PIs. Sixteen patients were infected with a virus considered resistant to one class of drug only, one to two classes and one to three classes of drug.

Among the patients who were recently HIV-infected at diagnosis among the Brighton dataset, 12.6% (20/159, 95%CI 8-18%) of patients sampled were infected with a drug resistant HIV variant. Nine patients had a virus considered to be resistant to one or more of NRTIs, 14 to NNRTIs and four had resistance to PIs. Fifteen patients were infected with a virus considered to be resistant to one class of drug only (five to NRTIs, nine to NNRTIs and one to PIs), with three to two classes and two to all three main classes of drugs.

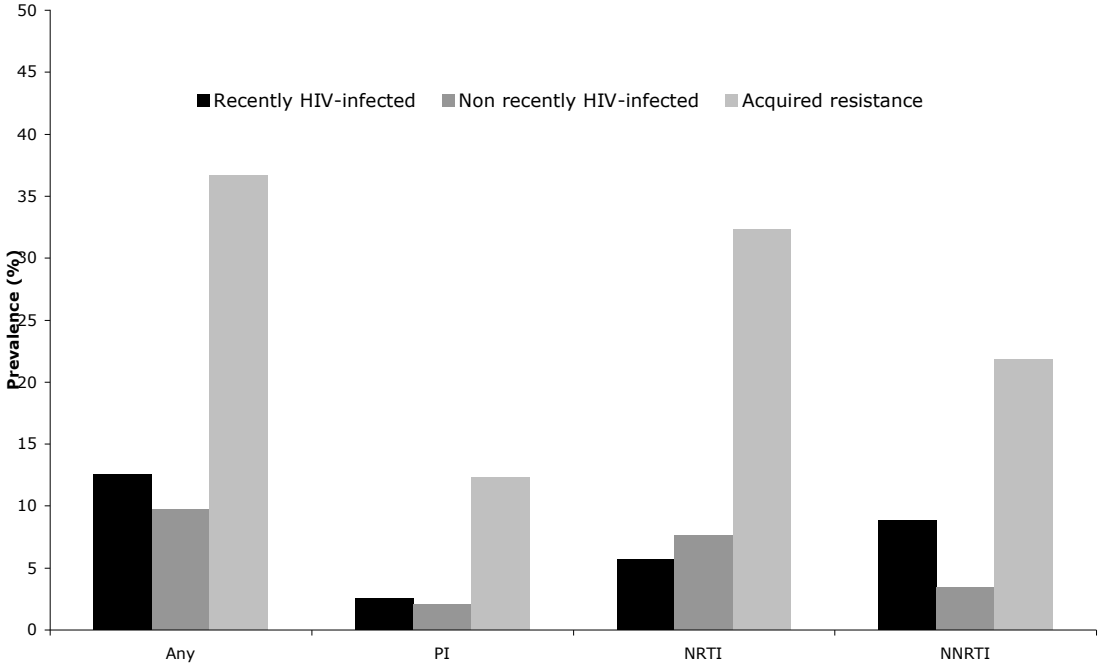
Differences between transmitted and acquired drug resistance

Among the Brighton dataset, overall, 22.2% (191/859) of patients sampled were infected with a drug resistant HIV variant. Of these, 148 patients had a virus considered to be resistant to one or more of NRTIs, 102 carried resistance mutations to NNRTIs and 58 had resistance to PIs. Of these, 106 patients were infected with a virus considered to be resistant to one class of drug only (67 to NRTIs, 34 to NNRTIs and five to PIs), 53 to two classes and 32 to all three main classes.

Brighton patients with drug resistant viruses were further subdivided into three categories: those recently HIV-infected at diagnosis; the drug naïve, but chronically HIV-infected at diagnosis; and the drug experienced at the time the sequence was taken. The differences in prevalence was 12.6% (20/159), 9.7% (34/352) and 39% (74/396) respectively. For 63 patients with drug resistance mutations, the category of drug resistance could not be ascertained from their clinical data.

There were differences in class of drug resistance between drug naïve patients who were recently and chronically HIV-infected at diagnosis. While the prevalence of resistance to PIs was similar between both groups, the prevalence of resistance to NNRTIs may have been higher than the prevalence of resistance to NRTIs among the recently HIV-infected (8.8% vs. 5.7% respectively) (p =0.14). The converse was true of the chronically HIV-infected at diagnosis (3.4% vs. 7.6% respectively) (p=0.007) **(Figure 7.2).**

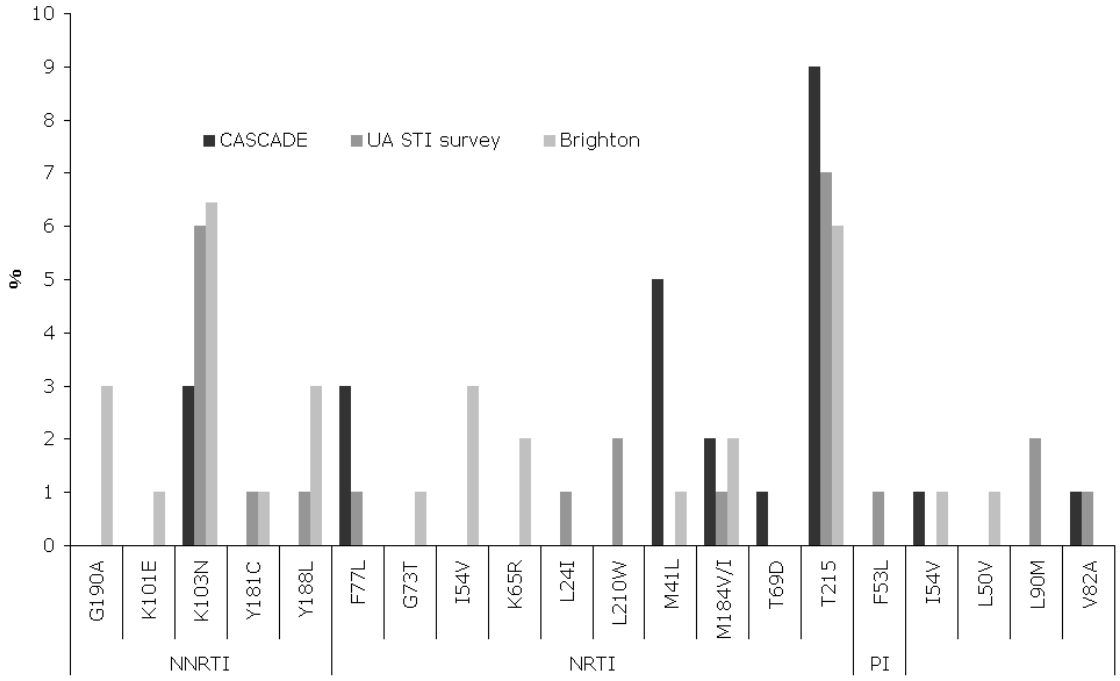
Figure 7.2: The prevalence of drug resistance mutations among the recently HIV-infected, the untreated chronically HIV-infected and the treated chronically HIV-infected population, Brighton: 2000-2006



Distribution of specific mutations

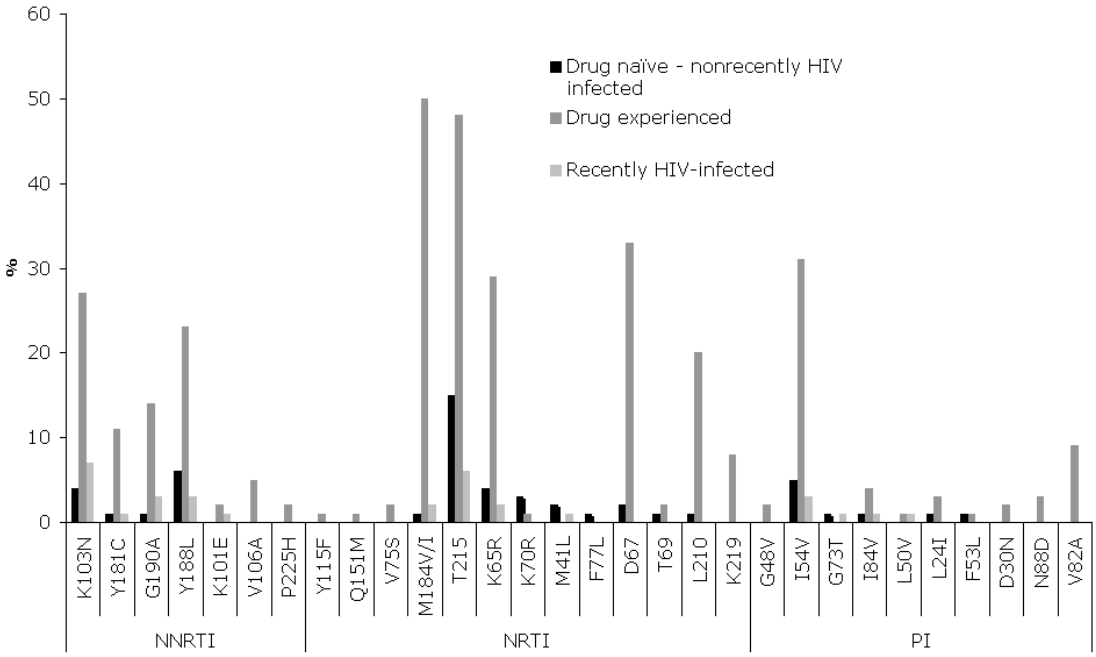
The distribution of specific mutations among those recently HIV-infected at diagnosis is described in **Figure 7.3.** The most frequent mutation overall was the T215 variant followed by the K103N. Among the CASCADE dataset, five patients had M41L, but this mutation was not identified in the other two datasets.

Figure 7.3: Distribution of specific mutations found among recently HIV-infected MSM, CASCADE, UA survey, and Brighton: multiple years



The differences between the specific mutations for each infection category are described in **Figure 7.4.** Where specific mutations were observed in all three categories, the frequency of mutations was substantially higher among the treated population. There were mutations from all three classes that were observed in the treated population that were not apparent in the naïve populations including: V106A (NNRTI), K219E (NRTI), and V82A (PI). Conversely, mutations were observed among the drug naïve populations that were not apparent in the treated population, for instance: M41L and F77L (NRTIs) and G73T (PI).

Figure 7.4: Distribution of specific drug resistance mutations between the recently HIV-infected, the untreated chronically HIV-infected and the treated chronically HIV-infected population, Brighton: 2000-2006



7.3.2 Phylogenetic reconstructions of HIV transmissions of TDR from the recently HIV-infected

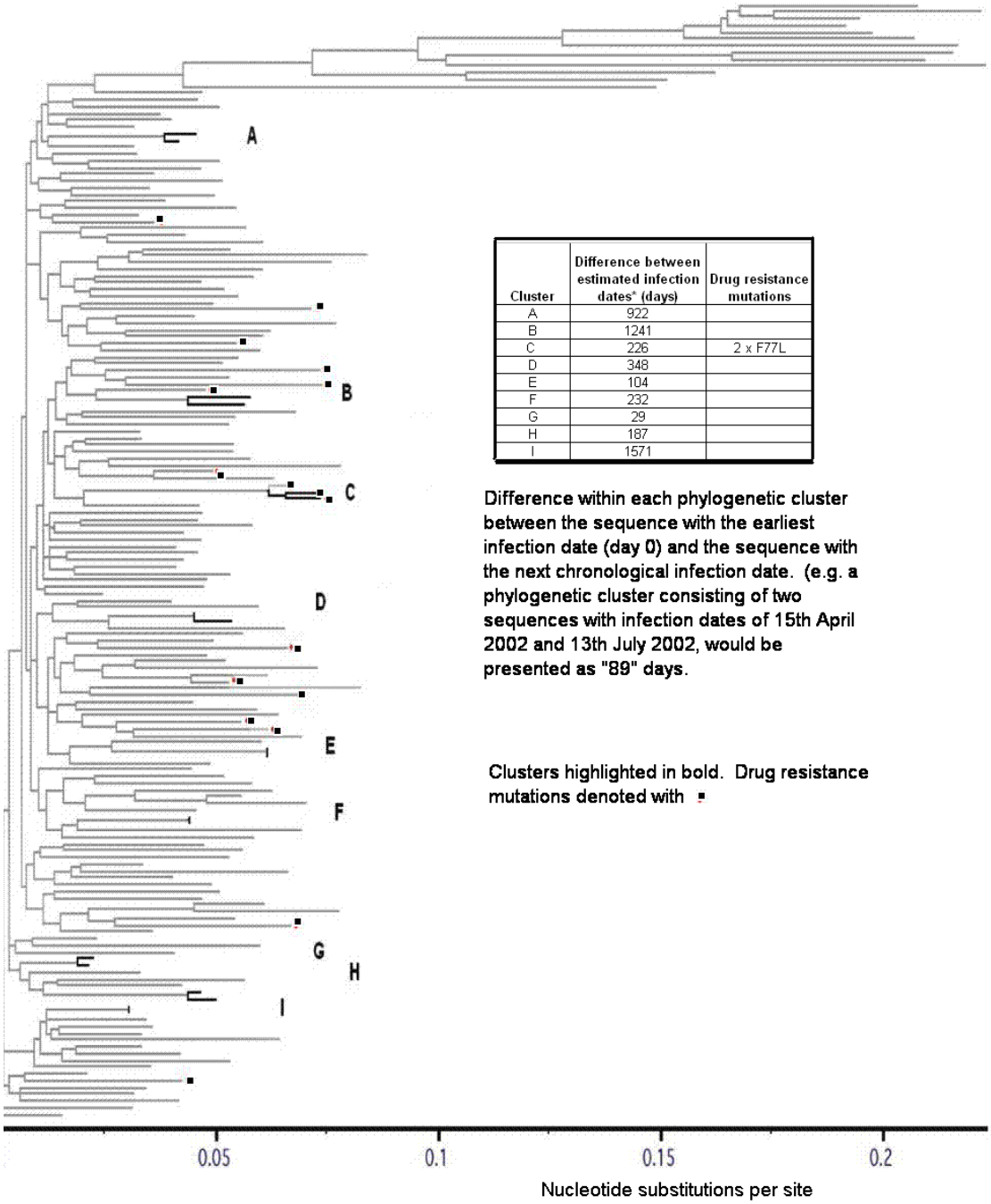
Of the 165 sequences from the CASCADE dataset, 18 (10.9%) formed a phylogenetic relationship with at least one other sequence **(Figure 7.5)**. These formed nine robust clusters, with two sequences within each cluster. Of the 18 sequences with drug resistance mutations, only two formed one cluster (C). Both sequences shared the F77L mutation, and were derived from patients with a difference of 226 days between their infection dates (Brown 2009a).

Of the 127 sequences from the UA STI dataset, 16 (12.6%) formed a robust cluster with at least one other sequence **(Figure 7.6)**. Of the 18 sequences with drug resistance mutations, six formed two clusters (cluster A and cluster

E). All four sequences that formed cluster A contained the K103N mutation, and both sequences that formed cluster E contained T215V/D with one sequence also contained F77L. Cluster A contained two sequences that had the same calendar quarter, and two with a difference of two calendar quarters. Cluster E comprised sequences obtained from patients who had a difference in calendar quarter of nine.

Considering only the recently HIV-infected at diagnosis (n=159) from Brighton, 31 sequences fell into a cluster with at least one other sequence **(Figure 7.7)**. Of the 20 sequences with drug resistance mutations, four fell into a cluster with at least one other sequence. One cluster contained sequences that shared the K103N mutation. Two other sequences (with G190A/E) each formed a cluster with a wild-type sequence.

Figure 7.5: Phylogenetic reconstruction of HIV transmission events and distribution of drug resistance mutations between MSM recently HIV-infected at diagnosis, CASCADE: 1989-2004



Cluster	Difference between estimated infection dates* (days)	Drug resistance mutations
A	922	
B	1241	
C	226	2 x F77L
D	348	
E	104	
F	232	
G	29	
H	187	
I	1571	

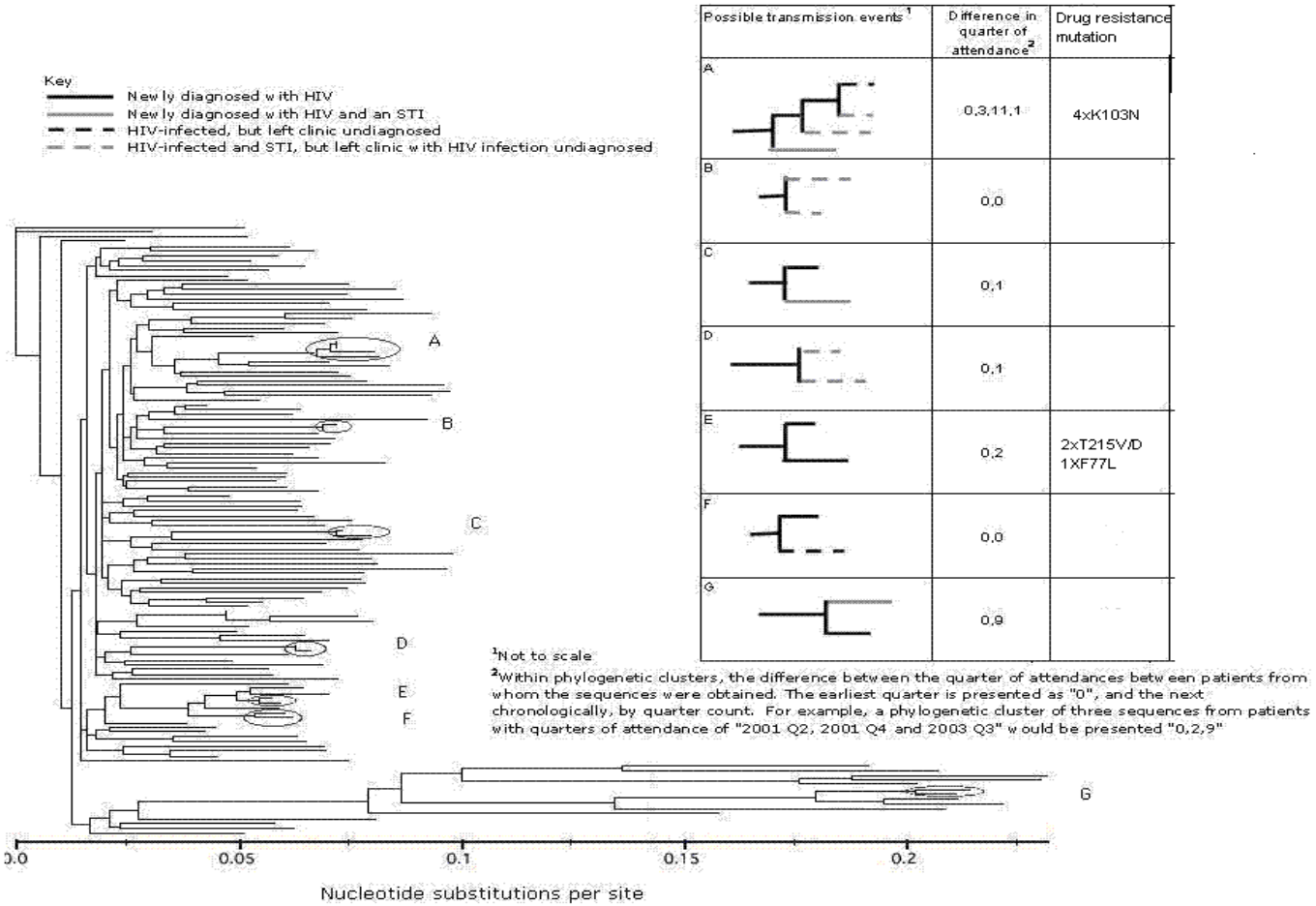
Difference within each phylogenetic cluster between the sequence with the earliest infection date (day 0) and the sequence with the next chronological infection date. (e.g. a phylogenetic cluster consisting of two sequences with infection dates of 15th April 2002 and 13th July 2002, would be presented as "89" days.

Clusters highlighted in bold. Drug resistance mutations denoted with ▪

Using the whole Brighton dataset, overall, 15% (129/859) sequences fell into a robust cluster with another sequence (figure not shown)**.** These formed 62 phylogenetic relationships, five clusters had three sequences per cluster; the remaining 57 consisted of sequence pairs. Of the 191 sequences with drug resistance mutations, 23 fell into a cluster with another sequence. These formed eleven clusters and, of these, four sequences formed a cluster with a wild-type sequence. The remaining nine sequences clustered with another sequence sharing an identical mutation. Three clusters each comprised K103N, and three comprised T215 variants. Other mutations included 154L, G73T, I84V, I54L, and Y188L.

For the CASCADE and Brighton dataset, there was no significant difference between the proportion of clusters that formed between sequences with drug resistance mutations compared to the proportion formed between wild-type sequences **(Figure 7.8).** For CASCADE, overall, 11.1% (2/18) of sequences with mutations clustered compared with 10.1% (16/147) of wild-type sequences (p=0.7). Among the recently HIV-infected Brighton population, 20% (2/20) of sequences with drug resistance mutations clustered compared with 19% (27/139) of wild type sequences (p=0.3). For the UA STI survey, 33.3% (6/18) of sequences with drug resistance mutations clustered, compared with 9.2% (10/109) of wild-type sequences (p=0.03). For all datasets, mutations that were found in clusters tended to be those that occurred most frequently in terms of prevalence.

Figure 7.6: Phylogenetic reconstruction of HIV transmission events and distribution of drug resistance mutations between MSM recently HIV-infected at diagnosis, UA survey: 1999-2002

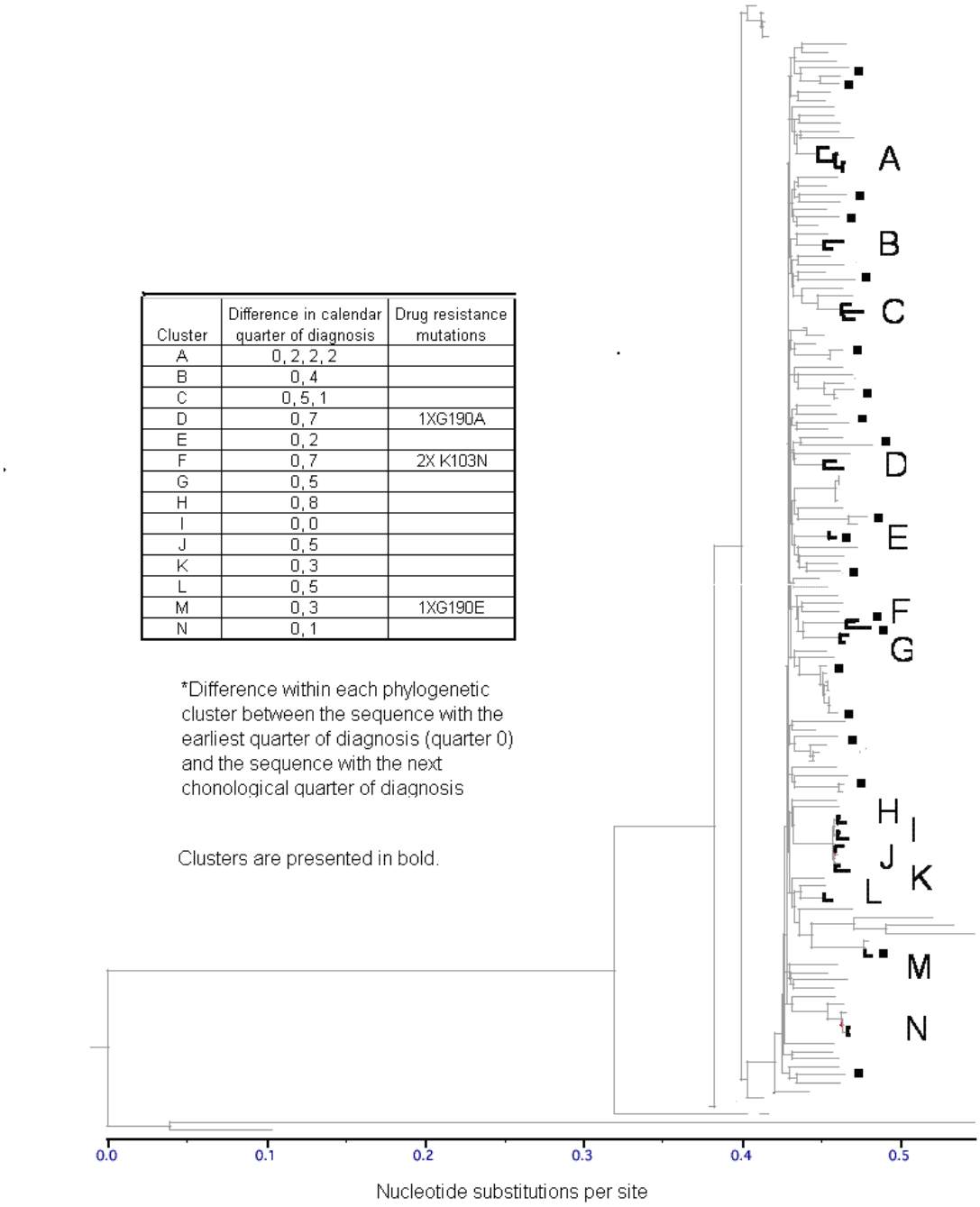


Possible transmission events[1]	Difference in quarter of attendance[2]	Drug resistance mutation
A	0,3,11,1	4xK103N
B	0,0	
C	0,1	
D	0,1	
E	0,2	2xT215V/D 1XF77L
F	0,0	
G	0,9	

[1]Not to scale

[2]Within phylogenetic clusters, the difference between the quarter of attendances between patients from whom the sequences were obtained. The earliest quarter is presented as "0", and the next chronologically, by quarter count. For example, a phylogenetic cluster of three sequences from patients with quarters of attendance of "2001 Q2, 2001 Q4 and 2003 Q3" would be presented "0,2,9"

Figure 7.7: Phylogenetic reconstruction of HIV transmission events and distribution of drug resistance mutations between MSM recently HIV-infected at diagnosis, Brighton: 2000-2006

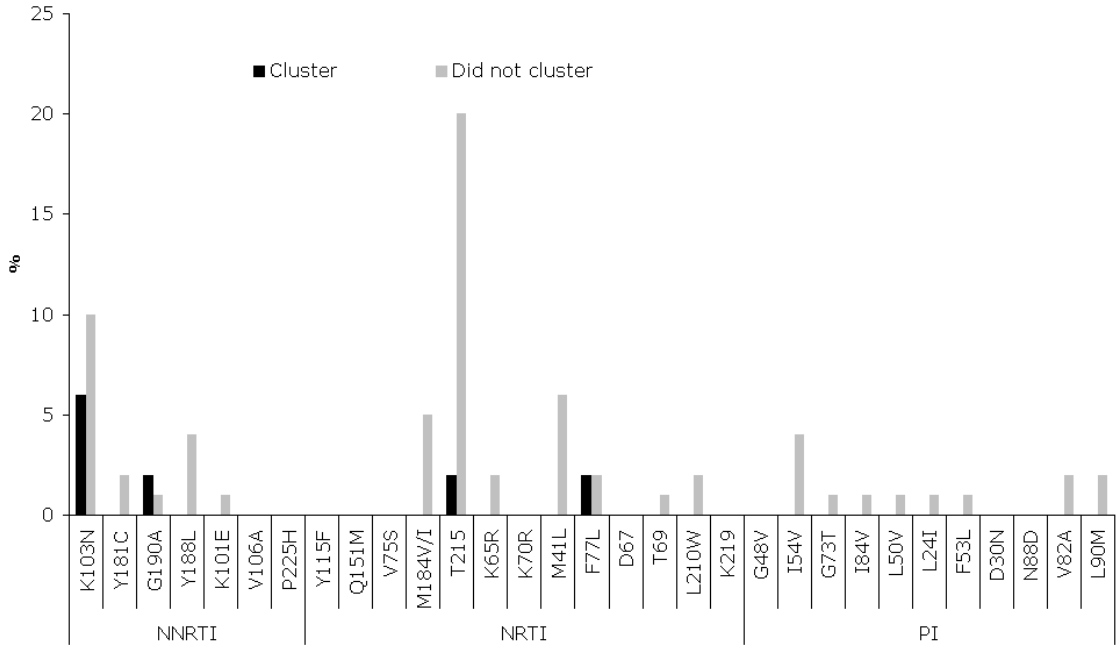


Cluster	Difference in calendar quarter of diagnosis	Drug resistance mutations
A	0, 2, 2, 2	
B	0, 4	
C	0, 5, 1	
D	0, 7	1XG190A
E	0, 2	
F	0, 7	2X K103N
G	0, 5	
H	0, 8	
I	0, 0	
J	0, 5	
K	0, 3	
L	0, 5	
M	0, 3	1XG190E
N	0, 1	

*Difference within each phylogenetic cluster between the sequence with the earliest quarter of diagnosis (quarter 0) and the sequence with the next chonological quarter of diagnosis

Clusters are presented in bold.

Figure 7.8: Distribution of specific mutations between clustering and non clustering sequences from patients recently HIV-infected at diagnosis, CASCADE, UA survey and Brighton; multiple years



7.3.3 Attributes and transmission potential of patients with transmitted and acquired drug resistance

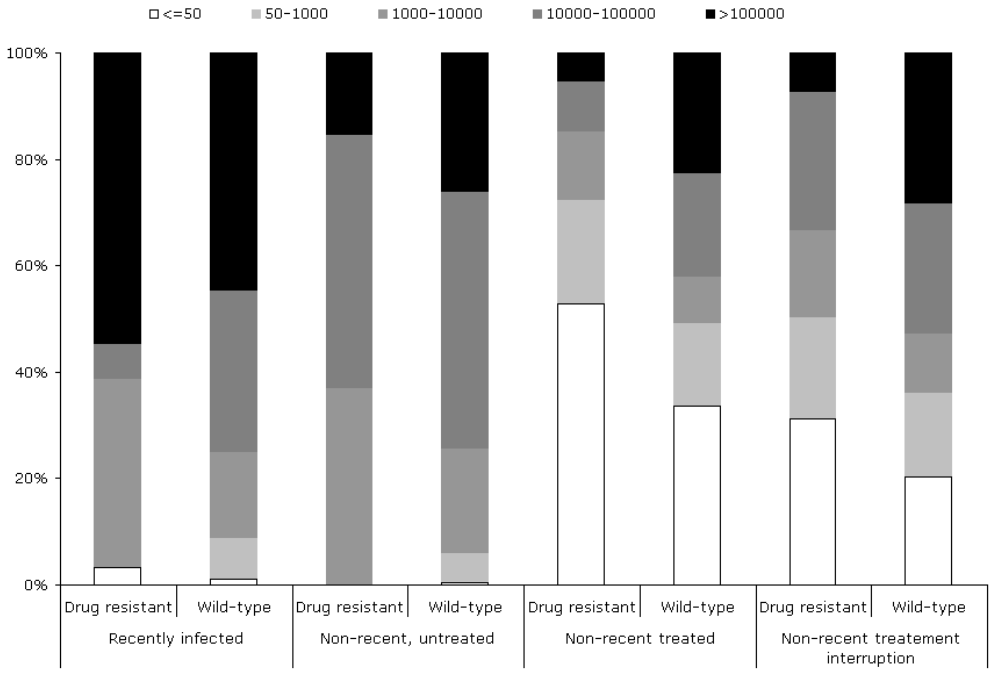
Using all calendar quarters of the study period, the viral load and STI distribution were compared between the patients with drug resistance mutations, and those with wild-type sequences, by infection category.

With the exception of those currently untreated, there was no statistical difference in the mean viral load between patients whose sequences had drug resistance mutations and those with wild-type sequences **(Table 7.1**). The distribution of viral load by infection category is summarized in **Figure 7.9**.

Table 7.1: Mean viral load (copies per mL) by infection category, between drug resistant and wild-type sequences from diagnosed HIV-infected MSM, Brighton: 2000-2006

Infection category	Drug resistant	Wild-type	p-value (t-test)
Recent	235883	209928	p=0.8
Chronic, untreated	62961	78642	p=0.02
Chronic, treated	14643	22624	p=0.1
Chronic, treatment interruption	58408	50643	p=0.2

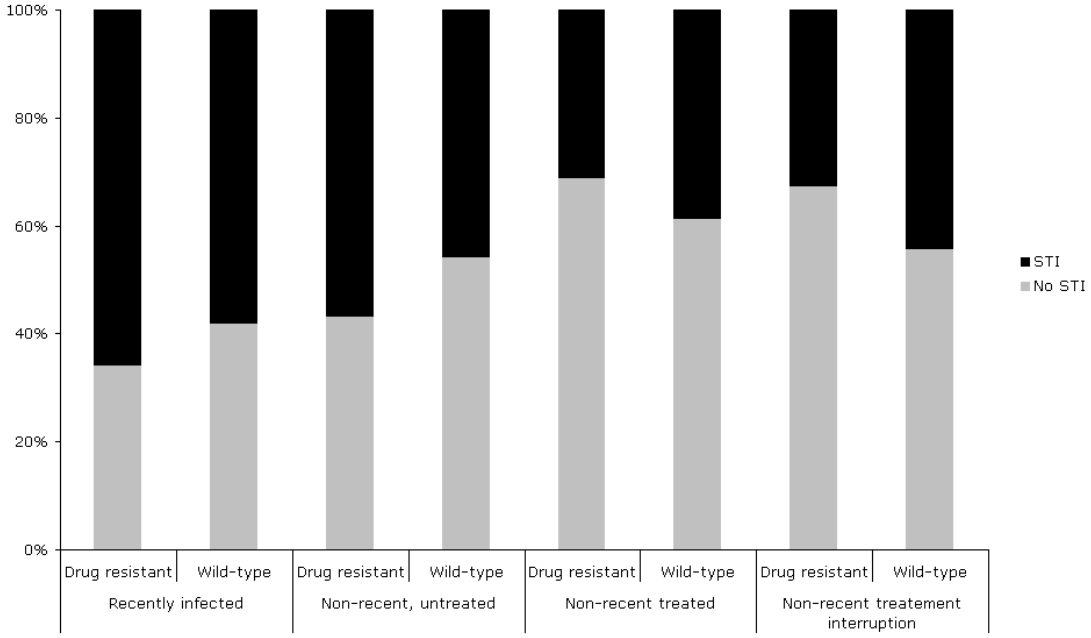
Figure 7.9: Viral load distribution by infection category between drug resistant and wild-type sequences from diagnosed HIV-infected MSM, Brighton: 2000-2006



Including every calendar quarter of the study period, (approach three, table 6.2) overall, 34% (1521/4428) calendar quarters were linked to patients with drug resistance mutations who also had an STI diagnosis compared to 43% (5847/13750) (p<0.0001) of quarters linked to patients with wild-type sequences. Among the recently HIV-infected the proportions were 65.9% (29/44) and 58.2% (181/311) respectively (p=0.17), and among the chronically infected, 56% (235/413) and 45.8% (2067/4510) (p<0.0001) respectively. The inverse was true for the treated population; those with wild-type were more

likely to have STIs. Of patient calendar quarters with wild-type viruses, 38.7% (2500/6457) had an STI diagnosis compared with 31.3% (943/3012) ($p<0.0001$) of those with drug resistance mutations **(Figure 7.10).**

Figure 7.10: STI distribution by infection and drug resistance category among diagnosed HIV-infected MSM: Brighton, 2000-2006



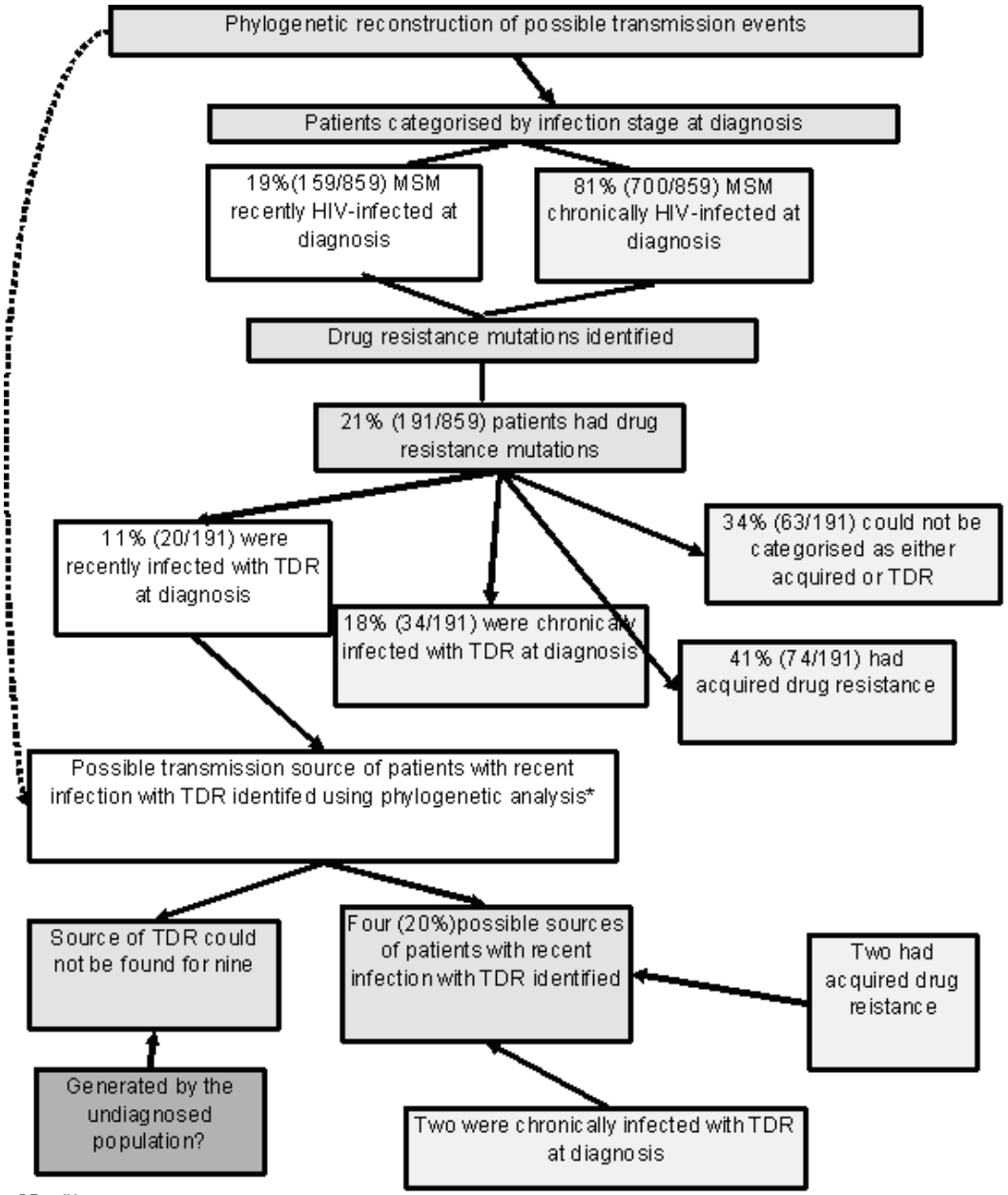
7.3.4 Where does TDR come from?

Of the 20 patients who were recently HIV-infected at diagnosis, and who had drug resistance mutations, the most likely transmission source was identified in four cases (20%) **(Figure 7.11)**. The source of transmission could not be found for nine. For the remaining seven, the direction of transmission could not be ascertained, or data was not available during the calendar quarter of transmission. Two of the most likely transmitters had chronic TDR (A and B) and two had acquired resistance (C and D). The mutation sites were identical

between the transmission sources and the patients with recently acquired TDR **(Figure 7.12).**

Figure 7.13 describes the attributes of the four transmission sources at the time of transmission**.** The untreated transmission sources had viral loads of >20,000 copies/mL (transmitters A and B) and the transmitters with acquired resistance had interrupted therapy with viral loads of >550,000 copies/mL (transmitters C and D). The most likely transmitter could not be identified for nine (69%).

Figure 7.11: Flow diagram of ascertainment of transmission sources of patients recently HIV-infected with TDR HIV strain



* Possible source:
- had to form a robust cluster with the sequence from the patient who was recently HIV-infected at diagnosis (99% bootstrap support and genetic distance <0.015 nucleotide substitutions per site)
- where there was more than one candidate transmitter, the candidate with the shortest genetic distance was chosen as the most likely transmitter
- candidate transmitters had to be diagnosed BEFORE the patient with recent HIV infection (or the same calendar quarter if the transmitter was chronically infected at diagnosis), or the direction of transmission could not be ascertained

Figure 7.12: Phylogenetic reconstruction of possible HIV transmission events that generated patients recently HIV-infected with TDR HIV strain, Brighton: 2000-2006

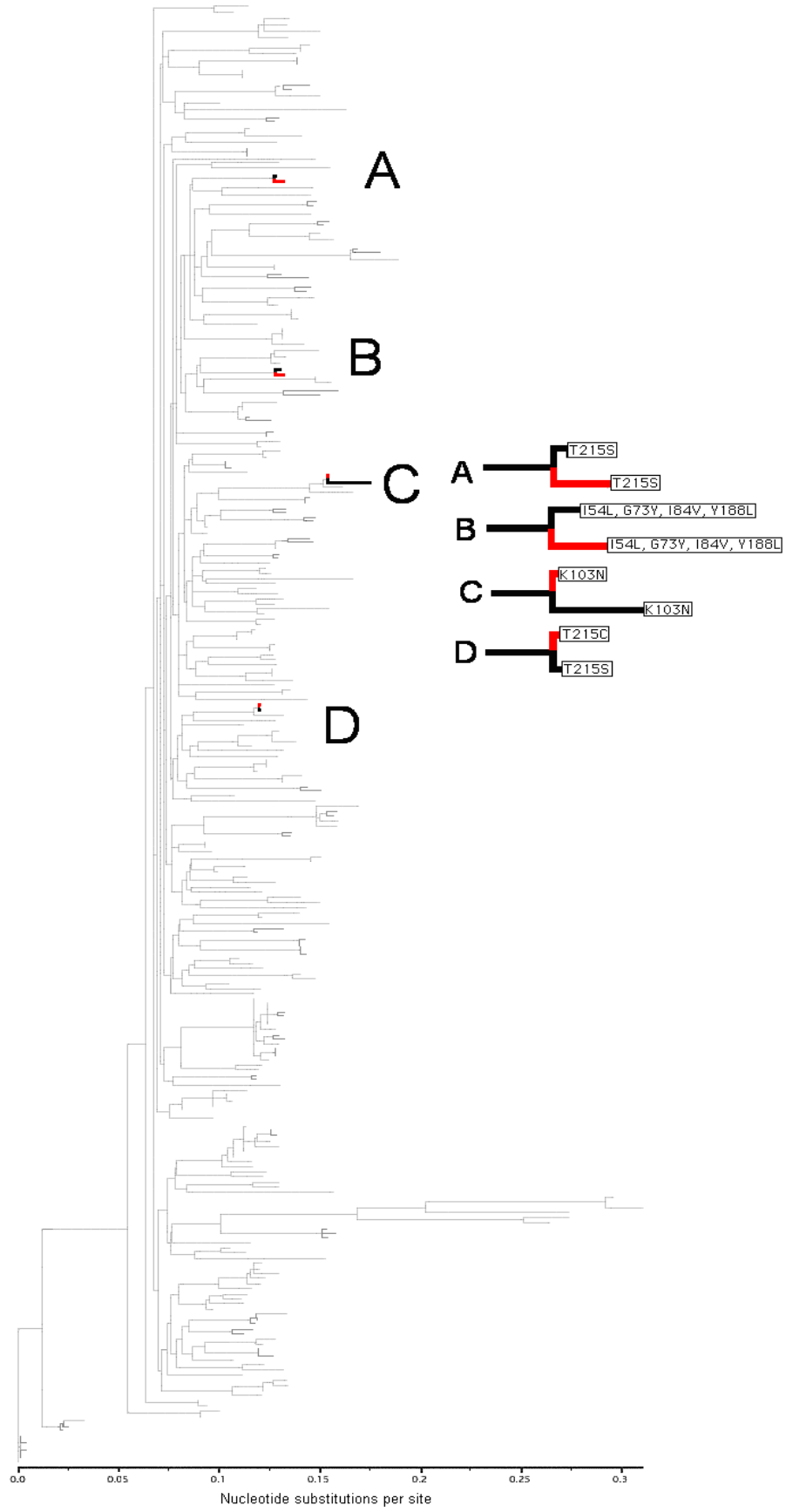
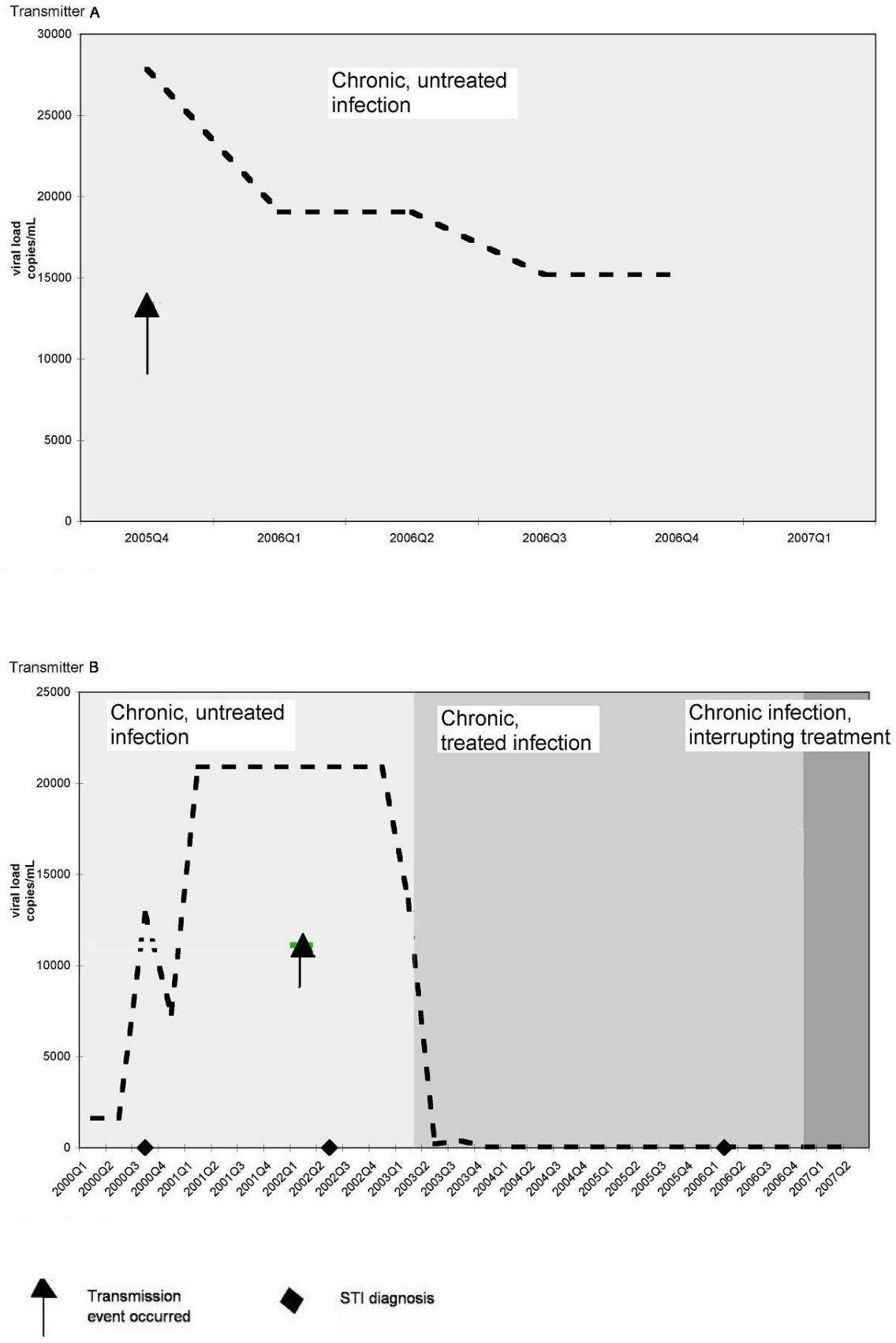
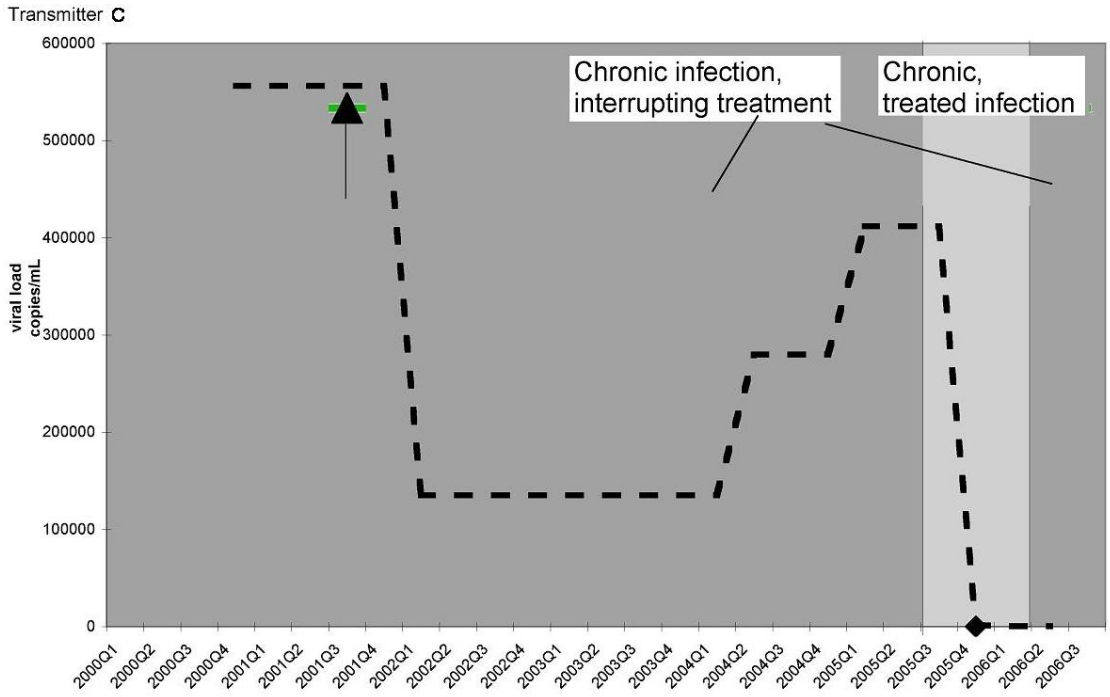
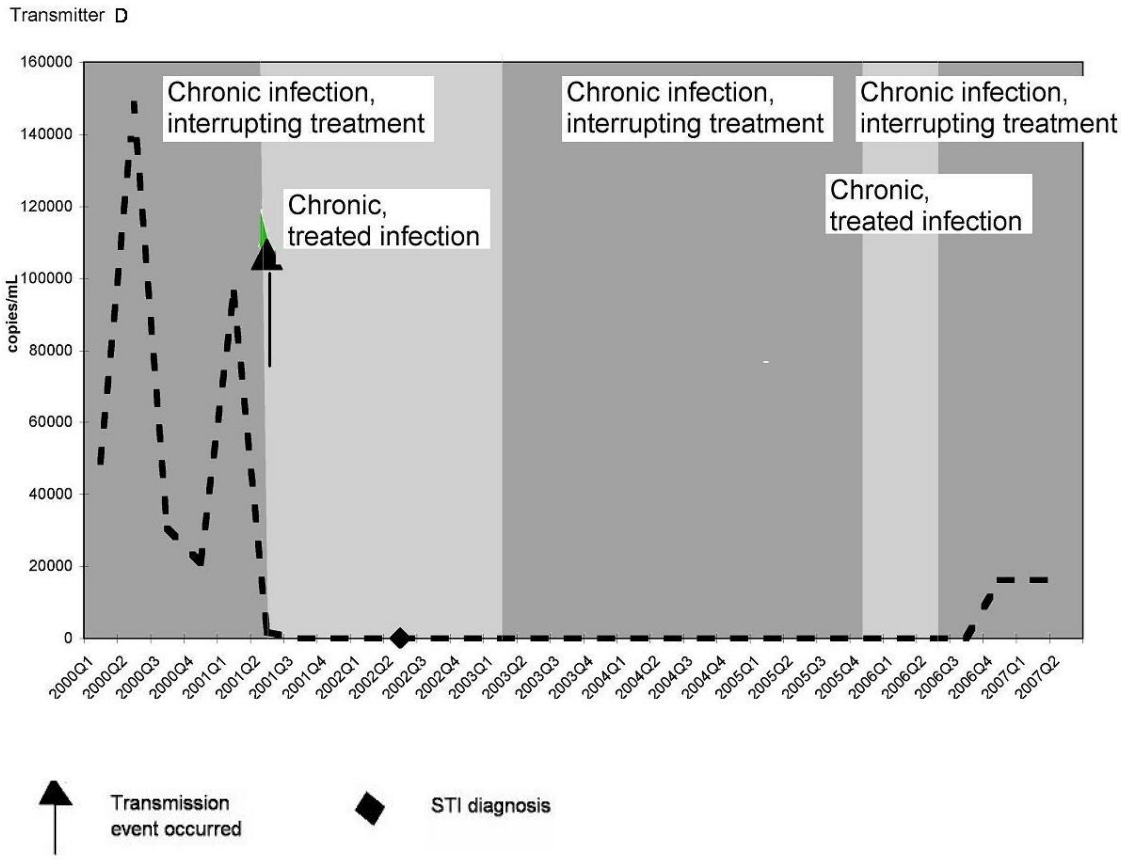


Figure 7.13: Attributes of transmission sources of TDR at around the time of transmission




Transmitter C
viral load copies/mL
600000
500000
400000
300000
200000
100000
0
Chronic infection, interrupting treatment
Chronic, treated infection
2000Q1
2000Q2
2000Q3
2000Q4
2001Q1
2001Q2
2001Q3
2001Q4
2002Q1
2002Q2
2002Q3
2002Q4
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3


Transmitter D
copies/mL
160000
140000
120000
100000
80000
60000
40000
20000
0
Chronic infection, interrupting treatment
Chronic, treated infection
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic, treated infection
2000Q1
2000Q2
2000Q3
2000Q4
2001Q1
2001Q2
2001Q3
2001Q4
2002Q1
2002Q2
2002Q3
2002Q4
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2

Transmission event occurred
STI diagnosis

7.4 Discussion

This chapter examined the level of TDR within the available datasets, and described the distribution and transmission of specific mutations between patients recently HIV-infected at diagnosis. It described the distribution of specific drug resistance mutations between datasets, and for the Brighton dataset, between infection categories. The distribution of viral load and STI diagnoses was compared between infection categories. It also aimed to ascertain the sources of TDR and their attributes at around the time of transmission.

7.4.1 Prevalence of HIV drug resistance

Transmitted drug resistance

The prevalence of TDR mutations among the recently HIV-infected at diagnosis was measured at: 10.9% for CASCADE, 14.2% for the UA STI survey and 12.6% among the (recently HIV-infected) Brighton dataset. These are broadly consistent with the literature for the UK (UKCollaborativeGroup 2007) and Western Europe (Wensing, van de Vijver et al. 2005).

The variation in the prevalence of TDR between the three datasets is likely to be due to three factors. Firstly the definitions used to gauge recent infections. While the UA STI survey used the serological testing algorithm for recent HIV seroconversion (STARHS) only, Brighton and CASCADE used more than one algorithm to identify recent infection, some of which have shorter window periods (section 3.6.4). Consequently Brighton and CASCADE may be more likely to identify drug resistant mutations.

Secondly, the period of time from which the sequences were taken may explain the variation between the three datasets. The UK data suggest that prevalence of TDR peaked at 14% in 2001/2 and decreased to 8% in 2004 (UKCollaborativeGroup 2007). This reduction was largely due to a decline in NRTI mutations, while NNRTI mutations increased in prevalence. Data from the UA survey and CASCADE date from earlier years. Brighton data extends from 2000-2006 and also showed a higher prevalence of NNRTI mutations. The small samples sizes precluded the presentation of prevalence over time for each dataset.

Finally, the geographic range from which the sequences were taken is likely to affect the prevalence. Specifically, CASCADE constitutes a broad range of countries and the sequences were taken during a time period spanning over a decade: each country represented is likely to have different underlying transmission networks. While the UA STI survey is taken from a narrower time period, it is less likely to include transmission events since it is taken from several national sites.

Differences between transmitted and acquired drug resistance

Using the entire Brighton dataset, the overall prevalence of drug resistance mutations (transmitted or acquired) was 22.2%. This is substantially higher than that found among sequences from MSM who were recently HIV-infected at diagnosis. This is expected since the overall Brighton population will contain patients with acquired resistance. Among the treated Brighton population, 36% had drug resistance mutations. Among the drug naïve population, the prevalence of drug resistance mutations was higher among those recently HIV-

infected at diagnosis compared with those with a chronic infection at diagnosis. While this difference may be explained by the relative weaker fitness of drug resistance mutations (Devereux, Youle et al. 1999) other considerations related to the timing of HIV testing require exploration. Those who were diagnosed during recent infection may have been prompted to come forward for HIV testing either through recent risk exposure or seroconversion illness. Higher risk sexual mixing (some HIV diagnosed MSM with sub-optimal treatment adherence may also be more likely to have risky sex) and any clinical effects of the specific drug resistant viruses respectively could both be independently associated with being infected with drug resistant strains.

While sexual factors and symptoms possibly associated with resistant viruses may be a factor, the uneven distribution of specific classes of drug resistance mutations between the recently and chronically HIV-infected population suggests the relative fitness of specific mutations may have a major role. This work demonstrates that, within one population, there is a difference in the specific mutations found depending on the time the sequence was taken relative to infection. This adds further evidence of the potential for under-detection of specific drug resistance mutations in the chronic population. Consequently the use of deep sequencing or other methods to detect minority species should be considered for MSM who are not recently HIV-infected at diagnosis (Johnson, Li et al. 2008).

Distribution of specific mutations

The distribution of specific mutations varied between infection categories. Mutations V106A, K219E and V82A were not observed in the drug naïve

populations, only in the treated. Conversely, M41L, G73T and F77L were observed among the drug naïve but not among the treated populations.

This analysis demonstrates that the likelihood of identifying drug resistant mutations is affected by the specific mutations under investigation. It indicates that there is a difference in the TDR prevalence between infection categories but also that this may be directional to specific mutations. The results remain exploratory and need to be interpreted with caution due to small sample sizes.

7.4.2 Phylogenetic reconstructions

Using phylogenetic reconstructions comprised of sequences from patients who were diagnosed during recent infection, the extent of clustering between sequences from patients with drug resistance mutations was fairly low. For CASCADE, 11% formed a cluster, for UA STI 33.3% formed a cluster and from Brighton (recent) 20% formed a cluster. The sequences with drug resistance mutations that clustered were limited to K103N, F77L, and G190 and T215 variants. All of them shared identical mutations within clusters.

Including all three datasets, the formation of a robust cluster that consisted of a sequence with a drug resistance mutation, and a wild-type sequence, occurred only three times. Two cases occurred within the Brighton dataset and involved the G190 site and one occurred within the UA STI survey involving the F77L mutation. The G190 mutation has been known to revert after infection (Bezemer, de Ronde et al. 2006). If the phylogenetic reconstruction did not reflect reality, by chance, it would be expected that a higher proportion of

clusters would be formed between wild-type and drug resistant mutations. Consequently this adds some support to the reconstruction.

Comparing the estimated infection dates within clusters demonstrates whether transmission could have occurred during recent infection (section 5.2.3). The one instance of a cluster containing sequences with an identical drug resistance mutation, identified in CASCADE could not have been generated during recent infection because there was a difference of 226 days between the infection dates). Cluster A could have been generated during recent infection for the UA survey (difference of one calendar quarter), but cluster E could not (difference of three calendar quarters). For the Brighton dataset (the recently HIV-infected only) no transmission could have been generated during recent infection.

With the exception of the UA survey, there was no significant difference between the proportion of phylogenetic clusters formed between sequences with mutations compared with clusters formed between wild-type sequences. This suggests that patients with drug resistant viruses may not be any more likely to pass on their infection compared to those infected with wild-type sequences. This is despite those with TDR mutations being more likely to have been undiagnosed and more likely to be experiencing recent HIV infection.

7.4.3 Attributes and transmission potential of patients with transmitted and acquired drug resistance

With the exception of the untreated, chronic population, there was no significant difference in mean viral load between Brighton patients with drug resistant viruses and those with wild-type viruses, by each category. Untreated patients

with drug resistant mutations may be less likely to transmit their infection onwards compared to the untreated wild-type population.

Patients with drug resistance mutations were less likely to have an STI diagnosis during the study period compared to those with wild-type sequences ($p<0.0001$). This may be because those with mutations are disproportionately the treated population who may have been attending treatment and care for long period of time and may have therefore modified their sexual behaviour. However, among the recently HIV-infected, those with drug resistance mutations may have been more likely to have STIs compared to those with wild-type sequences. This occurrence may be associated with risk behaviours: those with TDR may have acquired their infection through sex between men with acquired resistance. This may suggest that HIV-infected MSM with sub-optimal ARV adherence may be more “risky” in relation to treatment and sex.

7.4.4 Where does TDR come from?

Identifying the transmission source of patients with recently acquired TDR allows the direction and approximate date of transmission to be ascertained. This in turn allows the infection stage and attributes of the likely transmission source to be ascertained at around the time of infection (section 5.2.3).

The identified transmission sources of patients with recently acquired TDR were untreated individuals, and those going through a period of treatment interruption. The transmission source could not be identified for the majority of patients with recently acquired TDR, indicating that a substantial proportion may have been derived from MSM with undiagnosed infection. However, the

unidentified transmission sources of the patients with TDR may also have come from diagnosed MSM with no sequence available, or MSM attending clinical care elsewhere. Further limitations include the small sample size (making it difficult to extrapolate to the wider population) and the inability of the methods employed to detect minority species.

7.5 Conclusion

The prevalence estimates of TDR mutations found in this chapter are broadly consistent with other studies. The chapter also lends further evidence that the relative fitness of drug resistance mutations, and their likelihood to be transmitted, is differential to specific mutations. The refined definition of the recently HIV-infected population and drug resistant mutations has allowed improvements on the methods used in previous analyses.

This chapter suggests that patients with viruses with drug resistance mutations do not seem any more likely to transmit their infection onwards compared with those with wild-type viruses. However, those with TDR may represent riskier populations in terms of acquiring STIs. The exploratory analysis suggests high viral load as a risk factor for the transmission, including the transmission of resistant strains. It also implicates individuals with undiagnosed infection as an important source of TDR. Our data therefore support earlier HIV testing and initiation of ARV as public health measures.

8 Chapter Eight: Discussion and conclusions

This chapter describes how the thesis contributes to research and links with the current literature. It also outlines the thesis limitations and describes the future research needed in this field. Finally, the implications for policy and practice are discussed.

8.1 Contributions to research

The objective of the thesis was to enhance understanding of HIV transmission between MSM through combining HIV *pol* sequences with laboratory, clinical and demographic information. The specific contributions that the thesis has made to research are divided into two categories: increased understanding of factors associated with HIV transmission and methodological developments.

The thesis has contributed several findings related to HIV transmission that are of public health importance. Chapter four suggested that while the consistency of phylogenetic reconstructions of HIV transmission events was broadly good, the need to interpret results with caution remains, as results are at least partially dependent on the sample size and heterogeneity. The phylogenetic reconstructions of transmission events from MSM diagnosed during recent infection, as described in chapter five confirmed the recently HIV-infected as an important group in generating HIV transmission. Improved methods in chapter six demonstrated that the recently HIV-infected are disproportionately generating onward transmission, with a rate ratio of 3.04 (compared with 1 from the untreated, chronically HIV-infected). However, the majority of transmissions are generated by the currently untreated population – regardless of infection category – with the risk of transmission from the treated population being extremely low. Importantly, 70% of transmissions were generated from patients with CD4 counts over 350 cells/mm^3. Chapter seven provided further evidence that specific drug resistance mutations are different with respect to transmission risk and patients with drug resistant mutations may represent riskier populations in terms of acquiring STIs. The chapter also found

that transmitted drug resistance (TDR) may be disproportionately generated by the undiagnosed population.

The methodological improvements can be subdivided into three. Firstly, a method was developed to assess the consistency of phylogenetic reconstructions of transmission events for public health purposes (chapter four). Secondly, chapter five demonstrated that previous phylogenetic studies may have overestimated the transmission risk from the recently HIV-infected through failing to recognize that “recent” infection is a transient attribute: a phylogenetic relationship between sequences from two patients recently HIV-infected at diagnosis does not mean both patients were recently HIV-infected at the time of transmission (Brown 2009a). Thirdly, on this basis, an improved method for ascertaining risk factors associated with transmission events through phylogenetic reconstructions (including recent infection) was developed in chapter six. This method can be applied to factors that may vary over the course of an individual’s infection (e.g. STI, viral load, and ARV treatment). This method was applied to study the transmission of drug resistant viruses in chapter seven.

8.2 How the research fits into the current literature

The phylogenetic reconstructions presented in this thesis are compared to other phylogenetic reconstructions of HIV transmission events. The specific findings from the reconstructions will then be discussed in the context of other relevant literature.

The thesis has been written at a time when an increasing number of phylogenetic reconstructions of HIV transmission events are being published (Brenner, Roger et al. 2007; Lewis, Hughes et al. 2008; Paraskevis, Pybus et al. 2009; Yerly, Junier et al. 2009). All use anonymous datasets so that it is impossible to identify any individual involved in a transmission event. Such reconstructions can be subdivided according to their methods. Firstly those (like that presented in this thesis) that use maximum likelihood methods to reconstruct transmission events (Pao, Fisher et al. 2005; Brenner, Roger et al. 2007), link such events to the patients' clinical and demographic information (usually data on recent infection at diagnosis) and explore associations between transmission events and the specific risk factors. As previously discussed, our methods are an improvement on previous methods since they take account of the fact that infection stages (and other risk factors) are transient and change over the course of an HIV infection.

The second type of phylogenetic reconstructions use Bayesian methods (Drummond and Rambaut 2007) that permit possible transmission events to be "dated" through calibrating nucleotide substitution rates against known HIV transmission events (Drummond, Ho et al. 2006) using software such as BEAST (Bayesian evolutionary analysis by sampling trees) (Drummond and Rambaut 2007). Using these methods, the dates that certain lineages appeared, or the time intervals between transmission events can be estimated (Lewis, Hughes et al. 2008).

Through such methods, Lewis *et al*. estimated that 25% of transmissions occurred at intervals of under six months of each other, and interpreted this as evidence of the recently HIV-infected driving transmission (Lewis, Hughes et al. 2008). While the methods are not the same as those used in this thesis, it is encouraging that the results are broadly consistent.

Following on from the Lewis *et al* paper, Hughes *et al.* presented a paper in CROI 2009 of a phylogenetic reconstruction of HIV transmission events among heterosexuals within the UK. Using BEAST they calculated the intertransmission intervals and found them to be substantially longer than those identified among MSM (Lewis, Hughes et al. 2008). From this they concluded that recent infection was not as an important driver of transmission among the heterosexual population as it is among the MSM population, and that less transmission occurred within this population. However, rather than representing transmission from the chronically infected, it may be that the longer time intervals are due to the reduced probability of identifying transmission events among heterosexual populations; sampled MSM are more likely to be drawn from the same transmission networks, since more transmission occurs within UK for this risk group. However, this finding is consistent with the lower rates of sexual partner change found among heterosexual populations (Fenton, Mercer et al. 2005) compared with the homosexual male population (Dodds, Mercey et al. 2004). The inclusion of epidemiological and clinical data through this approach could have assisted with interpretation.

A disadvantage of the Bayesian/BEAST approach is its inability to link transmission events to the attributes of the patient at around the time of transmission. This is because it can ascertain ancestral transmission events (with no representation in the dataset) or reconstruct events that occurred while patients were undiagnosed (and consequently no information is known at the time of transmission). The only factors that can be used to assist with interpretation are the time intervals between transmissions and the sources from which the sequences were derived.

The thesis confirmed recent HIV infection as a risk factor associated with an increased transmission risk; the reasons for this have been well documented (Gray, Wawer et al. 2001; Wawer, Gray et al. 2005). Through phylogenetic methods, it is estimated that approximately a quarter of transmissions were generated by MSM with recent infection; this is lower than the half estimated by Brenner *et al*. (Brenner, Roger et al. 2007). This is partially explained through Brenner *et al*. interpreting transmission events between patients recently infected at diagnosis as “recent to recent transmissions”, even though the transmission events may have occurred during chronic infection (Brown 2009a).

Secondly, the work presented found that a substantial proportion of transmission is coming from the currently untreated, diagnosed population. It is known that the diagnosed HIV-infected population continues to have unprotected anal intercourse (Dodds, Mercey et al. 2004; Elford and Hart 2005). This is to a greater extent with regular partners and to a lesser extent with casual partners (Elford 2009 ISTTDR). Additionally, those on treatment may

have unprotected sex due to a perceived reduced infectiousness (Crepaz, Hart et al. 2004; Elford and Hart 2005).

The third main finding is that treatment reduces the risk of transmission. It is well established that HIV-infected patients who have been treated with, and adhere to combinations of treatment typically achieve a suppressed viral load (Cu-Uvin, Caliendo et al. 2000; Vernazza, Troiani et al. 2000). In studies of serodiscordant couples (where researchers study transmission within monogamous couples with different HIV serostatuses) treatment has also been shown to substantially reduce transmission (section 1.1.10). The current literature recognizes that treatment greatly reduces transmission, but cannot state definitively that it eliminates the risk of transmission. There have been instances where vertical transmission has occurred where patients have undergone treatment – however all mothers that transmitted had advanced HIV disease (EuropeanCollaborativeStudy 2005). It has also been noted that blood viral load is not synonymous with genital fluid viral load. Specifically it has been found that among individuals who achieve undetectable blood viral load through ARV, viral shedding has been observed in genital fluids (Zhang, Dornadula et al. 1998; Neely, Benning et al. 2007). Plasma viral load only reflects the level of cell-free virus in the blood. Latently infected cells might transmit infection in the absence of virus. Consequently, using viral load as a definitive marker for infectivity may be inaccurate (Adelisa L. Panlilio 2005). Finally, patients fully adherent to ARVs may experience transient low-level viral rebound (blips) (Garcia-Gasco, Maida et al. 2008). The transmission risk from such blips is not clear.

8.3 Limitations

There are several limitations to the thesis. These relate to: the datasets used; the phylogenetic methods employed; the techniques used to link infection stage and clinical data; and the lack of behavioural data. These will be taken in turn.

The selection of the specific dataset for use in phylogenetic reconstructions will influence the number of transmission events that potentially can be identified. There are additional biases to consider. For instance, every sequence included will have been derived from a patient with a diagnosed HIV infection. Consequently the undiagnosed population is difficult to include (unless samples are obtained from Unlinked Anonymous testing). Furthermore, the sequences that are selected may not be representative of the HIV-infected population: they may over-represent the recently HIV-infected. This population may have been prompted to come forward due to a recent risk exposure, or have symptoms of seroconversion illness (Burchell, Calzavara et al. 2003). They may differ in important ways to the HIV-infected population as a whole. Chapter four suggests that sample heterogeneity might affect the accuracy of reconstructions, suggesting the extent of sexual mixing within one geographic locality is likely to be an important factor.

The limitations and concerns with phylogenetic reconstructions were examined in detail in chapter four. However, there are additional problems. The technique used only considers pair-wise relationships; there have been no attempts to look at alternative methods of interpreting larger clusters. The size of the sample available is also a limitation to phylogenetic reconstructions. It is

difficult to predict a sample size that would generate an interpretable number of phylogenetic reconstructions. Additionally, phylogenetic reconstructions that use a larger number of sequences are limited by the extensive computational power that would be needed to undertake such analyses.

The thesis exclusively studies *pol* sequences from MSM. Firstly, this is because, historically, this population has tended to have subtype B virus. Arguably this increases the likelihood of finding transmission events within a population sample of sequences, since circulating viruses are less likely to contain viruses acquired abroad. Secondly, the analyses frequently include infection stage, derived from laboratory algorithms. These have been developed and validated with subtype B (Janssen, Satten et al. 1998), however use of the avidity assay means the serological testing algorithm for recent HIV seroconversion (STARHS) can now be applied to heterosexual populations (Suligoi, Butto et al. 2008). With approximately half of the HIV-infected population comprising heterosexuals, many of whom are infected with non-B virus, there is a need to see if the results are also applicable to this population.

The greatest limitation of the thesis is the lack of behavioural data. Sexual risk behaviour is likely to have a vital role in mitigating the effect of patient infectivity; transmission cannot occur without an exposure event. Behaviour is known to vary between diagnosed HIV positive, undiagnosed HIV positive and HIV negative men, and also with respect to whether partners are casual or regular (Dodds, Mercey et al. 2004). Elford *et al.* (ISSTDR 2009) found that UAI was more likely to occur between MSM in stable relationships. This has important

implications both in terms of interpretation of phylogenetic reconstruction events, and also for prevention messages. Markers of sexual behaviour linkable to HIV sequences are needed to assist with the interpretation of transmission rates by infection category. Patterns of sexual mixing would also facilitate interpretation (e.g. gauging the extent that sexual mixing is random).

8.4 Future research

Two main areas of further research are needed. Chapter six indicated that the recently HIV-infected are disproportionately likely to generate infection, and that transmission is unlikely to occur from the treated population with suppressed viral load. It is important to determine whether these results can be replicated in other settings, or whether the results are applicable to Brighton MSM only. Investigation is required among heterosexual populations in particular; this is now possible since avidity assays used in conjunction with STARHS, enables recent infection to be consistently identified regardless of subtype (Suligoi, Butto et al. 2008). The Health Protection Agency is rolling out incidence tests for every new HIV diagnosis made in the UK. This will increase the availability of infection category data linkable to HIV *pol* sequences nationally. Further analyses should also attempt to integrate behavioural data to the sequences from patients to assist with interpretation. The integration of sexual behaviour could help in three ways: patterns of sexual behaviour by infection category would help adjust the transmission risks from each category; linking markers of risky sexual behaviour to patients from whom the sample was obtained would help to ascertain the extent to which infectivity, and the frequency of exposure events, are drivers of transmission; and the inclusion of sequences from patients with known transmission histories to inform the phylogenetic accuracy.

It is also important that further work is undertaken to develop phylogenetic methods that can be applied appropriately for public health purposes. Further work is needed to improve definitions of what constitutes a definite transmission event (particularly with regard to heterogeneity). Work is needed to generate some measure of the degree of error that can be expected from population level phylogenetic reconstructions.

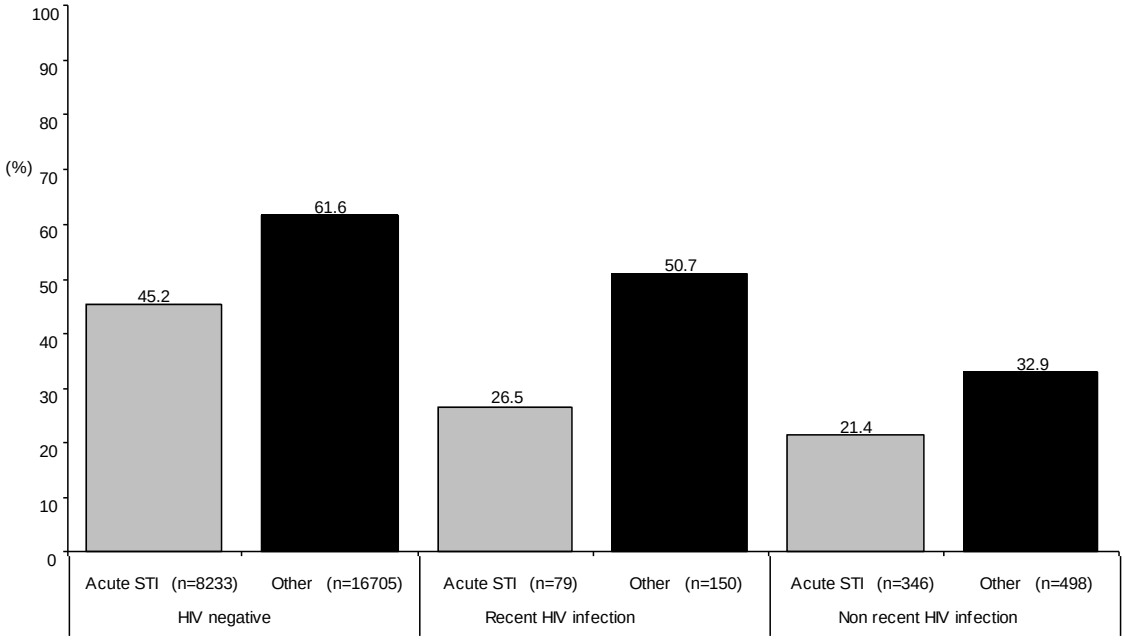
8.5 Implications for policy and practice

There are two main implications for policy and practice: the need for the recently HIV-infected to be targeted for testing; and the possibility that wide-spread treatment for HIV could be used to prevent transmission from the HIV diagnosed population.

The obvious solution to the disproportionate level of transmission arising from the recently HIV-infected is to encourage testing in this population. However, this is not necessarily easy to achieve. Work undertaken using data from the Unlinked Anonymous (UA) sexually transmitted infection (STI) survey, but not presented in the main part of the thesis, looked at the rate of HIV testing by HIV status, by STI presence and, for the HIV-infected, and whether the infection was recently or non recently acquired. Between 1999-2002 more than half of recently HIV-infected MSM and nearly 80% of all undiagnosed HIV-infected MSM with an STI did not receive HIV tests (Brown, Murphy et al. 2009b) **(Figure 8.1**). The rate of HIV test uptake has since increased from 45% in 1996 to 86% in 2007. However, in the same year it was measured at 42% among the HIV-infected with an STI and 63% among the recently HIV-infected. Therefore despite an increased voluntary confidential HIV testing (VCT) uptake

generally, a substantial proportion of those at highest risk of onward transmission left the clinic remaining undiagnosed.

Figure 8.1: Proportion of MSM attending sentinel STI clinics survey receiving voluntary confidential HIV testing, by diagnosis status, UA survey:1999-2002.



Source: (Brown, Murphy et al. 2009b)

The reason for poor VCT uptake among the recently HIV-infected may have been due to a practice exemplified in older BASHH guidelines (BHIVA 2006) that suggested that recently HIV-exposed patients receive two HIV tests: one immediately and one 3 months after exposure (the “three month rule”). This is because of concerns that HIV testing shortly after exposure may not yield accurate results. However, recent improvements in HIV test sensitivity mean that the latest (so-called ‘fourth generation” tests) can detect both anti-HIV and HIV p24 antigen within four weeks of infection (Busch, Glynn et al. 2005). However, delaying the second test for three months may be detrimental; it will include a subset of MSM with recent infection remaining unaware of their infection at a time when they are highly infectious. A BASHH audit found that

the most commonly cited reason for not receiving VCT was deferral because of concerns relating to the accuracy of HIV testing shortly after exposure (Munro 2007). The rate of reattendance for the second test was low. The BASHH guidelines have since been updated to include every patient attending, including the targeting of those with recent risk exposure (Gazzard 2008).

This suggests that VCT may be limited in its ability to prevent HIV transmission. An alternative solution may be to target those with seroconversion illness presenting at settings outside of STI clinics. This has been found to be a feasible practice in Brighton, achieved through raising awareness of seroconversion illness symptoms (and its association with elevated infectivity) among health care workers and among MSM at risk in areas of high HIV prevalence (Sudarshi, Pao et al. 2008) (Fox, White et al. 2009).

Secondly, the finding that treatment is protective of transmission risk adds further weight to the argument that treating the entire diagnosed HIV-infected population could substantially impact on reducing transmission. In recent times, the possibility has been considered that wide-spread treatment could prevent HIV transmission from the diagnosed population. In 2008, the Swiss federation on HIV and AIDS published a controversial paper stating that HIV serodiscordant heterosexual partners may have unprotected sex, provided the infected partner is ARV treated with viral load <40 copies/mL (Vernazza 2008).

Table 8.1 presents viral load by treatment for MSM attending care services in the UK during 2005-7 (SOPHID, personal communication). The table

demonstrates the elevated viral load among the untreated population. It can be seen that in 2007, almost one third of the diagnosed HIV-infected MSM population were untreated, with a viral load over 10,000/mL; this level was found to be associated with an increased transmission risk in chapter six. Treatment of this population could therefore substantially impact upon transmission at the population level.

However, the Swiss report has been criticized as inconclusive and dangerous, jointly by the World Health Organization and UNAIDS (UNAIDS 2008), and the American Centers for Disease Control (CDC 2008). The mass treatment of the diagnosed population is controversial for several reasons. Firstly, it cannot be ascertained definitively that patients fully adherent to treatment will have their risk of transmission eliminated. Even if patients achieved undetectable viral load in their blood, it does not mean that their genital fluid will also have an undetectable viral load. For instance, a recent paper (Sheth, Kovacs et al. 2009) demonstrated that among 25 HAART treated HIV positive MSM, blood viral load was undetectable in all subjects by week 16; however, HIV RNA shedding was detected in semen among 12/25 subjects, and at a high level (>5000 copies/mL) in 4 of the 25. Secondly, even if treatment was 100% efficacious in preventing transmission, it cannot be guaranteed that patients will fully adhere to treatment. Thirdly, there are concerns about the mass treatment of patients, regardless of clinical need. This is because of the side effects for the patients, but also it increases the risk of acquiring drug resistance. Fourthly, mass treatment among the diagnosed population will ignore the transmission risk from the undiagnosed population, an elevated proportion of

whom will be experiencing recent HIV infection. However, if large-scale treatment is successful in reducing transmission, the proportion of the HIV-infected population who are recently HIV infected should diminish.

In late 2008, Wilson *et al*. (Wilson, Law et al. 2008) published an exercise based on modelling transmission risks at certain viral loads to ascertain the risk of transmission from HIV-infected patients over a prolonged period. Assuming monogamous serodiscordant relationship with 100 sexual encounters per year, for MSM they calculated a transmission risk of 0.043 per sex act. While this seems low, they point out that this translates as of 10,000 HIV negative MSM exposed to HIV each year, 3524 would become infected per year. In 2009, Granich *et al.* published a mathematical model that demonstrated that universal VCT in combination with immediate ARV could reduce HIV incidence within 10 years, and drive global prevalence to less than 1% within 50 years (Granich, Gilks et al. 2009). A recent meta-analysis of HIV transmission risk through unprotected sexual intercourse by ARV and/or viral load found that there are few studies that examine the effect of ARV and viral load (Attia, Egger et al. 2009) on transmission. However, on the limited data, it concluded that the overall HIV transmission rate in the presence of ARV was around 0.5 per 100 person years – substantially higher than that presented in this thesis **(Table 6.4).** While data are lacking, preliminary results show the risk of transmission from the treated population cannot be eliminated. A randomized control trial (START) has been launched by the US National institutes of Health to ascertain the magnitude and durability of the benefits of early treatment to prevent HIV

transmission between serodiscordant couples in practice (Cohen, Mastro et al. 2009; De Cock, Gilks et al. 2009).

Table 8.1: Viral load by treatment status among diagnosed HIV-infected MSM attending treatment services, UK: 2005-2007Table 8.2: Viral load by treatment status among diagnosed HIV-infected MSM attending treatment services, UK: 2005-2007

viral load/mL	ARV level	2005	2006	2007	Grand total
<50	None	163	229	343	1069
	Mono	25	64	78	333
	Dual	74	112	170	596
	Triple	430	678	886	3172
	Quadruple +	207	310	453	1589
	Not known	15	130	70	218
50-999	None	1115	1757	1881	6051
	Mono	61	107	95	361
	Dual	125	223	199	688
	Triple	1615	2430	2772	9093
	Quadruple +	774	1285	1271	4271
	Not known	27	252	60	339
1000-9999	None	2696	4507	4506	14666
	Mono	26	50	34	135
	Dual	29	55	54	168
	Triple	657	944	870	3101
	Quadruple +	306	445	369	1356
	Not known	34	415	148	598
10000-99999	None	4062	6315	6415	20514
	Mono	29	42	27	117
	Dual	33	81	62	206
	Triple	714	1020	1024	3452
	Quadruple +	288	431	383	1317
	Not known	59	713	232	1005
>100000	None	1256	1955	1738	5997
	Mono	10	18	20	60
	Dual	17	38	20	88
	Triple	425	609	564	1961
	Quadruple +	171	285	237	837
	Not known	23	199	96	320
Not known	None	201	153	168	750
	Mono	50	19	33	123
	Dual	69	31	70	198
	Triple	217	167	180	756
	Quadruple +	156	89	146	510
	Not known	58	53	84	206
Grand Total		16217	26211	25758	86221

Source: Personal communication, SOPHID

Whilst this debate needs to continue, it also needs to become more sophisticated. It is likely that the risk of transmission cannot be eliminated

through treatment, but this needs to be weighed against the very real reductions that could be achieved through wide-spread treatment, regardless of clinical need. Instead debate should focus upon what threshold of transmission risk from the treated population is acceptable. If such a threshold can be agreed upon, it is equally important to consider what level of monitoring of viral load and therapy adherence would be necessary to ensure that the risk of transmission remains below this threshold. Considerations should also include the feasibility and cost benefits of implementing such a monitoring system among the UK HIV-infected population. In the meantime behavioural interventions should promote awareness of the elevated transmission risk among the untreated, diagnosed population.

8.6 Thesis conclusions

The combination of sequence based, laboratory and clinical data can provide more insight into HIV transmission than can be gleaned from each source considered individually. The thesis has confirmed recent infection as an important risk factor in causing onward HIV transmission and provided further evidence for the debate of whether population level treatment of the diagnosed population – regardless of an individual's clinical need – should be implemented for public health purposes. There is a very real need for behavioural interventions targeted both at the at-risk HIV negative population (promoting awareness of elevated infectivity during seroconversion) and among the HIV diagnosed population (particularly emphasizing the increased transmission potential among the untreated population). The phylogenetic approach remains in its infancy and it is essential that methods continue to be assessed rigorously and results be interpreted with caution.

9 Appendix A - List of tables and figures

List of Tables

List of Figures

10 Appendix B - Consent letter for Brighton dataset

Lawson Unit
Department of GU Medicine
Eastern Road
Brighton
BN2 5BE

Study title: Understanding HIV transmission in men who have sex with men

You are being invited to take part in a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

1. What is the purpose of the study?
We are seeing a worrying increase in the number of new cases of HIV infection diagnosed in the UK, and the reasons for this are not clear. The aim of this study is to find out which factors contribute to this increase, and to use this knowledge to help stop more people becoming infected and in turn reduce the growth of the HIV epidemic. What we know so far suggests that patients recently infected with HIV (i.e. within 12 months) are much more likely to transmit the virus to another person (when there is a much higher level of HIV in the semen or vaginal secretions). This perhaps begins a chain of infection that grows wider and larger with time. Crucially, most people with early HIV have no indication that they are infected and feel perfectly well, and so will not be aware of the risk of infecting others. Thus, recognising people who may be at risk of early HIV infection may be very useful in preventing the transmission of HIV and thus prevent a much larger number of people being infected.

2. Why have I been chosen?
You are being invited to take part in this study because you are HIV positive, over 16 years of age and attend the Lawson Unit regularly. In order to carry out this research, we need to take an additional blood sample from you.

3. Do I have to take part?
It is up to you to decide whether or not to take part. If you decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. A decision not to take part will not affect in any way the standard of care you receive.

4. What will happen to me if I take part?
If you agree to join the study and sign the consent form at the back of this leaflet, the study doctor will then ask you to provide a sample of blood (10 ml). The blood sample will be obtained during your clinic visit.

The virus contained in this blood sample will be analysed* to identify any potential resistance you may have to anti-HIV drugs. This can help in deciding the best treatment choice for you, if and when you need to start HIV treatment

We would also like to look for similarities and differences that exist between different individuals' virus samples. If viruses from people who were recently HIV-infected are very similar, this may show that recently infected people may be more likely to infect others. We should have completed the testing of all samples by July 2007.

5. What are the possible disadvantages and risks of taking part?
The blood samples we ask you to provide, about a couple of teaspoonfuls, may cause some discomfort and/or bruising.

6. What if something goes wrong?

If you are harmed by taking part in this research project, there are no special compensation arrangements. If you are harmed due to someone's negligence, then you may have grounds for a legal action but you may have to pay for it. Regardless of this, if you wish to complain, or have any concerns about any aspect of the way you have been approached or treated during the course of this study, the normal National Health Service complaints mechanisms should be available to you.

7. What are the possible benefits of taking part?
Your participation will increase our knowledge of how HIV is transmitted within a population and will help us to develop ways of slowing or halting the epidemic. If we find that your virus is resistant to some anti-HIV drugs we will feed this back to you and it will help us make better decisions about your future anti-HIV treatment.

8. Will my taking part in this study be kept confidential?

Yes, we can guarantee that you will remain unidentifiable at any time during the study or after study completion. Confidentiality and anonymity will be highly protected, as follows.

Your blood sample will be sent to the laboratory for genotypic resistance analysis. The analysis will show whether your virus is resistant to any anti-HIV drugs. The results of this analysis will be fed back to you and will be included in your clinic notes. This will help us make better decisions about your future anti-HIV treatment.

We will then use the genotypic resistance analysis results for a study to see what factors are important in causing HIV transmission. Before we start this study will give your genotypic resistance analysis results an anonymous study number and delete any information that identifies you.

We will also take some very limited information from your clinic records to help interpret your resistance analysis results – no information will be taken that could identify you directly or indirectly.

The resistance analysis results and the limited clinical data will be sent to an independent centre (Health Protection Agency) together with an anonymous study number.

The Health Protection Agency will examine what factors are shared by viruses that are very similar to each other. It will be impossible to identify any individual throughout the duration of the study.

9. What will happen to the results of the research study?
Before the main analysis, your virus will be tested (genotypic resistance testing) to see if there are any drugs it is resistant to. These results will be fed back to you.

The main study looking at virus similarity will not be feedback to you as an individual since the data are completely anonymised before the analysis is carried out.

Similarly, neither Brighton HIV services nor the laboratory will be informed of individual results,

The overall results of the study will be made available to you at its conclusion, and may be published in a medical journal,

10. Contact for further information

You may ask questions about this study at any time. If you have any questions about the informed consent process or you require any further information, please contact your regular clinic doctor.

Alternatively you can contact Dr Martin Fisher who is the principal researcher for this project.

Title of Project: Understanding HIV transmission in men who have sex with men

Patient Identification Number for this trial:

Name of Researcher: Dr Martin Fisher

Please initial box

1. I confirm that I have read and understand the information sheet dated 27th November 02 ☐
(version 1.1) for the above study and have had the opportunity to ask questions.

2. I understand that my participation is voluntary and that I am free to withdraw at any time, ☐
without giving any reason, without my medical care or legal rights being affected.

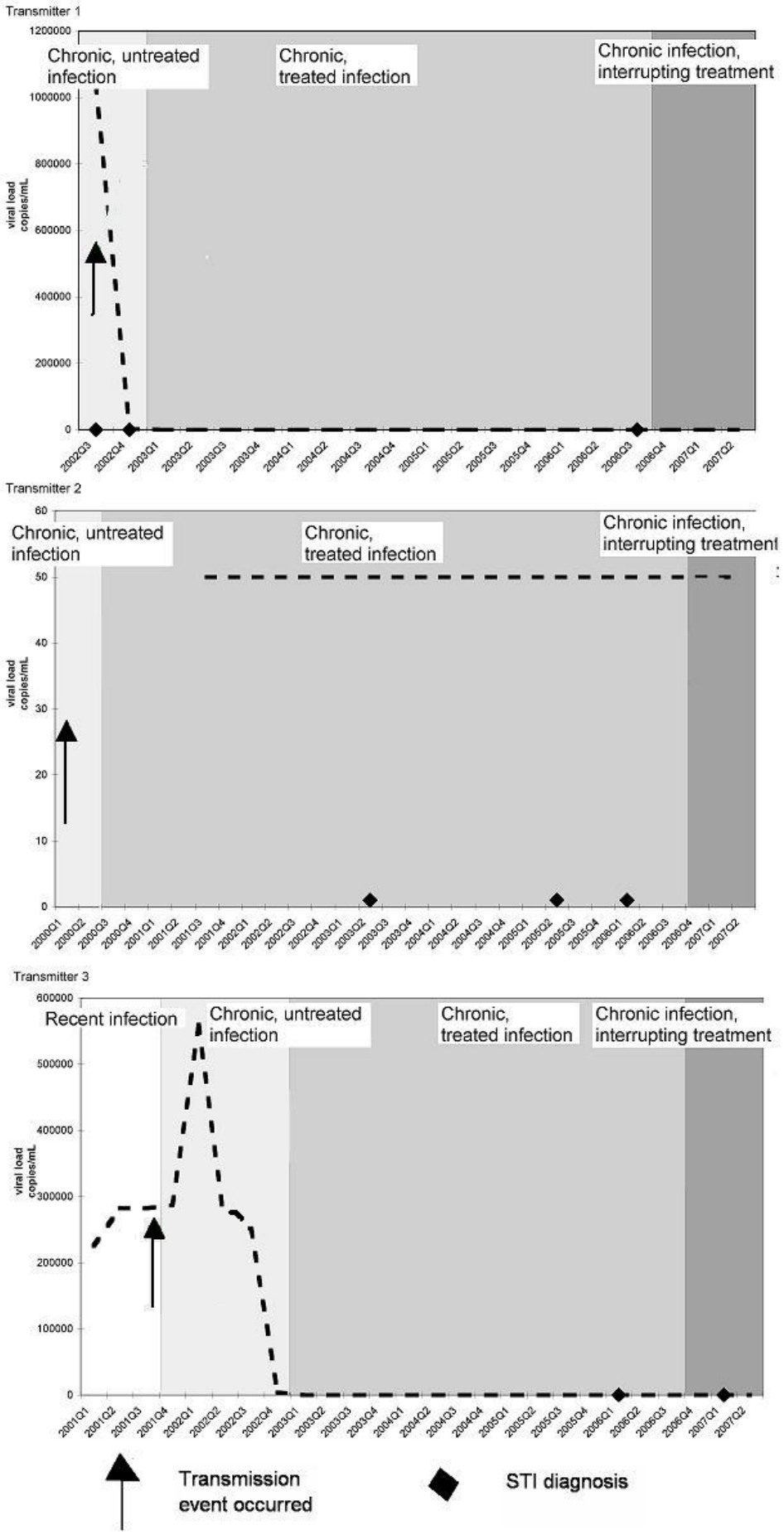
3. I understand that sections of my medical notes may be looked at by researchers ☐
from the Lawson Unit as is relevant to my taking part in research. I understand that my identity will remain unknown to the researchers at all times, up to and including the conclusion of the research. No data collected from the notes will relate either directly or indirectly to my identity. I give permission for these individuals to have access to my records.

4. I agree to take part in the above study.
☐

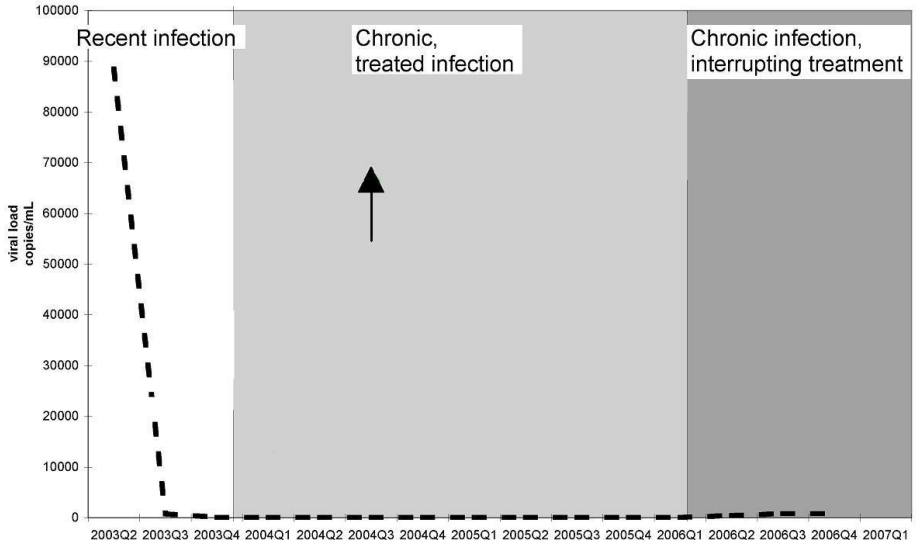
_____ _____
Name of Patient Date Signature

_____ _____
Name of Person taking consent Date Signature
(If different from researcher)

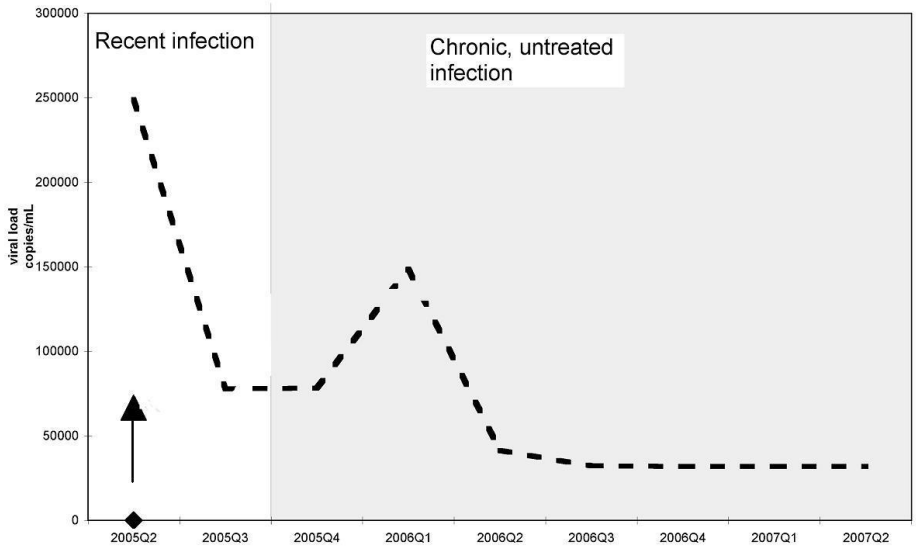
11 Appendix C - Life histories of transmission sources with estimated transmission dates


Transmitter 1
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
Transmitter 2
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
Transmitter 3
Recent infection
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
Transmission event occurred
STI diagnosis

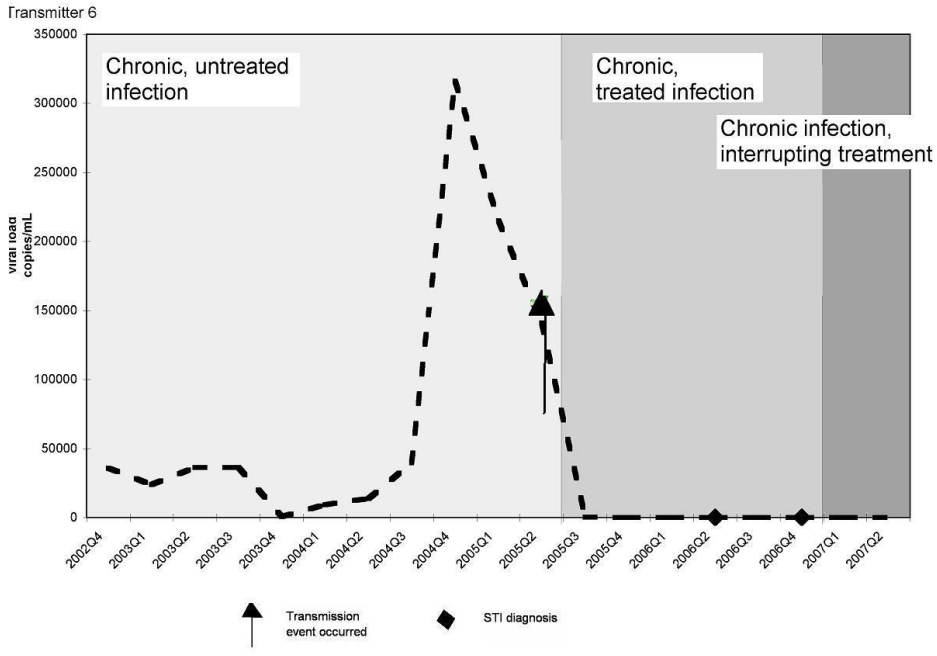
Transmitter 4



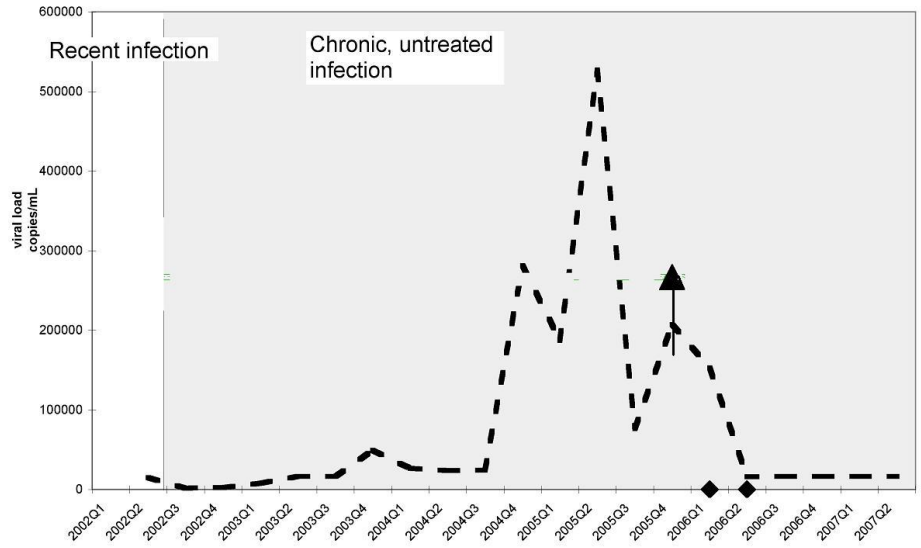
Transmitter 5



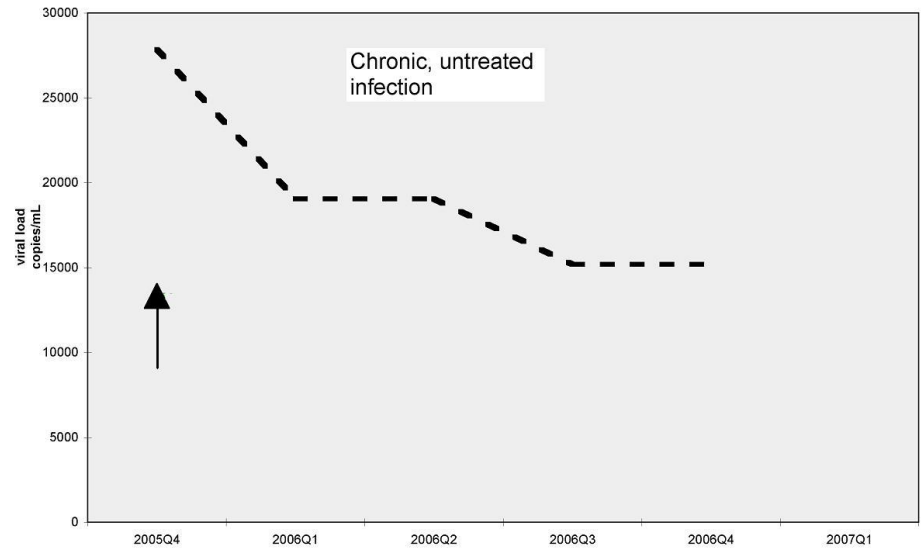
Transmitter 6



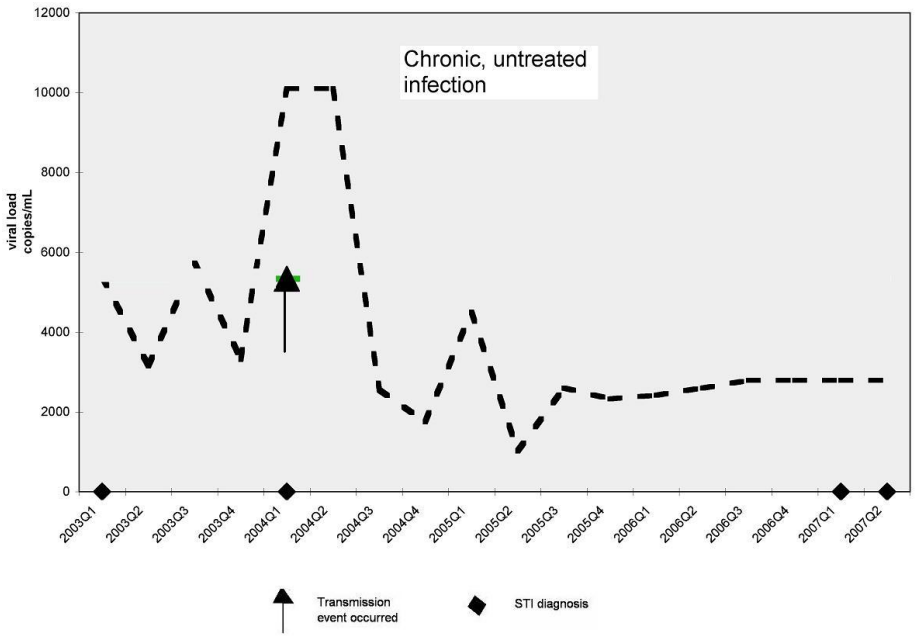
Transmitter 7

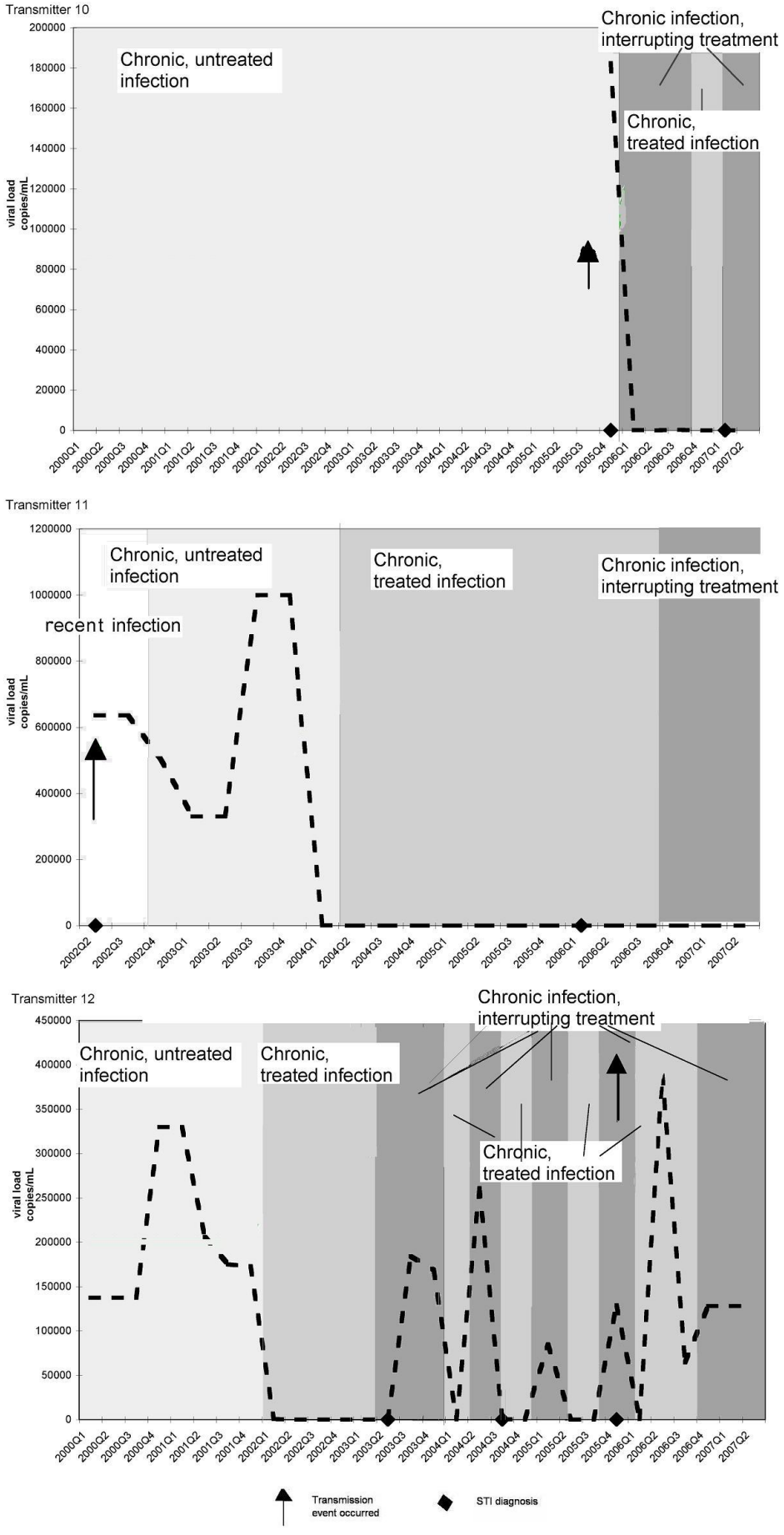


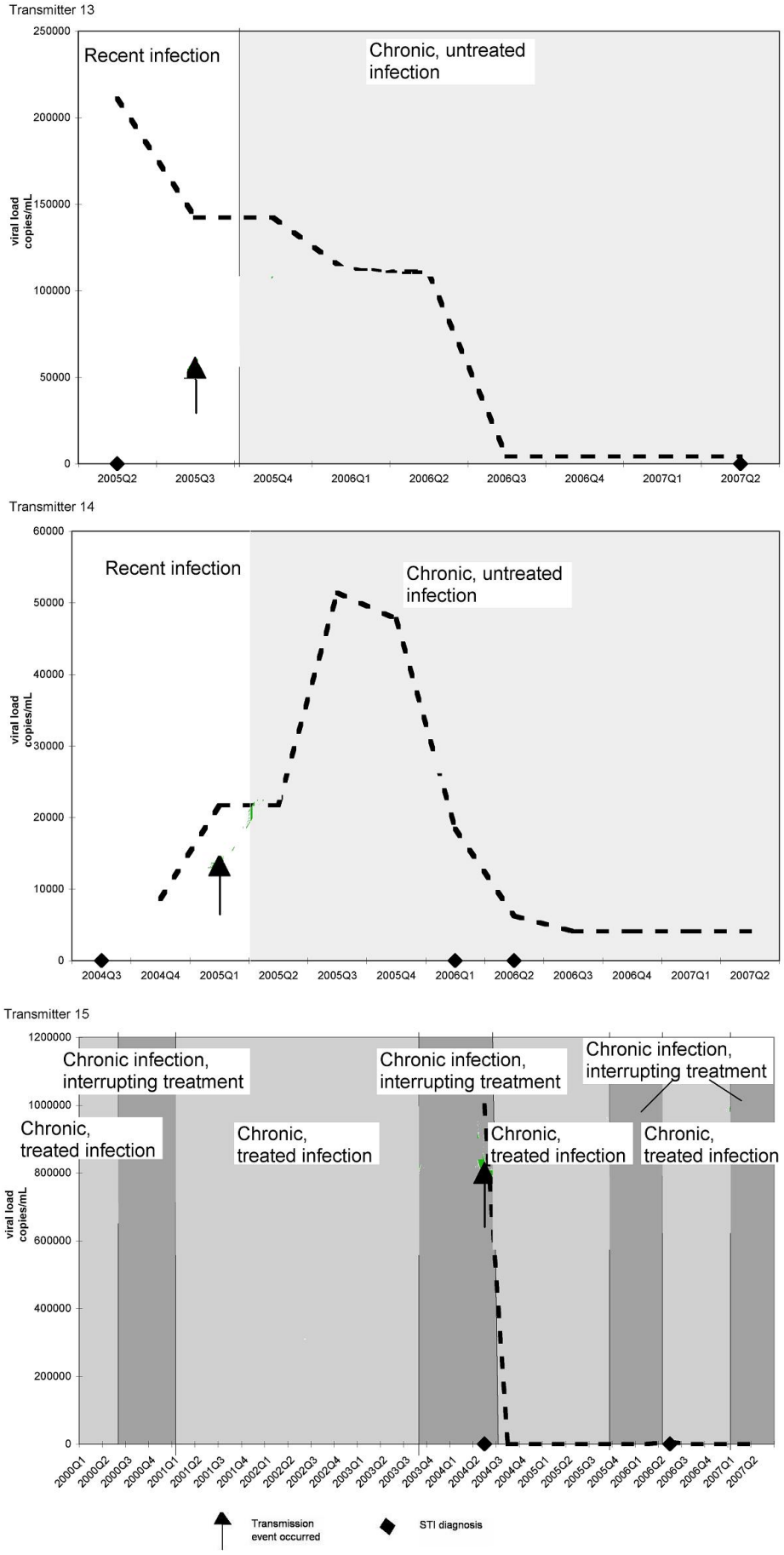
Transmitter 8



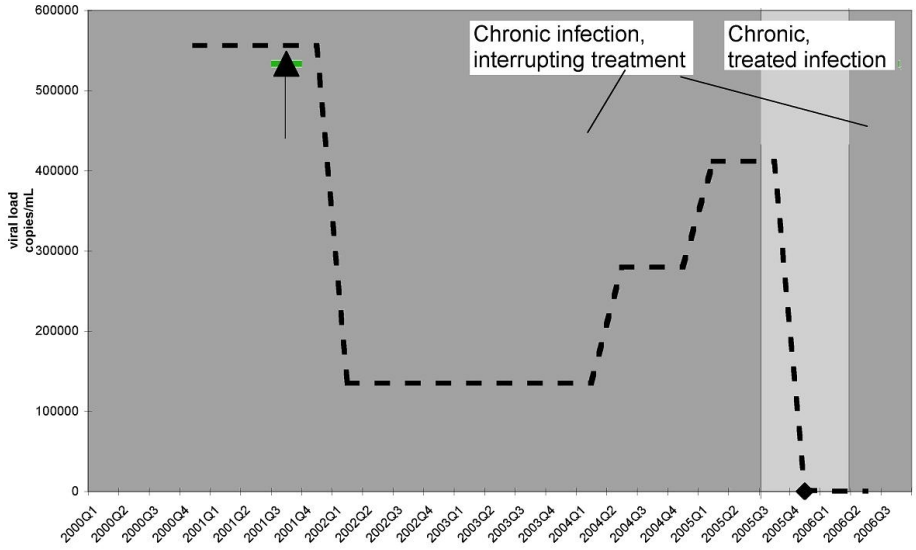
Transmitter 9



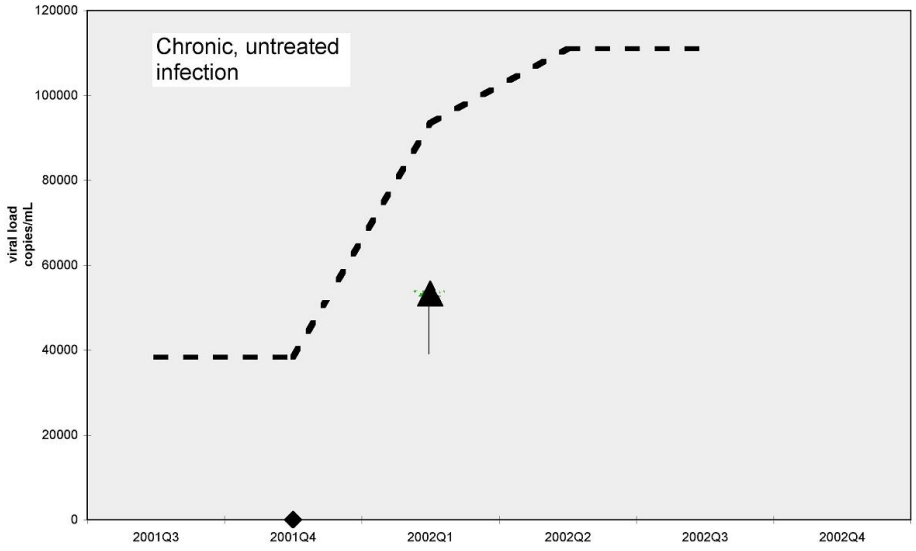

Transmitter 10
Chronic, untreated infection
Chronic infection, interrupting treatment
Chronic, treated infection
viral load copies/mL
Transmitter 11
Chronic, untreated infection
recent infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
Transmitter 12
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
Chronic, treated infection
viral load copies/mL
Transmission event occurred
STI diagnosis


Transmitter 13
Recent infection
Chronic, untreated infection
viral load copies/mL
250000
200000
150000
100000
50000
0
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmitter 14
Recent infection
Chronic, untreated infection
viral load copies/mL
60000
50000
40000
30000
20000
10000
0
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmitter 15
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic, treated infection
Chronic, treated infection
Chronic, treated infection
Chronic, treated infection
viral load copies/mL
1200000
1000000
800000
600000
400000
200000
0
2000Q1
2000Q2
2000Q3
2000Q4
2001Q1
2001Q2
2001Q3
2001Q4
2002Q1
2002Q2
2002Q3
2002Q4
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmission event occurred
STI diagnosis

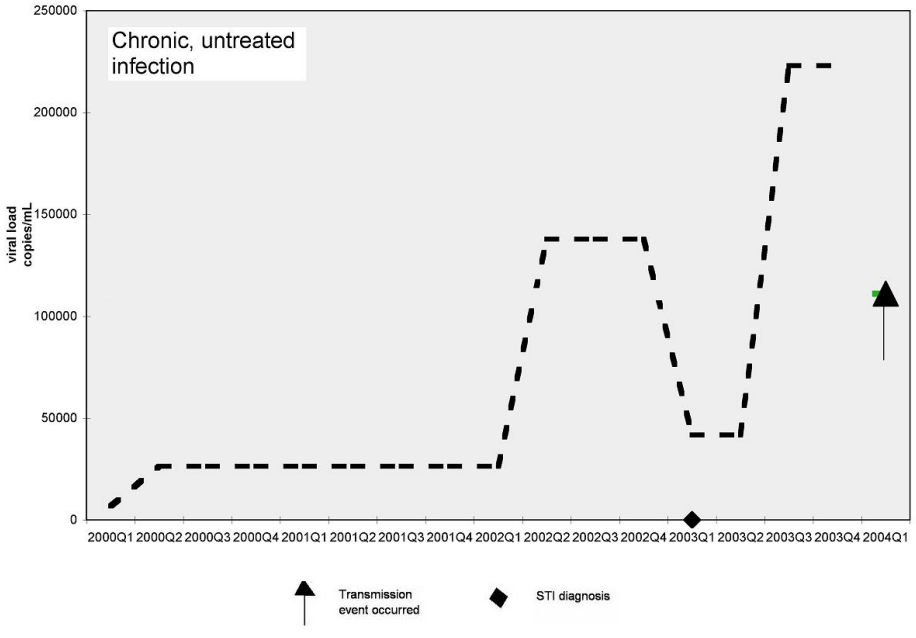
Transmitter 16



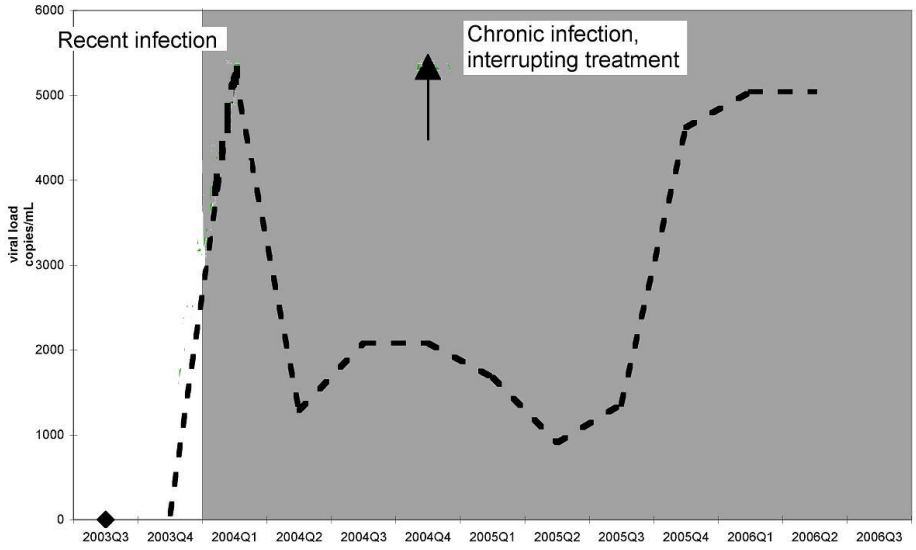
Transmitter 17



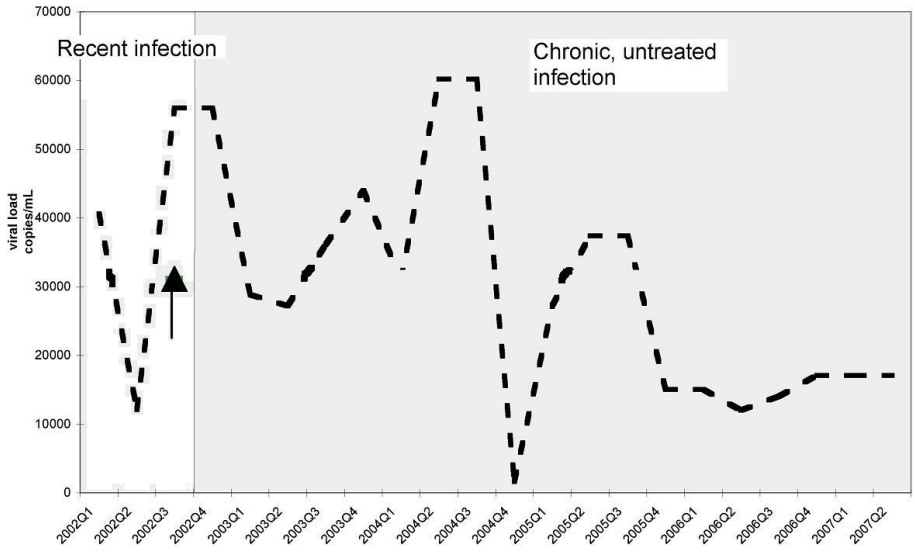
Transmitter 18



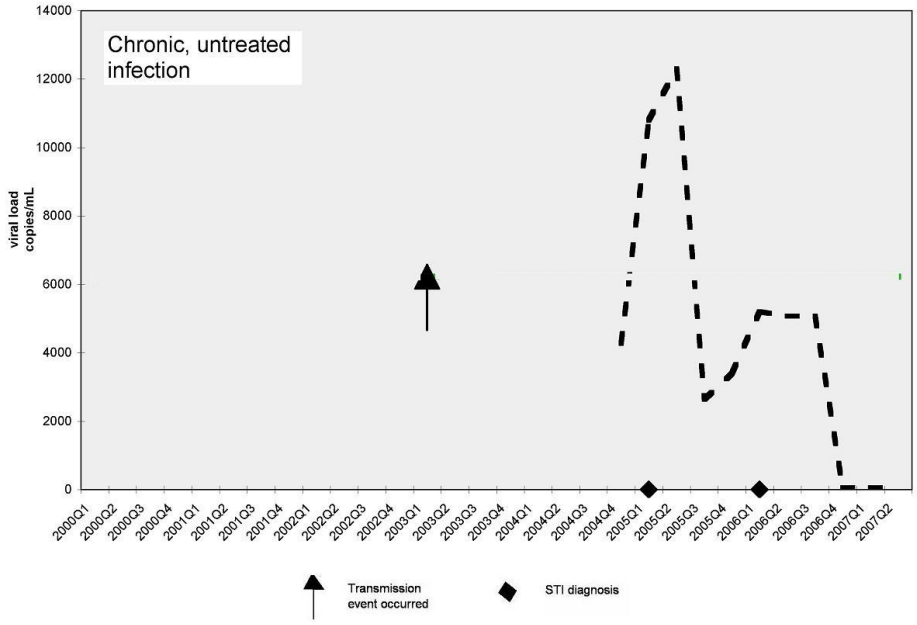
Transmitter 19



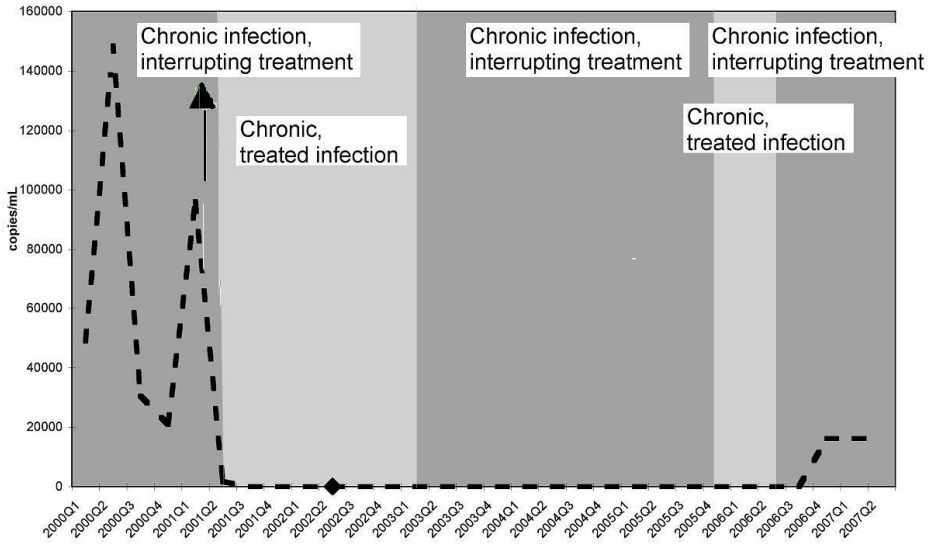
Transmitter 20



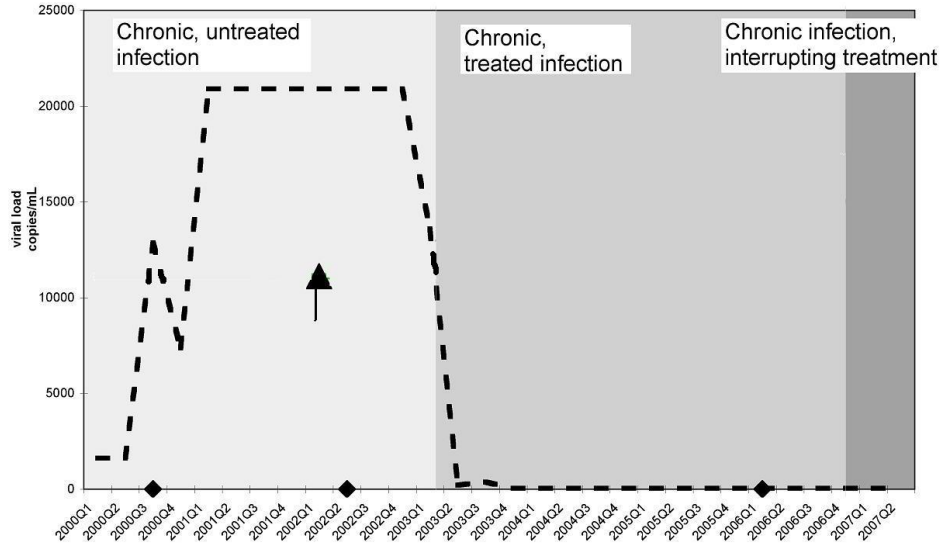
Transmitter 21



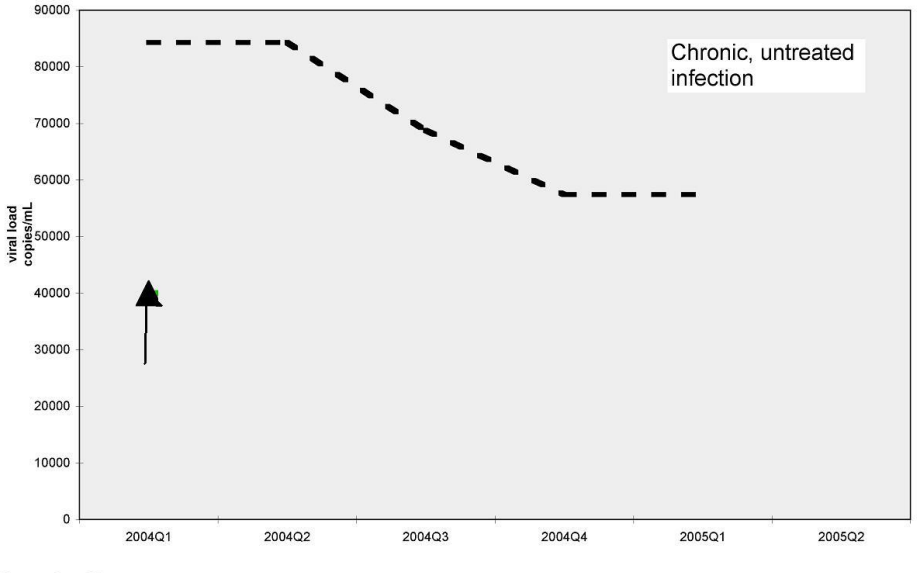
Transmitter 22

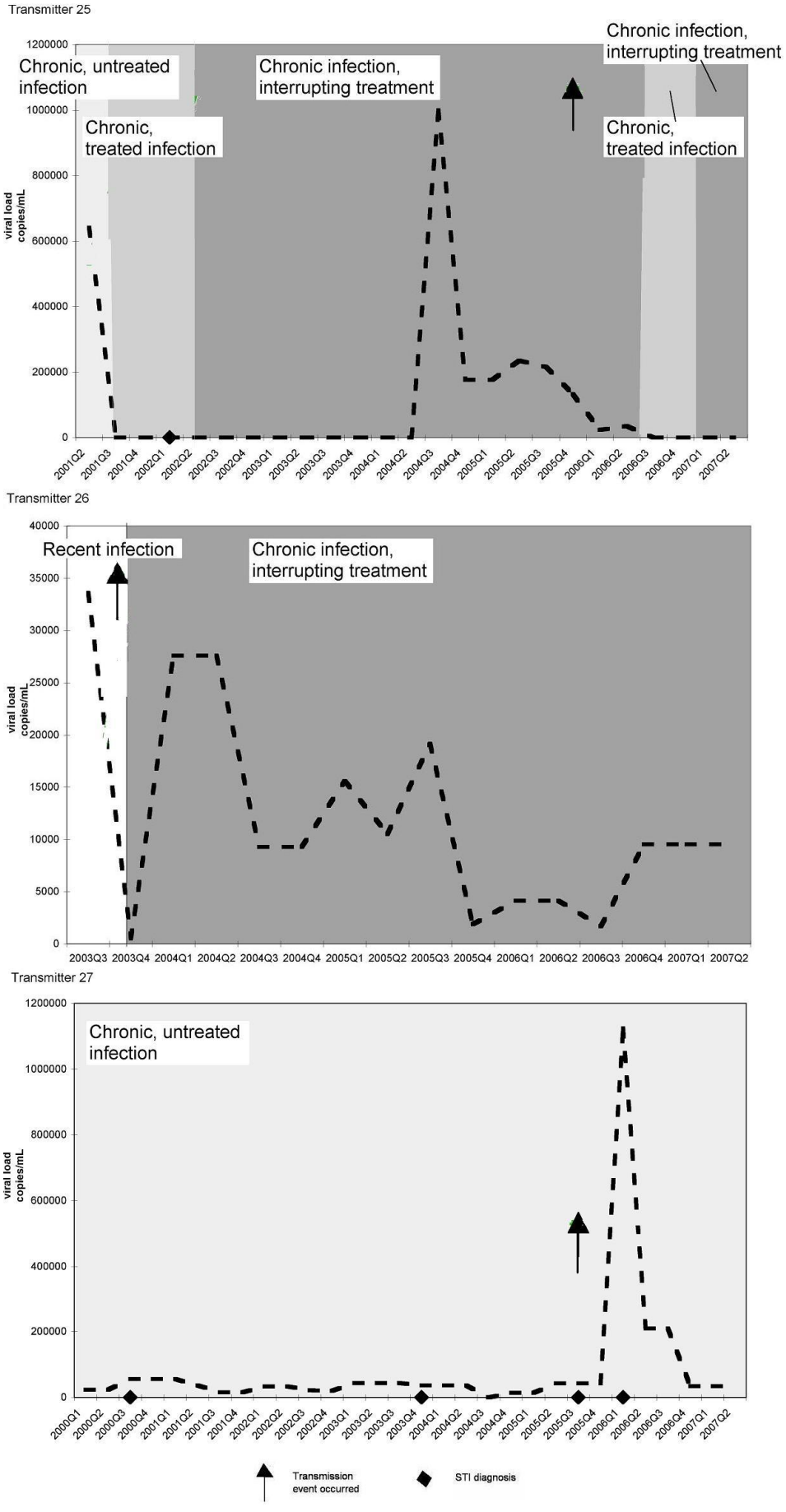


Transmitter 23

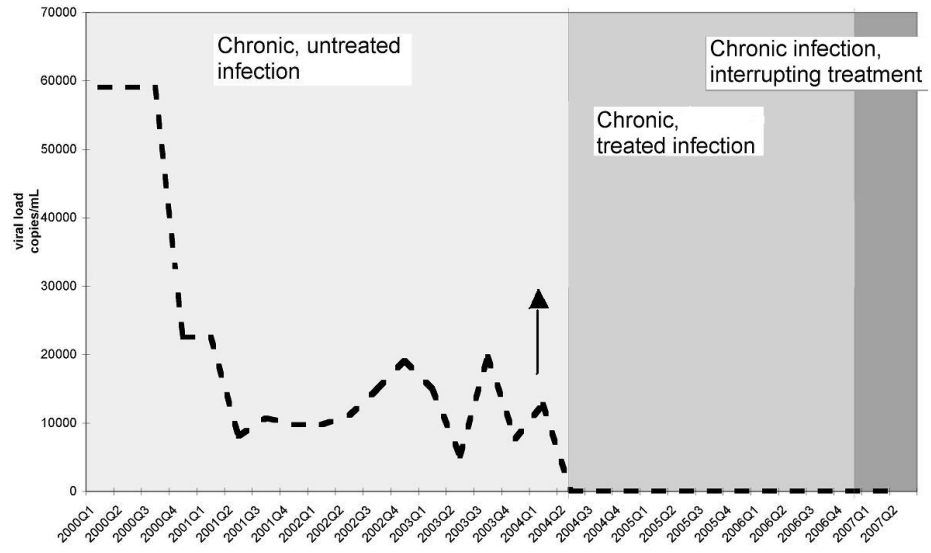


Transmitter 24

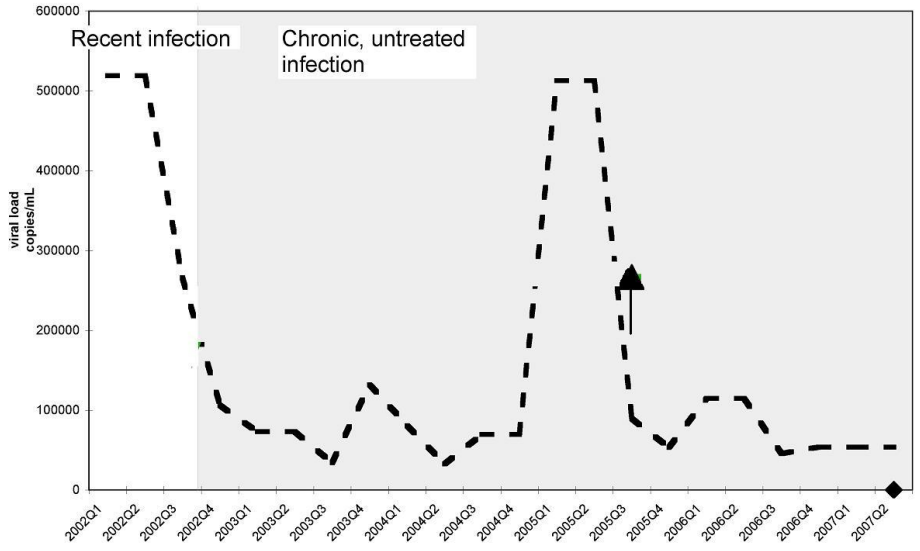



Transmitter 25
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic, treated infection
viral load copies/mL
Transmitter 26
Recent infection
Chronic infection, interrupting treatment
viral load copies/mL
Transmitter 27
Chronic, untreated infection
viral load copies/mL
Transmission event occurred
STI diagnosis

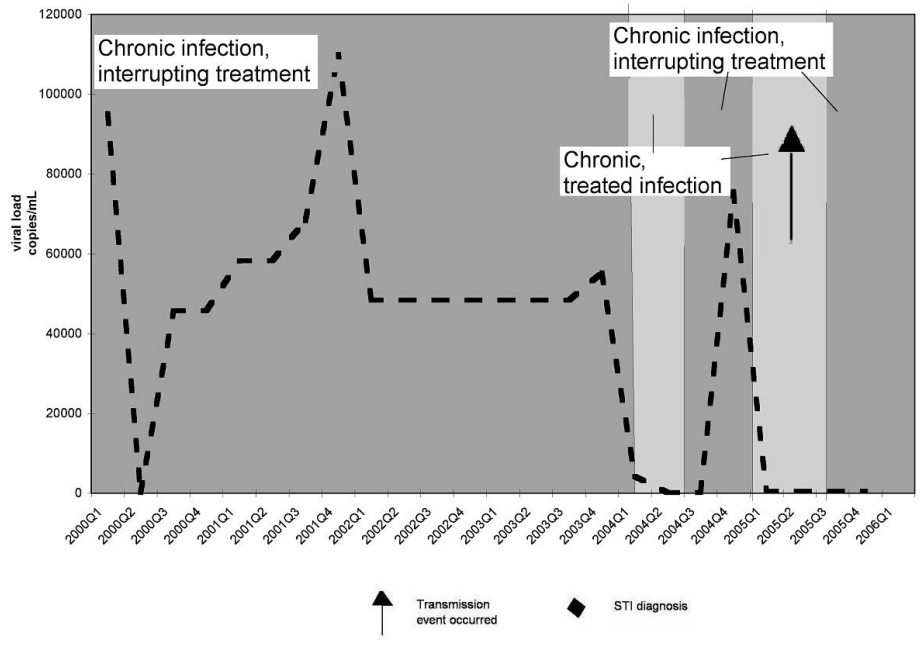
Transmitter 28

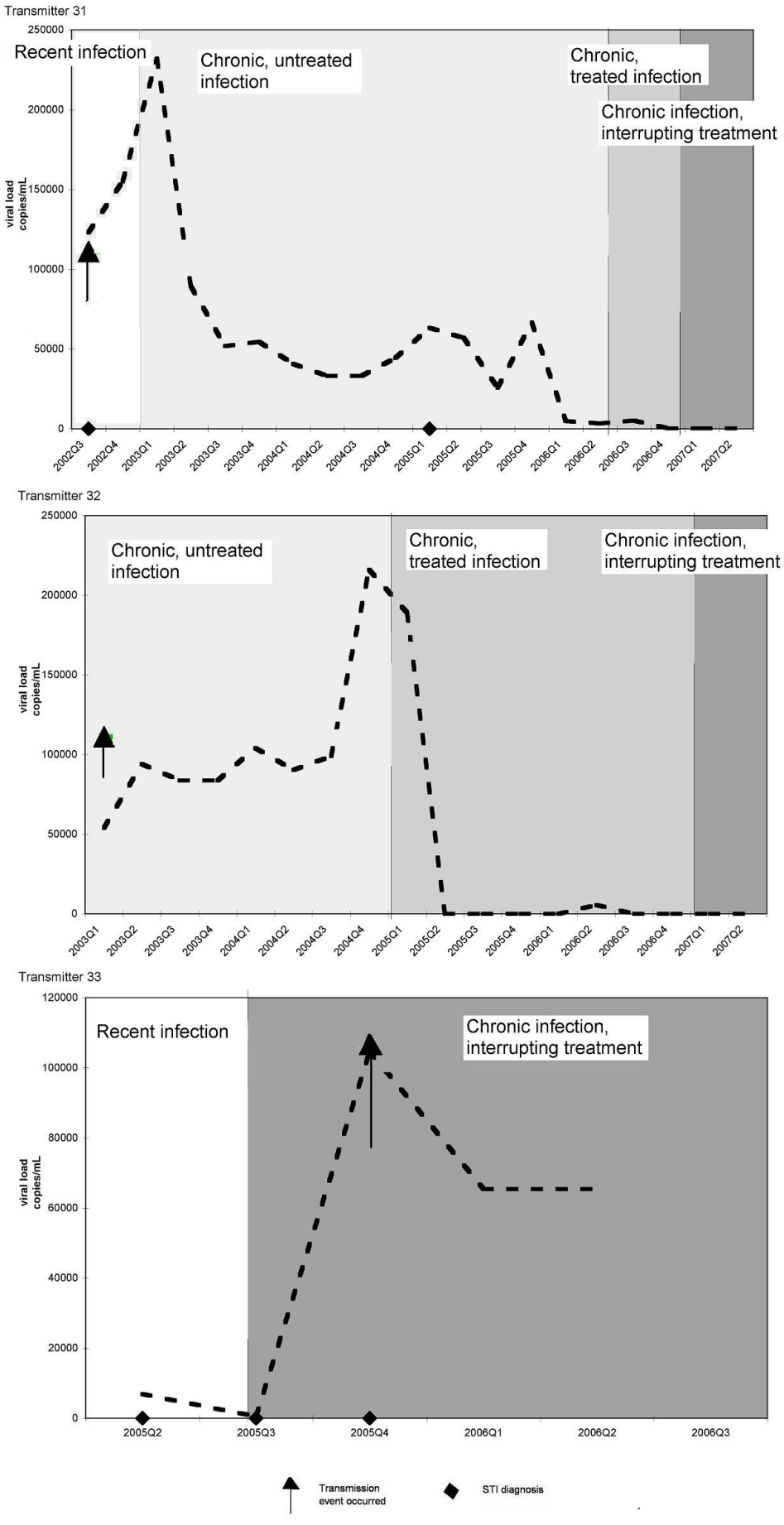


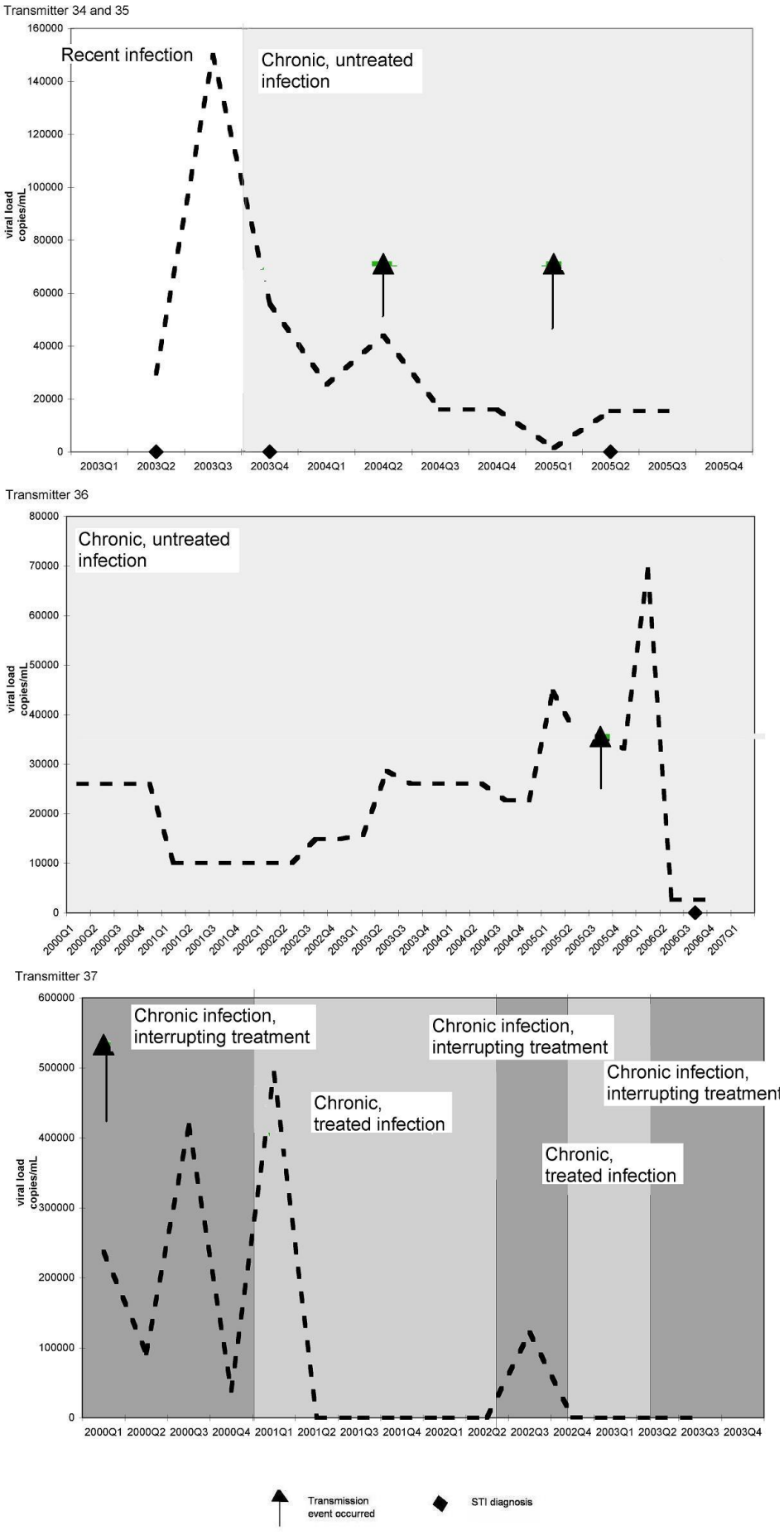
Transmitter 29

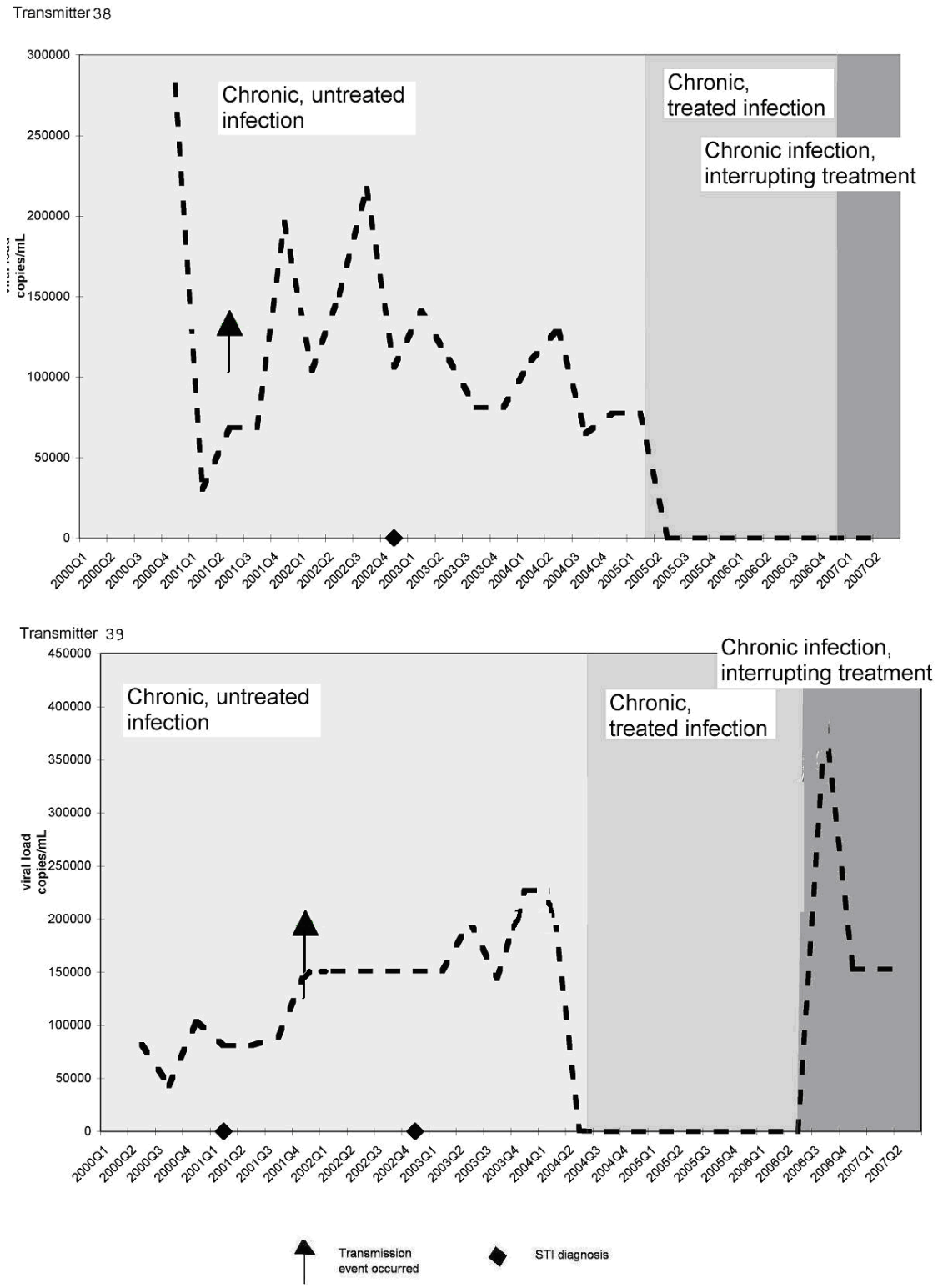


Transmitter 30




Transmitter 31
Recent infection
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
250000
200000
150000
100000
50000
0
2002Q3
2002Q4
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmitter 32
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
viral load copies/mL
250000
200000
150000
100000
50000
0
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmitter 33
Recent infection
Chronic infection, interrupting treatment
viral load copies/mL
120000
100000
80000
60000
40000
20000
0
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
Transmission event occurred
STI diagnosis


Transmitter 34 and 35
Recent infection
Chronic, untreated infection
viral load copies/mL
160000
140000
120000
100000
80000
60000
40000
20000
0
2003Q1 2003Q2 2003Q3 2003Q4 2004Q1 2004Q2 2004Q3 2004Q4 2005Q1 2005Q2 2005Q3 2005Q4
Transmitter 36
Chronic, untreated infection
viral load copies/mL
80000
70000
60000
50000
40000
30000
20000
10000
0
2000Q1 2000Q2 2000Q3 2000Q4 2001Q1 2001Q2 2001Q3 2001Q4 2002Q1 2002Q2 2002Q3 2002Q4 2003Q1 2003Q2 2003Q3 2003Q4 2004Q1 2004Q2 2004Q3 2004Q4 2005Q1 2005Q2 2005Q3 2005Q4 2006Q1 2006Q2 2006Q3 2006Q4 2007Q1
Transmitter 37
Chronic infection, interrupting treatment
Chronic, treated infection
Chronic infection, interrupting treatment
Chronic infection, interrupting treatment
Chronic, treated infection
viral load copies/mL
600000
500000
400000
300000
200000
100000
0
2000Q1 2000Q2 2000Q3 2000Q4 2001Q1 2001Q2 2001Q3 2001Q4 2002Q1 2002Q2 2002Q3 2002Q4 2003Q1 2003Q2 2003Q3 2003Q4
Transmission event occurred
STI diagnosis


Transmitter 38
Chronic, untreated infection
Chronic, treated infection
Chronic infection, interrupting treatment
copies/mL
300000
250000
200000
150000
100000
50000
0
2000Q1
2000Q2
2000Q3
2000Q4
2001Q1
2001Q2
2001Q3
2001Q4
2002Q1
2002Q2
2002Q3
2002Q4
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmitter 39
Chronic infection, interrupting treatment
Chronic, untreated infection
Chronic, treated infection
viral load copies/mL
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
2000Q1
2000Q2
2000Q3
2000Q4
2001Q1
2001Q2
2001Q3
2001Q4
2002Q1
2002Q2
2002Q3
2002Q4
2003Q1
2003Q2
2003Q3
2003Q4
2004Q1
2004Q2
2004Q3
2004Q4
2005Q1
2005Q2
2005Q3
2005Q4
2006Q1
2006Q2
2006Q3
2006Q4
2007Q1
2007Q2
Transmission event occurred
STI diagnosis

12 Appendix D - Published papers

BRIEF REPORT

Phylogenetic Reconstruction of Transmission Events from Individuals with Acute HIV Infection: Toward More-Rigorous Epidemiological Definitions

Alison E. Brown,[1,4] Robert J. Gifford,[2] Jonathan P. Clewley,[1] Claudia Kucherer,[3] Bernard Masquelier,[5] Kholoud Porter,[3] Claudia Balotta,[7] Nicole K. T. Back,[8] Louise Bruun Jorgensen,[9] Carmen de Mendoza,[10] Krishnan Bhaskaran,[3] O. Noel Gill,[1] Anne M. Johnson,[4] and Deenan Pillay,[1,2] on behalf of the Concerted Action on Seroconversion to AIDS and Death in Europe (CASCADE) Collaboration[a]

[1]Centre for Infections, Health Protection Agency, [2]Centre for Virology, University College London, [3]Medical Research Council Clinical Trials, and [4]Division of Population Health, University College London, London, United Kingdom; [5]Robert Koch Institute, Berlin, Germany; [6]Department de Virologie et Immunologie Biologique, Centre Hospitalier Universitaire, Bordeaux, France; [7]Institute of Infectious and Tropical Diseases, University of Milan, Milan, Italy; [8]Department of Medical Microbiology Academic Medical Centre, Amsterdam, the Netherlands; [9]Statens Serum Institut, Copenhagen, Denmark; [10]Service of Infectious Diseases, Hospital Carlos III, Madrid, Spain

Phylogenetic reconstructions of transmission events from individuals with acute human immunodeficiency virus (HIV) infection are conducted to illustrate this group's heightened infectivity. Varied definitions of acute infection and assumptions about observed phylogenetic clusters may produce misleading results. We conducted a phylogenetic analysis of HIV *pol* sequences from 165 European patients with estimated infection dates and calculated the difference between dates within clusters. Nine phylogenetic clusters were observed. Comparison of dates within clusters revealed that only 2 could have been generated during acute infection. Previous analyses may have incorrectly assigned transmission events to the acutely HIV infected when they were more likely to have occurred during chronic infection.

HIV-infected individuals may be more likely to transmit their virus onward during acute infection than during chronic infection [1]. This is because acute infection is associated with elevated blood and genital fluid viral loads [2] at a time when infection is frequently undiagnosed. Individuals who are acutely infected with drug-resistant HIV strains may also disproportionately drive further transmission of drug-resistant viruses because of a combination of increased infectivity and the persistence of resistant viruses in their plasma [3]. To ensure that prevention initiatives (such as promotion of HIV testing and adherence to antiretroviral therapy) are effective, it is important to measure the extent to which patients with acute HIV infection generate transmission at the population level, including the transmission of drug-resistant strains.

Sequence-based antiretroviral resistance testing has become commonplace to inform treatment options [4]; this has resulted in an accumulation of HIV *pol* sequences, the variability of which is adequate for phylogenetic reconstruction of transmission events [5]. Phylogenetic reconstructions of transmission events that combine HIV *pol* sequences with data on patients' infection stage (obtained either by comparing the time interval between the date of the last negative HIV test result and diagnosis [6] or by laboratory serological algorithms [7]) have the potential to enhance our understanding of the role that the acutely HIV infected play in generating HIV transmission. Through such methods, explorative phylogenetic analyses have identified possible transmission events generated by the acutely HIV infected, illustrating the elevated transmission potential among this group [8]. More recently, Brenner et al. [9] attempted to measure the extent to which the acutely HIV infected generate new HIV transmission events by comparing the proportion of new transmissions generated by the acutely and chronically infected in turn.

However, interpretation and comparison between studies is difficult because of varied sampling strategies and definitions of acute HIV infection used. Importantly, the latter may serve to overestimate the force of transmission from this population. Viral load is known to peak less than a month after infection but remains elevated for ~10 weeks [2]. However, definitions of acute infection and laboratory techniques used to identify it vary [7, 10, 11], and those frequently adopted mean that patients can be so categorized >6 months after infection [8, 9]. Moreover,

Received 18 March 2008; accepted 27 August 2008; electronically published 10 January 2009.
Potential conflicts of interest: none reported.

The Journal of Infectious Diseases 2009;199:427–31

0022-1899/2009/19903-0019$15.00
DOI: 10.1086/596049

Financial support: University College London Hospitals/University College London Comprehensive Biomedical Centre (support to this study). The CASCADE Collaboration has been funded through the European Union (grants BMH4-CT97–2550, QLK2–2000-01481, QLRT-2001–01709, and LSHP-CT-2006–018843).
[a] Study group members are listed after the text.
Reprints or correspondence: Alison Brown, HIV/STI Dept., Health Protection Agency Centre for Infections, 61 Colindale Ave., London NW9 5EQ, UK (alison.brown@hpa.org.uk).

studies do not consider the transient nature of acute infection: phylogenetic analyses frequently categorize patients as being acutely or chronically infected according to their infection stage at diagnosis [8, 9]. However, all patients experience acute and chronic infection, and the observation of possible transmission events (ascertained through phylogenetic reconstruction) between patients diagnosed during acute infection does not necessarily mean that transmission occurred while the patients were acutely HIV infected. For instance, Brenner et al. [9] showed that the mean interval between infection dates within clusters was 15 months—suggesting that most identified transmissions originated from chronically HIV-infected individuals—but interpreted these as acute-to-acute transmissions.

We conducted a phylogenetic analysis of a previously described HIV-infected European population [6] that used precise definitions of acute HIV infection to identify instances of possible transmission events, including transmission of drug-resistant strains. We calculated the difference between the estimated infection dates within the observed clusters to determine whether transmission could have occurred during acute infection. The analysis is not an attempt to measure the force of infection from the acutely HIV infected but a demonstration of the need for rigor in the design and interpretation of such analyses.

Methods. The study population was derived from CASCADE (http://www.cascade-collaboration.org) [12], an ongoing international collaboration that monitors events among HIV-infected individuals throughout their infections. HIV-1–infected patients were included provided they were >15 years old and had a reliable and precise date of infection. For the present analysis, this involved either an HIV-positive test result within 90 days of an HIV-negative test result or an HIV antibody–negative test result with polymerase chain reaction (PCR) positivity. The estimated infection date was taken as the midpoint between the diagnosis date and last negative test result for the former and as the date the sample was taken for the latter.

The first plasma sample available after the estimated HIV infection date was obtained for genotypic resistance testing. Samples obtained >18 months after the estimated infection date or from drug-experienced patients were excluded. Nucleotide sequence data spanned the protease gene and at least codons 41–236 of reverse transcriptase (RT). The sequencing methodology has been described previously [6]. Drug-resistance mutations were defined on the basis of the standardized list of mutations for use in epidemiological studies [13].

Before phylogenetic analysis, the mutation sites associated with drug resistance [13] were deleted from all sequences. Protease and RT sequences were aligned across 998 nt and imported into the tree-building software PAUP (version 4.0b10). A neighbor-joining tree was created under the general time-reversible (GTR) model with gamma rate heterogeneity set at 0.5. The alignment was run through Modeltest software (version 3.7) to select the best-fitting evolutionary model. The GTR+I+G model was selected, and a heuristic search was conducted for a maximum-likelihood tree by means of this model and its derived parameters (proportion of invariable sites, 0.43; gamma distribution, 0.79), using the initial neighbor-joining tree as the starting tree. The robustness of the tree was evaluated by bootstrap analysis with 100 replicates. Genetic distances were calculated from the maximum-likelihood tree topology. Branches comprising 2 or more sequences with a bootstrap value of >99% and a genetic distance of <0.015 nucleotide substitutions per site were considered to constitute a cluster, representing possible transmission [5]. To assess whether a possible transmission was generated by an individual during acute infection, clusters containing sequences with a difference of <180 days between estimated infection dates were interpreted as being a possible acute-to-acute transmission.

Ethics approval for cohorts contributing to the CASCADE Collaboration is maintained according to the respective national regulations in the countries concerned. The nucleotide sequences used in this study were deposited into GenBank under the accession numbers FJ030643–FJ030807.

Results. Of 8993 patients with an estimated infection date from data pooled in December 2004, HIV sequences were submitted from 10 European cohorts. Of these, 165 sequences were complete, were from drug-naive patients, and fulfilled the definition of acute HIV infection. Specifically, 131 (79%) had had a documented seronegative test result at least 90 days before their diagnosis (mean number of days between tests, 38 days), and 34 (21%) were HIV antibody negative with PCR positivity.

The majority of sequences were from men who had had sex with men (120/165 [73%]). The sample proportion from each country varied annually; overall, 23% (38/165) of the samples were from Italy, 23% (38/165) were from France, 20% (33/165) were from Germany, and 17% (28/165) were from the UK, with the remainder from Denmark, Spain, and the Netherlands.

Overall, 10% (17/165) of patients sampled were infected with a drug-resistant HIV-1 variant. Fourteen patients had a virus considered resistant to 1 or more nucleotide RT inhibitors (NRTIs), 3 carried a virus resistant to nonnucleotide RT inhibitors (NNRTIs), and 1 had a virus resistant to protease inhibitors. One patient carried virus with both NRTI and NNRTI mutations.

Of the 165 sequences, 18 (11%) formed a phylogenetic cluster with at least 1 other sequence (clusters A-I) (figure 1). These formed 9 phylogenetic clusters, with 2 sequences in each cluster. All 9 phylogenetic clusters were composed of sequences derived from patients sampled within the same country. Two of the 9 clusters contained sequences that had a difference between infection dates of <180 days (104 and 29 days for clusters E and G, respectively) (table 1). For the remaining 7 clusters, the average difference in infection dates within clusters was 675 days (range, 187–1571 days).

Of the 18 sequences with drug-resistance mutations, 2 formed a phylogenetic cluster (C) (figure 1). Both sequences shared the

Figure 1. Phylogenetic tree showing 9 clusters and sequences with drug-resistance mutations within a European cohort of HIV-infected patients with well-estimated infection dates (<1995–2004). Phylogenetic clusters are shown in boldface and are labeled A-I; red diamonds denote drug-resistance mutations.

F77L mutation and were derived from patients with a difference of 226 days between their infection dates table 1.

Discussion. We conducted a phylogenetic analysis of HIV-infected European patients with well-estimated infection dates and calculated the difference between infection dates within the possible transmission clusters identified. Nine transmissions may have occurred between individuals in this sample, including 1 possible instance of transmitted drug resistance. Of the possible transmissions events identified, only 2 could have been generated during acute HIV infection.

This analysis demonstrates the need for caution in the design and interpretation of phylogenetic analyses that link HIV *pol* sequences to data on patients' infection stage. For the first time, the time intervals between infection dates between patients involved in a possible transmission event have been explored. This comparison revealed that only 2 possible transmissions could have occurred during acute infection. The remainder are more likely to be transmissions from individuals with chronic infection. Previous analyses that use liberal definitions of acute HIV infection and do not take into account the transient nature of infection stage may inaccurately calculate the force of infection from this group, including the transmission of drug-resistant viruses [9], within the populations studied.

We recommend that future analyses should comprise sequences from patients from a local HIV-infected population and contain a large proportion of patients with well-estimated infection dates. Methods should then allow each patient's infection stage to change from acute to chronic to reflect the natural progression of HIV infection. Consequently, both acute and chronically HIV-infected patients will be represented in the sample.

Table 1. Difference between estimated infection dates and drug-resistance mutations for the phylogenetic clusters shown in figure 1.

Cluster	Difference between estimated infection dates,[a] days	Drug-resistance mutations
A	922	...
B	1241	...
C	226	2 × F77L
D	348	...
E	104	...
F	232	...
G	29	...
H	187	...
I	1571	...

[a] Difference within each phylogenetic cluster between the sequence with the earliest infection date (day 0) and the sequence with the next chronological infection date (e.g., a phylogenetic cluster consisting of 2 sequences with infection dates of 15 April 2002 and 13 July 2002 would be presented as 89 days).

Additionally, we suggest that identified transmission events that involve patients with acute infection could be used to approximate the transmission date (i.e., transmissions involving an acutely HIV-infected individual are likely to have occurred around that patient's time of diagnosis). This method of dating will allow researchers to ascertain whether the patient most likely to have generated the acute infection was acutely or chronically infected at the transmission date (provided he or she also has an estimated infection date).

Our analysis is not a fresh attempt to measure the force of transmission from patients with acute HIV infection. The data were selected because they contained well-estimated infection dates. However, the size and the inconsistent geographic and demographic makeup of the population sample prohibits the extrapolation of results to the HIV-infected population and the calculation of the rate of transmission from the acutely HIV-infected population.

Further methodological challenges are applicable to all phylogenetic analyses of HIV-infected patients. First, although the bootstrapping and genetic distance cutoffs used are conservative and have been shown to reduce the proportion of false-positive clusters among an analysis of UK HIV sequences [5], it is not possible to conclusively demonstrate that the possible transmission events identified in the present analysis represent actual transmissions [14]. The clusters observed could each have been generated by a third party (not sampled) infecting each individual sequenced in the cluster or acting as an intermediary between the sequenced persons. Similarly, transmissions from the patients represented in this study would not be identified if the sequence from the infection generated was not included in the sample. Additionally, such studies cannot include sequences from the population with undiagnosed infections (a higher proportion of the acutely HIV infected are likely to have undiagnosed infections compared with those with chronic infection). Finally, the phylogenetic transmissions observed may be relative to the genetic diversity among the sampled sequences [14] rather than reflect absolute transmission events. For instance, Brenner et al. [9] found that the phylogenetic clusters observed among the acutely HIV infected were disrupted when additional sequences were added.

We have demonstrated the need for precise definitions of acute HIV infection in phylogenetic transmission reconstructions and the requirement that the possible transmission events observed through such analyses take into account the transient nature of acute infection. Future analyses should (1) use population data that contain a large proportion of patients with well-estimated infection dates, (2) allow each patient to progress from acute to chronic infection, (3) use transmissions involving acute infections to approximate the date of transmission and thereby ascertain the infection stage of the most likely transmitter at around the time of transmission, and (4) use a sample that is broadly representative of a local HIV-infected population with respect to geography and time. Only then can we attempt to measure with any precision the extent to which the acutely HIV infected drive HIV transmission at the population level.

CASCADE Collaboration. Steering Committee: Julia Del Amo (chair), Laurence Meyer (vice-chair), Heiner Bucher, Geneviève Chêne, Deenan Pillay, Maria Prins, Magda Rosinska, Caroline Sabin, and Giota Touloumi. *Coordinating Centre:* Kholoud Porter (project leader), Sara Lodi, Sarah Walker, Abdel Babiker, and Janet Darbyshire. *Clinical Advisory Board:* Heiner Bucher, Andrea de Luca, Martin Fisher, and Roberto Muga. *Collaborators:* Australia—Sydney AIDS Prospective Study and Sydney Primary HIV Infection Cohort (John Kaldor, Tony Kelleher, Tim Ramacciotti, Linda Gelgor, David Cooper, and Don Smith); Canada—South Alberta Clinic (John Gill); Denmark—Copenhagen HIV Seroconverter Cohort (Louise Bruun Jørgensen, Claus Nielsen, and Court Pedersen); Estonia—Tartu Ülikool (Irja Lutsar); France—Aquitaine Cohort (Geneviève Chêne, Francois Dabis, Rodolphe Thiebaut, and Bernard Masquelier), French Hospital Database (Dominique Costagliola and Marguerite Guiguet), Lyon Primary Infection Cohort (Philippe Vanhems), and SEROCO Cohort (Laurence Meyer and Faroudy Boufassa); Germany—German Cohort (Osamah Hamouda and Claudia Kucherer); Greece—Greek Haemophilia Cohort (Giota Touloumi, Nikos Pantazis, Angelos Hatzakis, Dimitrios Paraskevis, and Anastasia Karafoulidou); Italy—Italian Seroconversion Study (Giovanni Rezza, Maria Dorrucci, Benedetta Longo, and Claudia Balotta); Netherlands—Amsterdam Cohort Studies among Homosexual Men and Drug Users (Maria Prins, Liselotte van Asten, Akke van der Bij, Ronald Geskus, and Roel Coutinho); Norway—Oslo and Ullevål Hospital Cohorts (Mette Sannes, Oddbjorn Brubakk, Anne Eskild, and Johan N. Bruun); Poland—National Institute of Hygiene (Magdalena Rosinska);

Portugal—Universidade Nova de Lisboa (Ricardo Camacho); Russia—Pasteur Institute (Tatyana Smolskaya); Spain—Badalona IDU Hospital Cohort (Roberto Muga), Barcelona IDU Cohort (Patricia Garcia de Olalla), Madrid Cohort (Julia Del Amo and Jorge del Romero), Valencia IDU Cohort (Santiago Pérez-Hoyos and Ildefonso Hernandez Aguado); Switzerland—Swiss HIV Cohort (Heiner Bucher, Martin Rickenbach, and Patrick Francioli); Ukraine—Perinatal Prevention of AIDS Initiative (Ruslan Malyuta); United Kingdom—Edinburgh Hospital Cohort (Ray Brettle), Health Protection Agency (Valerie Delpech, Sam Lattimore, Gary Murphy, John Parry, and Noel Gill), Royal Free Haemophilia Cohort (Caroline Sabin and Christine Lee), UK Register of HIV Seroconverters (Kholoud Porter, Anne Johnson, Andrew Phillips, Abdel Babiker, Janet Darbyshire, and Valerie Delpech), University College London (Deenan Pillay), and University of Oxford (Harold Jaffe).

References

1. Gray RH, Wawer MJ, Brookmeyer R, et al. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. Lancet 2001; 357:1149–53.
2. Pilcher CD, Joaki G, Hoffman IF, et al. Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection. AIDS 2007; 21:1723–30.
3. Devereux H, Youle LM, Johnson MA, Loveday C. Rapid decline in detectibility of HIV-1 drug resistance mutations after stopping therapy. AIDS 1999; 13:F123–7.
4. Pozniak A, Gazzard B, Anderson J, et al. British HIV Association (BHIVA) guidelines for the treatment of HIV-infected adults with antiretroviral therapy. HIV Med 2003; 4:1–41.
5. Hue 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; 18:719–28.
6. Masquelier B, Bhaskaran K, Pillay D, et al. Prevalence of transmitted HIV-1 drug resistance and the role of resistance algorithms: data from seroconverters in the CASCADE Collaboration from 1987 to 2003. JAIDS 2005; 40:505–11.
7. Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. JAMA 1998; 280:42–8.
8. Pao D, Fisher M, Hue S, et al. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. AIDS 2005; 19:85–90.
9. Brenner B, Roger MN, Routy JP, et al. High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 2007; 195: 951–9.
10. Fiscus SA, Pilcher CD, Miller WC, et al. Rapid, real-time detection of acute HIV infection in patients in Africa. J Infect Dis 2007; 195:416–24.
11. Pilcher CD, Fiscus SA, Nguyen TQ, et al. Detection of acute infections during HIV testing in North Carolina. N Engl J Med 2005; 352:1873–83.
12. Porter K, Babiker A, Bhaskaran K, et al. Determinants of survival following HIV seroconversion after the introduction of HAART. Lancet 2003; 362:1267–74.
13. Shafer RW, Rhee SY, Pillay D, et al. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. AIDS 2007; 21:215–23.
14. Hue S, Clewley JP, Cane PA, Pillay D. Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. AIDS 2005; 19:449–50.

Clinical

EDITOR'S CHOICE

Implications for HIV testing policy derived from combining data on voluntary confidential testing with viral sequences and serological analyses

A E Brown,[1,2] G Murphy,[1] G Rinck,[1] J P Clewley,[1] C Hill,[1] J V Parry,[1] A M Johnson,[2] D Pillay,[1,3] O N Gill[1]

See Editorial, p 2

[1] Health Protection Agency Centre for Infections, London, UK; [2] Division of Population Health, University College London, London, UK; [3] Department of Virology, University College London, London, UK

Correspondence to: Ms A E Brown, HIV/STI Department, Health Protection Agency Centre for Infections, 61 Colindale Avenue, London NW9 5EQ, UK; Alison.brown@hpa.org.uk

Accepted 13 October 2008
Published Online First 27 October 2008

ABSTRACT

Objectives: Laboratory, clinical and sequence-based data were combined to assess the differential uptake of voluntary confidential HIV testing (VCT) according to risk and explore the occurrence of HIV transmission from individuals with recently acquired HIV infection, before the diagnostic opportunity.

Methods: Between 1999 and 2002, nearly 30 000 anonymous tests for previously undiagnosed HIV infection were conducted among men who have sex with men (MSM) attending 15 sentinel sexually transmitted infection (STI) clinics in England, Wales and Northern Ireland. Using a serological testing algorithm, undiagnosed HIV-infected men were categorised into those with recent and non-recent infection. VCT uptake was compared between HIV-negative, recently HIV-infected and non-recently HIV-infected men. A phylogenetic analysis of HIV *pol* sequences from 127 recently HIV-infected MSM was conducted to identify instances in which transmission may have occurred before the diagnostic opportunity.

Results: HIV-negative MSM were more likely to receive VCT at clinic visits compared with undiagnosed HIV-infected MSM (56% (14 020/24 938) vs 31% (335/1072); p<0.001). Recently HIV-infected MSM were more likely to receive VCT compared with those with non-recent infections (42% (97/229) vs 28% (238/844); p<0.001). 22% (95/425) of undiagnosed HIV-infected MSM with STI received VCT. Phylogenetic analysis revealed at least seven transmissions may have been generated by recently HIV-infected MSM: a group that attended STI clinics soon after seroconversion.

Conclusions: The integration of clinical, laboratory and sequence-based data reveals the need for specific targeting of the recently HIV exposed, and those with STI, for VCT. VCT promotion alone may be limited in its ability to prevent HIV transmission.

Sequence-based, laboratory and clinical data are collected routinely around HIV diagnoses for medical and surveillance purposes. In particular, sequence-based antiretroviral resistance testing has recently become routine to inform treatment options.[1] This has resulted in an accumulation of HIV *pol* sequences, the variability of which allows phylogenetic reconstruction of possible transmission events.[2] Clinical data (eg, co-infection with sexually transmitted infections (STI) etc) allows the ascertainment of risk factors associated with infection. Laboratory methods such as the serological testing algorithm for recent HIV seroconversion (STARHS)[3 4] can identify men who have sex with men (MSM) with recent HIV infection. Such men are likely to have heightened infectivity as a result of the elevated viral load shortly after infection.[5 6] Combining such data has the potential to enhance our understanding of HIV transmission and its associated risk factors to a greater extent than can be gleaned from each data source when considered individually.

Methods have combined laboratory and clinical data previously.[7 8] Pao *et al*[9] used phylogenetic analysis of HIV *pol* sequences to identify possible transmissions between MSM and, through linkage to clinical data, identified risk factors associated with transmission. Brenner *et al*[8] recently linked sequences to infection stage data in a phylogenetic analysis that highlighted the higher transmission potential from individuals who were recently HIV infected. Such studies remain exploratory, but demonstrate the benefits of combining data sources to enhance our understanding of HIV transmission.

Combined approaches could also be focused to examine specific public health problems. For example, assessing the effectiveness of voluntary confidential HIV testing (VCT) in preventing HIV transmission, which we explore in this paper. VCT promotion among MSM who may have had HIV exposure is an important part of the UK response to the HIV epidemic and remains a central component of best practice guidelines[10] and the sexual health strategy.[11] Prompt HIV diagnosis may prevent further HIV transmission through ensuring patients' viral load is low (through regular monitoring and/or administering antiretroviral therapy when clinically appropriate[12]) and providing earlier opportunities for partner notification and behaviour change counselling. Between 1996 and 2006 the proportion of MSM attendees of STI clinics receiving VCT rose from 45% to 85%.[13] However, during the same period, annual HIV incidence has remained at approximately 3%.[13] Therefore, the effectiveness of promoting VCT to prevent HIV transmission may be limited for reasons that are not fully understood.

We examine the benefits of integrating sequence and laboratory data derived from an unlinked anonymous (UA) survey of STI clinic attendees[14] to improve our understanding of the effectiveness of VCT in preventing HIV transmission. VCT uptake is compared between recently and non-recently HIV-infected MSM to assess differential uptake according to transmission risk. A phylogenetic analysis is conducted to explore the occurrence of HIV transmission from recently HIV-infected individuals before the diagnostic opportunity.

Sex Transm Infect 2009;85:4–9. doi:10.1136/sti.2008.031831

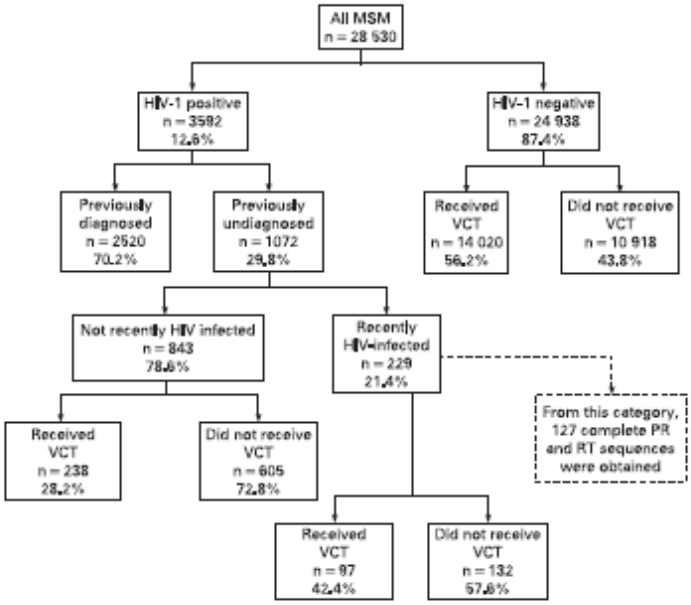


Figure 1 Flow diagram of all men who have sex with men (MSM) attending 15 sexually transmitted infection clinics: England, Wales and Northern Ireland, 1999–2002. PR, protease; RT, reverse transcriptase; VCT, voluntary confidential HIV testing.

METHODS

Study population

Data and sera were obtained from the UA survey between 1999 and 2002.[14] Established in 1989, this ongoing survey measures HIV prevalence (including undiagnosed prevalence) among attendees of sentinel STI clinics (15/232 clinics in England, Wales and Northern Ireland; seven in London, eight elsewhere). Such clinics provide advice, treatment and screening services for HIV/STI. Blood left over from syphilis testing is irreversibly unlinked from patient identifiers and anonymously HIV tested. Patients can only be included in the survey once per calendar quarter. Limited information is retained with each sample: risk group; age group; sex; calendar quarter of attendance; clinic; birth region; STI diagnosis; whether patient was HIV diagnosed before the clinic attendance or had received VCT during the clinic attendance.[15] The UA survey has had Public Health Laboratory Service ethical approval since its inception in 1989 and is compliant with the Human Tissue Act (2004).

Diagnosis status and HIV testing uptake

Patients found to be HIV positive through UA testing were categorised as "newly diagnosed" or "undiagnosed" according to whether they received VCT at the attendance (patients diagnosed with HIV before the clinic attendance were excluded). The "newly diagnosed" received VCT during the clinic attendance and the "undiagnosed" did not receive VCT, leaving the clinic remaining undiagnosed. T tests were used to supplement results when appropriate.

Serological testing algorithm for recent HIV seroconversion

All samples from MSM with a newly diagnosed or undiagnosed HIV infection were tested using STARHS.[8 9] STARHS distinguishes recent HIV infections through exploiting evolving antibody responses. The algorithm assumes anti-HIV titres increase at a similar rate between individuals over the months following infection: a less sensitive anti-HIV-1 enzyme immunoassay (EIA) is applied to test serum or plasma specimens confirmed to be anti-HIV-1 positive using a sensitive EIA. Reactivity below the cut-off (standardised optical density 1.0) in the less sensitive EIA indicates that the sample was derived from someone who was infected an average of 170 days (95% CI 162 to 183 days) before specimen collection.[16]

Phylogenetic analysis

Samples from recently HIV-infected individuals were genotyped as previously described.[17] The protease region of *pol* was sequenced along with the first 230 codons of reverse transcriptase (RT). Protease and RT sequences were aligned using sequence alignment version 2 across 998 nucleotides. Primary drug resistance mutation sites[18] were removed from the alignment to eliminate bias from convergent evolution. Phylogenetic analysis was performed using phylogenetic analysis using parsimony[19] with a neighbour-joining algorithm. The alignment was run through ModelTest to select the best fitting evolutionary model. A heuristic search was conducted for a maximum likelihood tree using the selected model (GTR + I + G) and its derived parameters (proportion of invariable sites 0.43 and gamma distribution 0.79) using the neighbour-joining tree as the starting tree. A bootstrap analysis

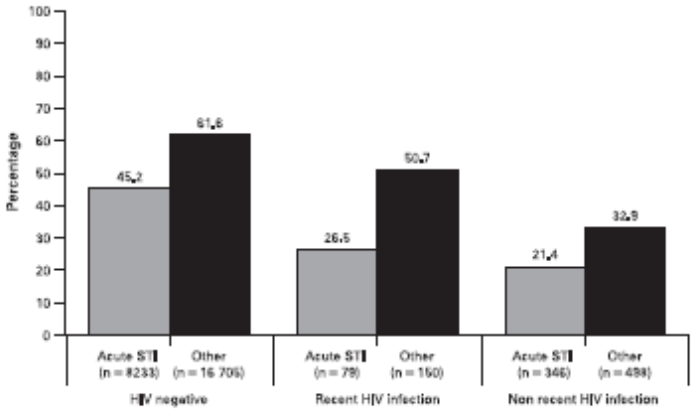


Figure 2 Proportion of men who have sex with men attending 15 sexually transmitted infection (STI) clinics receiving voluntary confidential HIV testing, by diagnosis status: England, Wales and Northern Ireland, 1999–2002.

(500 replicates) was used to obtain statistical support for branching patterns. Genetic distances were calculated from the consensus tree for each terminal cluster. Branches comprising two or more sequences with an average bootstrap value of over 99% and an average genetic distance under 0.015 nucleotide substitutions per site were considered to be "possible transmissions".[8]

RESULTS

Population characteristics

Between 1999 and 2002, 28 530 UA tests were conducted on samples from MSM attending 15 STI clinics in England, Wales and Northern Ireland (fig 1). Of these, 3592 samples were found to be HIV positive, of which 1072 were derived from patients with a previously undiagnosed infection (newly diagnosed or undiagnosed HIV infection). Among this group, 21% (229) of samples were derived from recently infected MSM. Complete protease and RT sequences were obtained from 127 samples from recently HIV-infected patients (the remaining 102 could not be amplified because of problems associated with using residual samples).

Overall, of the 229 samples from recently HIV-infected MSM, 86% (196) attended clinics in London, 56% (127) were UK born and 15% (35) were born elsewhere in Europe. For the 127 linkable to sequences, equivalent figures were 86% (109), 58% (74) and 15% (19), respectively.

VCT uptake

VCT uptake was higher among MSM from whom HIV-negative samples were collected than compared with samples from MSM with previously undiagnosed HIV infection: 56% (14 020/24 938, 95% CI 55.6 to 56.8) versus 31% (335/1072, 95% CI 28.5 to 34.1; p<0.001) (fig 1). Forty-two per cent (97/229, 95% CI 36.1 to 48.8%) of MSM with samples indicating recent HIV infection received VCT. This compares with 28% (238/843, 95% CI 25.3 to 31.3; p<0.001) among samples from previously undiagnosed HIV-infected MSM with a non-recent infection.

Figure 2 demonstrates the differential VCT uptake according to transmission risk. Regardless of HIV infection, all MSM were less likely to receive VCT if they had an STI: uptake was 22% (95/425, 95% CI 18.7 to 26.6) among the previously undiagnosed HIV infected and 45% (3721/8233, 95% CI 44.1 to 46.3) among the HIV negative (p<0.001).

Phylogenetic analysis

Of the 127 sequences from samples from recently HIV-infected MSM, 16 clustered with at least one other sequence with a bootstrap support of at least 99% and a genetic distance under 0.015 nucleotide substitutions per site (fig 3). These formed seven clusters (A–G): cluster A contained four sequences and clusters B–G each consisted of sequence pairs. Cluster B consisted of sequences obtained in Wales and the remaining six were from London. Clusters A–D and F contained sequences from samples obtained within the same, or successive, calendar quarters.

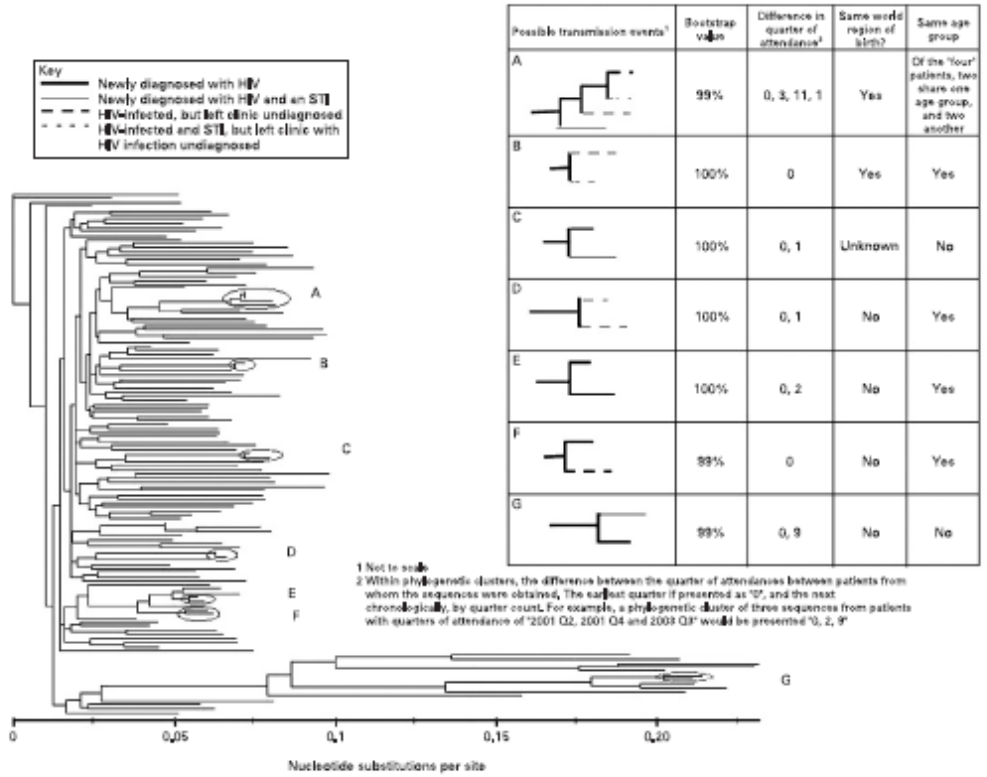
Clusters C, E and G were composed entirely of sequences from individuals whose HIV infection was diagnosed at that STI clinic attendance (fig 3). Clusters B and D comprised sequences from individuals who left the clinic remaining unaware of their HIV infection. Eight sequences that fell into a cluster were derived from patients who had an STI.

DISCUSSION

More than half of recently HIV-infected MSM and nearly 80% of all undiagnosed HIV-infected MSM with an STI, attending clinics collaborating in the UA survey, did not receive VCT. Phylogenetic analyses of HIV *pol* sequences from recently HIV-infected MSM identified at least seven possible instances of HIV transmission before the diagnostic opportunity.

VCT uptake

Combining VCT data with laboratory algorithms demonstrated that using VCT uptake as a prevention indicator lacks sensitivity in discriminating between groups with varied transmission risks. The diagnostic opportunity was missed among the recently HIV infected and those with an STI. We speculate that this could be for several reasons. First, recently HIV-infected individuals may have previously tested HIV negative and consequently believed themselves still to be HIV negative. Second, a concern that testing within a 3-month



Possible transmission events[1]	Bootstrap value	Difference in quarter of attendance[2]	Same world region of birth?	Same age group
A	99%	0, 3, 11, 1	Yes	Of the 'four' patients, two share one age group, and two another
B	100%	0	Yes	Yes
C	100%	0, 1	Unknown	No
D	100%	0, 1	No	Yes
E	100%	0, 2	No	Yes
F	99%	0	No	Yes
G	99%	0, 9	No	No

1 Not to scale
2 Within phylogenetic clusters, the difference between the quarter of attendances between patients from whom the sequences were obtained. The earliest quarter if presented as '0', and the next chronologically, by quarter count. For example, a phylogenetic cluster of three sequences from patients with quarters of attendance of '2001 Q2, 2001 Q4 and 2003 Q3' would be presented '0, 2, 9'

Figure 3 Possible transmission events among recently HIV-infected patients attending 15 sexually transmitted infection (STI) clinics: England, Wales and Northern Ireland, 1999–2002.

window period would affect diagnostic accuracy[16] may have led to test deferral among the recently HIV infected. Third, testing algorithms may be failing to pick up those at risk. Fourth, a clinical focus on STI management may have detracted from those with an STI receiving VCT. Finally, there may exist a subset of MSM who are serial VCT refusers.

It is only known whether VCT was received; it is not known whether and why VCT was not offered, or was refused. No data were collected concerning past HIV testing history. Furthermore, only patients undergoing syphilis tests are included: a population not necessarily representative of the clinic population. The survey design may mean a small subset of patients are included many times within the dataset. The size and attributes of this group are not known, but if they are associated with a particular risk or HIV testing behaviours, this may affect data accuracy.

Data are taken from 1999–2002, a period in which VCT uptake was lower among MSM. A British Association for Sexual Health and HIV (BASHH) audit demonstrated a recent rise in the proportion of STI clinic attendees being HIV tested (as a result of better targeted HIV testing efforts (including patients with STI) and more assertive testing invitations (eg, opt-out policies).[20] Although this increased VCT uptake is reflected within the UA survey, which found that VCT uptake increased to 85%[13] overall in 2006, in the same year VCT was measured at 42% among HIV-infected MSM with an STI and 63% among the recently HIV infected. Therefore, despite an increased VCT uptake generally, a substantial proportion of those at highest risk of onward transmission still leave the clinic remaining undiagnosed.

Transmission before diagnostic opportunity

The phylogenetic analysis was conducted to observe instances in which transmission events may have occurred before the diagnostic opportunity. No attempts were made to calculate the force of transmission from the recently HIV-infected population because the size and inconsistent geographical make-up of the dataset prohibits this, and indeed any extrapolation of results to the HIV-infected population.

We observed seven distinct possible HIV transmission clusters between recently HIV-infected MSM. Five clusters had closeness in the likely seroconversion dates between the sequences involved. This demonstrates the potential for HIV transmission to occur rapidly from recently HIV-infected MSM, even among those who attend clinics soon after infection. Combining

Key messages

- Between 1999 and 2002, VCT uptake was higher among HIV-negative MSM, compared with HIV-infected MSM. Low VCT uptake was observed among recently HIV-infected MSM.
- Phylogenetic reconstruction of HIV transmission events among recently HIV-infected MSM indicate that transmissions may have occurred before the opportunity for VCT.
- VCT promotion alone may be limited in its ability to prevent HIV transmission.
- Integrating clinical, laboratory and sequence-based HIV data can enhance our understanding of HIV transmission.

sequences with VCT data demonstrated that approximately half of the MSM involved in a possible transmission went on to miss the diagnostic opportunity at their clinic attendance, increasing the risk of further transmission and limiting the opportunity for prompt partner notification.

It cannot be conclusively demonstrated that the transmission clusters identified represent actual transmissions events. The bootstrapping and genetic distance cut-offs used are not definitive but have been shown to reduce the proportion of false-positive clusters among phylogenetic analyses of UK HIV sequences.[31] The analysis cannot rule out that each cluster could have been generated by a third party (not sampled) infecting each sequence in the cluster, or acting as an intermediary between the sequences. Similarly, transmissions from individuals in the dataset may not be represented if the infection generated was not in the study.

The survey methodology allows patients to be sampled (within the same clinic) up to four times annually, but not within the same calendar quarter. Therefore, theoretically, the transmission clusters identified could have been formed between sequences from the same individual, sampled many times. Furthermore, the UA survey is necessarily designed to prevent the identification of individuals. However, five of the possible transmissions (C–G) involve patient pairs who have a different birth region or age group (fig 3), making it unlikely that the sequences come from the same person. Similarly, there must be at least two individuals included in cluster A—but not necessarily four. The patients in cluster B both attended the same clinic within the same calendar quarter, meaning they should be different patients.

Implications for HIV testing policy

BASHH guidelines (2006) recommend that all patients should be offered VCT on their first clinical attendance,[10] including patients presenting with STI. Whereas the benefits of VCT are unquestionable for clinical purposes, the public health benefits require further consideration. Since 2002, HIV testing policy has not been adapted specifically to target patients with recent risk exposure.[10] BASHH guidelines recommend that recently HIV-exposed patients receive two HIV tests: one immediately and one 3 months after exposure (the "3-month rule"). This is because of concerns that HIV testing shortly after exposure may not yield accurate results. However, recent improvements in HIV test sensitivity mean that the latest (so-called "fourth generation" tests) can detect both anti-HIV and HIV p24 antigen within 4 weeks of infection.[33] We believe that delaying the second test for 3 months may be detrimental. The BASHH audit found the most commonly cited reason for not receiving VCT was deferral because of concerns relating to the accuracy of HIV testing shortly after exposure;[10] the rate of reattendance for the second test was low. We suggest the guidelines should be modified to emphasise the importance of undertaking VCT at first attendance. If the first test is negative, the second test should be brought forward to 1 month post-attendance (provided fourth generation tests are employed) to enable earlier interventions to reduce onward transmission.

Our phylogenetic analysis illustrates the potential for the recently HIV infected to generate rapid onward transmission, even among those who receive VCT soon after infection. Consequently, VCT alone may not have a large impact on the greater transmission potential of recently HIV-infected MSM. Alternative strategies include: more rigorous partner notification; post-exposure prophylaxis among recently exposed MSM; encouraging frequent, regular testing; educating MSM and healthcare providers about seroconversion illness.

This analysis demonstrates how the integration of clinical, laboratory and sequence-based data can be used to explore HIV transmission and related public health problems more extensively. Although the analysis has not measured the force of infection from the recently HIV-infected population, or the impact of VCT on HIV transmission, it highlighted the need for specific targeting of the recently exposed and those with STI for VCT. It also revealed that VCT alone may be limited in its ability to prevent HIV transmission.

Acknowledgements: The authors extend many thanks to their colleagues at the 15 participating STI clinics and associated laboratories, without whom this work would not have been possible.

Funding: This study was funded by the Department of Health, award ref UAS 6/1, for which the authors are very grateful.

Competing interests: None.

Ethics approval: The UA survey has had Public Health Laboratory Service ethical approval since its inception in 1999 and is compliant with the Human Tissue Act (2004).

The manuscript was submitted before the publication of updated BASHH HIV testing guidelines in September 2008. These guidelines can be found at: http://www.bashh.org/documents/1808.

REFERENCES

1. **Pozniak A**, Gazzard B, Anderson J, *et al*. British HIV Association (BHIVA) guidelines for the treatment of HIV-infected adults with antiretroviral therapy. *HIV Med* 2003;4(Suppl 1):1–41.
2. **Hue S**, Clewley JP, Cane PA, *et al*. HIV-1 *pol* gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 2004;18:719–28.
3. **Murphy G**, Charlett A, Jordan JF, *et al*. HIV incidence appears constant in men who have sex with men despite widespread use of effective antiretroviral therapy. *AIDS* 2004;18:265–72.
4. **Janssen RS**, Satten GA, Stramer SL, *et al*. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA* 1998;280:42–8.
5. **Wawer MJ**, Gray RH, Sewankambo NK, *et al*. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis* 2005;191:1403–9.
6. **Gray RH**, Wawer MJ, Brookmeyer R, *et al*. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *Lancet* 2001;357:1149–53.
7. **Pao D**, Fisher M, Hue S, *et al*. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS* 2005;19:85–90.
8. **Hue S**, Pillay D, Clewley JP, *et al*. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A* 2005;102:4425–9.
9. **Brenner BG**, Roger M, Routy JP, *et al*. High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis* 2007;195:951–9.
10. **British Association for Sexual Health and HIV British Infection Society**. UK National Guidelines for HIV Testing. British HIV Association, 2008.
11. **Department of Health**. *The national strategy for sexual health and HIV. 2001*. London. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4003133?IdcService=GET_FILE&dID=5538&Rendition=Web/ (accessed Dec 2008).

12. Ledergerber B, Egger M, Opravil M, *et al*. Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. Swiss HIV Cohort Study. *Lancet* 1999;353:863–8.
13. **The UK Collaborative Group for HIV and STI Surveillance**. *Testing times: HIV and other sexually transmitted infections in the United Kingdom: 2007*. London: Health Protection Agency, Centre for Infections, November 2007.
14. Catchpole MA, McGarrigle CA, Rogers PA, *et al*. Serosurveillance of prevalence of undiagnosed HIV-1 infection in homosexual men with acute sexually transmitted infection. *BMJ* 2000;321:1319–20.
15. Brown AE, Tomkins SE, Logan IE, *et al*. Monitoring the effectiveness of HIV and STI prevention initiatives in England, Wales, and Northern Ireland: Where are we now? *Sex Transm Infect* 2006;82:4–10.
16. Murphy G, Charlett A, Osmer N, *et al*. Reconciling HIV incidence results from two assays employed in the serological testing algorithm for recent HIV seroconversion (STARHS). *J Virol Methods* 2003;113:79–86.
17. Tatt ID, Barlow KL, Clewley JP, *et al*. Surveillance of HIV-1 subtypes among heterosexuals in England and Wales, 1997–2000. *J Acquir Immune Defic Syndr* 2004;36:1092–9.
18. Shafer RW, Rhee SY, Pillay D, *et al*. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS* 2007;21:215–23.
19. Swofford DL. *PAUP 4.0: Phylogenetic analysis using parsimony (and other methods)*. Version 4. Sunderland, MA: Sinaur Associates, 2001.
20. Murare HL, Lowndes CM, Daniels DS, *et al*. National study of HIV testing in men who have sex with men attending genitourinary clinics in the United Kingdom. *Sex Transm Inf* 2008;84:265–70.
21. Hue S, Clewley JP, Cane PA, *et al*. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 2004;18:719–28.
22. Busch MP, Lee LL, Satten G, *et al*. Time course of detection of viral and serologic markers preceding human immunodeficiency virus type 1 seroconversion: implications for screening of blood and tissue donors. *Transfusion* 1995;35:91–7.

13 References

Adelisa L. Panlilio, D. M. C., Lisa A. Grohskopf, (2005). Updated U.S. Public Health Service Guidelines for the Management of Occupational Exposures to HIV and Recommendations for Postexposure Prophylaxis. C. MMWR.

Adler, M. W. (1987). "ABC of AIDS. Range and natural history of infection." Br Med J (Clin Res Ed) **294**(6580): 1145-7.

Attia, S., M. Egger, et al. (2009). "Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis." Aids **23**(11): 1397-404.

Barre-Sinoussi, F., J. C. Chermann, et al. (1983). "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." Science **220**(4599): 868-71.

Barreiro, P., J. del Romero, et al. (2006). "Natural pregnancies in HIV-serodiscordant couples receiving successful antiretroviral therapy." J Acquir Immune Defic Syndr **43**(3): 324-6.

Beinker, N. K., D. L. Mayers, et al. (2001). "Genotypic drug resistance and cause of death in HIV-infected persons who died in 1999." J Acquir Immune Defic Syndr **28**(3): 250-3.

Bennett, D. E., R. J. Camacho, et al. (2009). "Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update." PLoS One **4**(3): e4724.

Bernard, E., Azed, Y, Vandamme, AM, Geretti, AM (2007). HIV Forensics: The use of phylogenetic analysis as evidence in criminal investigation of HIV transmission, NAM.

Berridge, V. (1996). AIDS in the UK, the making of policy' 1981-1994, Oxford University Press.

Bezemer, D., A. de Ronde, et al. (2006). "Evolution of transmitted HIV-1 with drug-resistance mutations in the absence of therapy: effects on CD4+ T-cell count and HIV-1 RNA load." Antivir Ther **11**(2): 173-8.

Bhaskaran, K., O. Hamouda, et al. (2008). "Changes in the risk of death after HIV seroconversion compared with mortality in the general population." Jama **300**(1): 51-9.

BHIVA (2006). "British Association for Sexual Health and HIV British Infection Society. UK National Guidelines for HIV Testing. British HIV Association, 2006."

BHIVA (2008). "British Association for Sexual Health and HIV British Infection Society. UK National Guidelines for HIV Testing 2008."

Blattner, W., R. C. Gallo, et al. (1988). "HIV causes AIDS." Science **241**(4865): 515-6.

Brennan, R. O. and D. T. Durack (1981). "Gay compromise syndrome." Lancet **2**(8259): 1338-9.

Brenner, B. G., M. Roger, et al. (2008). "Transmission networks of drug resistance acquired in primary/early stage HIV infection." Aids **22**(18): 2509-15.

Brenner, B. G., M. Roger, et al. (2007). "High rates of forward transmission events after acute/early HIV-1 infection." J Infect Dis **195**(7): 951-9.

Brown, A. E., Gifford, R.J., et al. (2009a). "Phylogenetic Reconstruction of Transmission Events from Individuals with Acute HIV Infection: Toward More-Rigorous Epidemiological Definitions." Journal of Infectious Disease **199**(3): 427-431.

Brown, A. E., G. Murphy, et al. (2009b). "Implications for HIV testing policy derived from combining data on voluntary confidential testing with viral sequences and serological analyses." Sex Transm Infect **85**(1): 4-9.

Brown, A. E., S. E. Tomkins, et al. (2006). "Monitoring the effectiveness of HIV and STI prevention initiatives in England, Wales, and Northern Ireland: where are we now?" Sex Transm Infect **82**(1): 4-10.

Brust, S., H. Duttmann, et al. (2000). "Shortening of the diagnostic window with a new combined HIV p24 antigen and anti-HIV-1/2/O screening test." J Virol Methods **90**(2): 153-65.

Burchell, A. N., L. Calzavara, et al. (2003). "Symptomatic primary HIV infection or risk experiences? Circumstances surrounding HIV testing and diagnosis among recent seroconverters." Int J STD AIDS **14**(9): 601-8.

Busch, M. P. and A. M. Courouce (1997). "Relative sensitivity of United States and European assays for screening blood donors for antibodies to human immunodeficiency viruses." Transfusion **37**(3): 352-4.

Busch, M. P., S. A. Glynn, et al. (2005). "Relative sensitivities of licensed nucleic acid amplification tests for detection of viremia in early human immunodeficiency virus and hepatitis C virus infection." Transfusion **45**(12): 1853-63.

Calzavara, L., A. N. Burchell, et al. (2002). "Increases in HIV incidence among men who have sex with men undergoing repeat diagnostic HIV testing in Ontario, Canada." Aids **16**(12): 1655-61.

Cane, P., I. Chrystie, et al. (2005). "Time trends in primary resistance to HIV drugs in the United Kingdom: multicentre observational study." Bmj **331**(7529): 1368.

CASCADE. (2009). Retrieved 09/08/09, from http://www.ctu.mrc.ac.uk/cascade/study_objectives.asp.

Castilla, J., J. Del Romero, et al. (2005). "Effectiveness of highly active antiretroviral therapy in reducing heterosexual transmission of HIV." J Acquir Immune Defic Syndr **40**(1): 96-101.

Catchpole, M. A., C. A. McGarrigle, et al. (2000). "Serosurveillance of prevalence of undiagnosed HIV-1 infection in homosexual men with acute sexually transmitted infection." Bmj **321**(7272): 1319-20.

CDC. (2008). "CDC Underscores Current Recommendation for Preventing HIV Transmission." Retrieved 29th June 2009, 2009.

Chadborn, T. R., K. Baster, et al. (2005). "No time to wait: how many HIV-infected homosexual men are diagnosed late and consequently die? (England and Wales, 1993-2002)." Aids **19**(5): 513-20.

Chadborn, T. R., V. C. Delpech, et al. (2006). "The late diagnosis and consequent short-term mortality of HIV-infected heterosexuals (England and Wales, 2000-2004)." Aids **20**(18): 2371-9.

Chawla, A., G. Murphy, et al. (2007). "Human immunodeficiency virus (HIV) antibody avidity testing to identify recent infection in newly diagnosed HIV type 1 (HIV-1)-seropositive persons infected with diverse HIV-1 subtypes." J Clin Microbiol **45**(2): 415-20.

Clark, S. (2003). Mutations in retroviral genes associated with resistance., Group Los Alamos National Laboratory 2005

Clumeck, N., J. Sonnet, et al. (1984). "Acquired immunodeficiency syndrome in African patients." N Engl J Med **310**(8): 492-7.

Cohen, M. S., T. D. Mastro, et al. (2009). "Universal voluntary HIV testing and immediate antiretroviral therapy." Lancet **373**(9669): 1077; author reply 1080-1.

Crepaz, N., T. A. Hart, et al. (2004). "Highly active antiretroviral therapy and sexual risk behavior: a meta-analytic review." Jama **292**(2): 224-36.

Cu-Uvin, S., A. M. Caliendo, et al. (2000). "Effect of highly active antiretroviral therapy on cervicovaginal HIV-1 RNA." Aids **14**(4): 415-21.

De Cock, K. M., C. F. Gilks, et al. (2009). "Can antiretroviral therapy eliminate HIV transmission?" Lancet **373**(9657): 7-9.

de Oliveira, T., O. G. Pybus, et al. (2006). "Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak." Nature **444**(7121): 836-7.

Devereux, H. L., M. Youle, et al. (1999). "Rapid decline in detectability of HIV-1 drug resistance mutations after stopping therapy." Aids **13**(18): F123-7.

DH, D. o. H. (1987). Don't die of ignorance.

DH, D. o. H. (2001). The national strategy for sexual health and HIV. London.

Dodds, J. P., A. M. Johnson, et al. (2007). "A tale of three cities: persisting high HIV prevalence, risk behaviour and undiagnosed infection in community samples of men who have sex with men." Sex Transm Infect **83**(5): 392-6.

Dodds, J. P., D. E. Mercey, et al. (2004). "Increasing risk behaviour and high levels of undiagnosed HIV infection in a community sample of homosexual men." Sex Transm Infect **80**(3): 236-40.

Dougan, S., J. Elford, et al. (2007). "Does the recent increase in HIV diagnoses among men who have sex with men in the UK reflect a rise in HIV incidence or increased uptake of HIV testing?" Sex Transm Infect **83**(2): 120-5; discussion 125.

Dougan, S., V. L. Gilbart, et al. (2005). "HIV infections acquired through heterosexual intercourse in the United Kingdom: findings from national surveillance." Bmj **330**(7503): 1303-4.

Drummond, A. J., S. Y. Ho, et al. (2006). "Relaxed phylogenetics and dating with confidence." PLoS Biol **4**(5): e88.

Drummond, A. J. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees." BMC Evol Biol **7**: 214.

Dubois, R. M., Braitwaite, M.A., Mikhail, J.R. et al., (1981). "'Primary Pneumocystis Carinii and Cytomegalovirus Infections'." Lancet **ii**: 1339.

Efron, B., E. Halloran, et al. (1996). "Bootstrap confidence levels for phylogenetic trees." Proc Natl Acad Sci U S A **93**(14): 7085-90.

Elford, J., G. Bolding, et al. (2007). "Barebacking among HIV-positive gay men in London." Sex Transm Dis **34**(2): 93-8.

Elford, J. and G. Hart (2005). "HAART, viral load and sexual risk behaviour." Aids **19**(2): 205-7.

EuropeanCollaborativeStudy (2005). "Mother-to-child transmission of HIV infection in the era of highly active antiretroviral therapy." Clin Infect Dis **40**(3): 458-65.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J Mol Evol **17**(6): 368-76.

Felsenstein, J. (1996). "Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods." Methods Enzymol **266**: 418-27.

Fenton, K. A., C. H. Mercer, et al. (2005). "Ethnic variations in sexual behaviour in Great Britain and risk of sexually transmitted infections: a probability survey." Lancet **365**(9466): 1246-55.

Fideli, U. S., S. A. Allen, et al. (2001). "Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa." AIDS Res Hum Retroviruses **17**(10): 901-10.

Fiebig, E. W., D. J. Wright, et al. (2003). "Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection." Aids **17**(13): 1871-9.

Fiscus, S. A., C. D. Pilcher, et al. (2007). "Rapid, real-time detection of acute HIV infection in patients in Africa." J Infect Dis **195**(3): 416-24.

Fox, J., P. J. White, et al. (2009). "Reductions in HIV transmission risk behaviour following diagnosis of primary HIV infection: a cohort of high-risk men who have sex with men." HIV Med **10**(7): 432-8.

Friedland, G. H. and A. Williams (1999). "Attaining higher goals in HIV treatment: the central importance of adherence." Aids **13 Suppl 1**: S61-72.

Gaines, H., M. von Sydow, et al. (1988). "Detection of immunoglobulin M antibody in primary human immunodeficiency virus infection." Aids **2**(1): 11-5.

Garcia-Gasco, P., I. Maida, et al. (2008). "Episodes of low-level viral rebound in HIV-infected patients on antiretroviral therapy: frequency, predictors and outcome." J Antimicrob Chemother.

Gazzard, B. G. (2008). "British HIV Association Guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy 2008." HIV Med **9**(8): 563-608.

Gifford, R. J. (2007). "Phylogenetic Surveillance of Viral Genetic Diversity and the Evolving Molecular Epidemiology of HIV-1." J Virol.

Gilbart, V. L., B. G. Evans, et al. (2004). "HIV transmission among men who have sex with men through oral sex." Sex Transm Infect **80**(4): 324.

Gill, O. N., M. W. Adler, et al. (1989). "Monitoring the prevalence of HIV." Bmj **299**(6711): 1295-8.

Granich, R. M., C. F. Gilks, et al. (2009). "Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model." Lancet **373**(9657): 48-57.

Grant, R. M., F. M. Hecht, et al. (2002). "Time trends in primary HIV-1 drug resistance among recently infected persons." Jama **288**(2): 181-8.

Gray, R. H., M. J. Wawer, et al. (2001). "Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda." Lancet **357**(9263): 1149-53.

Gurtler, L. (1996). "Difficulties and strategies of HIV diagnosis." Lancet **348**(9021): 176-9.

Gurtler, L. G., L. Zekeng, et al. (1996). "HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV." Arch Virol Suppl **11**: 195-202.

Hammer, S. M., J. J. Eron, Jr., et al. (2008). "Antiretroviral treatment of adult HIV infection: 2008 recommendations of the International AIDS Society-USA panel." Jama **300**(5): 555-70.

Hasegawa, M., H. Kishino, et al. (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." J Mol Evol **22**(2): 160-74.

Hillis D, M. (1996). Molecular Systematics. Sunderland, Massachussets, Sinauer Associates.

Hirsch, M. S., F. Brun-Vezinet, et al. (2003). "Antiretroviral drug resistance testing in adults infected with human immunodeficiency virus type 1: 2003 recommendations of an International AIDS Society-USA Panel." Clin Infect Dis **37**(1): 113-28.

Ho, D. D., A. U. Neumann, et al. (1995). "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection." Nature **373**(6510): 123-6.

Holmes, P. R. D. M. a. E. C. (1998). Molecular Evolution: a phylogenetic approach. Oxford, Blackwell Publishing.

HPA (2008). HIV in the UK 2007. London.

Hue, S., J. P. Clewley, et al. (2004). "HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy." Aids **18**(5): 719-28.

Hue, S., D. Pillay, et al. (2005). "Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups." Proc Natl Acad Sci U S A **102**(12): 4425-9.

Huelsenbeck, J. P. (1995). "The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining." Mol Biol Evol **12**(5): 843-9.

Janssen, R. S., G. A. Satten, et al. (1998). "New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes." Jama **280**(1): 42-8.

Johnson, J. A., J. F. Li, et al. (2008). "Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy." PLoS Med **5**(7): e158.

Johnson, V. A., F. Brun-Vezinet, et al. (2008). "Update of the Drug Resistance Mutations in HIV-1: Spring 2008." Top HIV Med **16**(1): 62-8.

Jukes, T. a. C. C. (1969). Evolution of protein molecules., Academic Press.

Kahn, J. O. and B. D. Walker (1998). "Acute human immunodeficiency virus type 1 infection." N Engl J Med **339**(1): 33-9.

Kamradt, T., D. Niese, et al. (1985). "Slim disease (AIDS)." Lancet **2**(8469-70): 1425.

Kantor, R., D. A. Katzenstein, et al. (2005). "Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration." PLoS Med **2**(4): e112.

Kearney, M., F. Maldarelli, et al. (2009). "Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals." J Virol **83**(6): 2715-27.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol **16**(2): 111-20.

Lanave, C., G. Preparata, et al. (1984). "A new method for calculating evolutionary substitution rates." J Mol Evol **20**(1): 86-93.

Lane, T., A. Pettifor, et al. (2006). "Heterosexual anal intercourse increases risk of HIV infection among young South African men." Aids **20**(1): 123-5.

Le Vu, S., J. Pillonel, et al. (2008). "Principles and uses of HIV incidence estimation from recent infection testing--a review." Euro Surveill **13**(36).

Ledergerber, B., M. Egger, et al. (1999). "Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. Swiss HIV Cohort Study." Lancet **353**(9156): 863-8.

Leigh Brown, A. J., S. D. Frost, et al. (2003). "Transmission fitness of drug-resistant human immunodeficiency virus and the prevalence of resistance in the antiretroviral-treated population." J Infect Dis **187**(4): 683-6.

Leitner, T., D. Escanilla, et al. (1996). "Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis." Proc Natl Acad Sci U S A **93**(20): 10864-9.

Leitner, T. A., J (2000). "Reconstruction of HIV-1 transmission chains for forensic purposes." AIDS Rev **2**: 241-51.

Lewis, F., G. J. Hughes, et al. (2008). "Episodic sexual transmission of HIV revealed by molecular phylodynamics." PLoS Med **5**(3): e50.

Li, W. H. (1993). "So, what about the molecular clock hypothesis?" Curr Opin Genet Dev **3**(6): 896-901.

Little, S. J., S. Holte, et al. (2002). "Antiretroviral-drug resistance among patients recently infected with HIV." N Engl J Med **347**(6): 385-94.

Louwagie, J., F. E. McCutchan, et al. (1993). "Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes." Aids **7**(6): 769-80.

Lucas, G. M., R. E. Chaisson, et al. (1999). "Highly active antiretroviral therapy in a large urban clinic: risk factors for virologic failure and adverse drug reactions." Ann Intern Med **131**(2): 81-7.

Mann, J. M., K. Bila, et al. (1986). "Natural history of human immunodeficiency virus infection in Zaire." Lancet **2**(8509): 707-9.

Masquelier, B., K. Bhaskaran, et al. (2005). "Prevalence of transmitted HIV-1 drug resistance and the role of resistance algorithms: data from seroconverters in the CASCADE collaboration from 1987 to 2003." J Acquir Immune Defic Syndr **40**(5): 505-11.

Masur, H., M. A. Michelis, et al. (1981). "An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction." N Engl J Med **305**(24): 1431-8.

Mau, B., M. A. Newton, et al. (1999). "Bayesian phylogenetic inference via Markov chain Monte Carlo methods." Biometrics **55**(1): 1-12.

MMWR (1982). "Unexplained Immunodeficiency and Opportunistic Infections in Infants- New York, New Jersey, California." MMWR Weekly **31**(49): 665-667.

Mocroft, A., S. Vella, et al. (1998). "Changing patterns of mortality across Europe in patients infected with HIV-1. EuroSIDA Study Group." Lancet **352**(9142): 1725-30.

Montaner, J. S., P. Reiss, et al. (1998). "A randomized, double-blind trial comparing combinations of nevirapine, didanosine, and zidovudine for HIV-infected patients: the INCAS Trial. Italy, The Netherlands, Canada and Australia Study." Jama **279**(12): 930-7.

MRC-Resistance-Database. (2009). Retrieved 09/08/09, 2009, from http://www.ctu.mrc.ac.uk/research_areas/study_details.aspx?s=13.

Munro, H. L. L., C.M., Daniels, D., Sullivan, A.K., Robinson, A. (2007). National BASHH/HPA study of HIV testing in men who have sex with men (MSM) attending genitourinary (GUM) clinics in the UK., Health Protection Agency Centre for Infections.

Murphy, G., A. Charlett, et al. (2004). "Is HIV incidence increasing in homo/bisexual men attending GUM clinics in England, Wales and Northern Ireland?" Commun Dis Public Health **7**(1): 11-4.

Murphy, G., A. Charlett, et al. (2004). "HIV incidence appears constant in men who have sex with men despite widespread use of effective antiretroviral therapy." Aids **18**(2): 265-72.

Murphy, G. and J. V. Parry (2008). "Assays for the detection of recent infections with human immunodeficiency virus type 1." Euro Surveill **13**(36).

Murphy, G., J. V. Parry, et al. (2001). "Test of HIV incidence shows continuing HIV transmission in homosexual/bisexual men in England and Wales." Commun Dis Public Health **4**(1): 33-7.

Musicco, M., A. Lazzarin, et al. (1994). "Antiretroviral treatment of men infected with human immunodeficiency virus type 1 reduces the incidence of heterosexual transmission. Italian Study Group on HIV Heterosexual Transmission." Arch Intern Med **154**(17): 1971-6.

Neely, M. N., L. Benning, et al. (2007). "Cervical shedding of HIV-1 RNA among women with low levels of viremia while receiving highly active antiretroviral therapy." J Acquir Immune Defic Syndr **44**(1): 38-42.

Nicoll, A., O. N. Gill, et al. (2000). "The public health applications of unlinked anonymous seroprevalence monitoring for HIV in the United Kingdom." Int J Epidemiol **29**(1): 1-10.

Novitsky, V., U. R. Smith, et al. (2002). "Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design?" J Virol **76**(11): 5435-51.

O'Connor, B. H., M. B. McEvoy, et al. (1983). "Kaposi's sarcoma/AIDS surveillance in the UK." Lancet **1**(8329): 872.

Osmanov, S., C. Pattou, et al. (2002). "Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000." J Acquir Immune Defic Syndr **29**(2): 184-90.

Page, R. D. M., Holmes, E C. (1998). Molecular Evolution A phylogenetic approach, Blackwell.

Palella, F. J., Jr., K. M. Delaney, et al. (1998). "Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators." N Engl J Med **338**(13): 853-60.

Palmer, S., D. Vuitton, et al. (2002). "Reverse transcriptase and protease sequence evolution in two HIV-1-infected couples." J Acquir Immune Defic Syndr **31**(3): 285-90.

Pantaleo, G., C. Graziosi, et al. (1993). "New concepts in the immunopathogenesis of human immunodeficiency virus infection." N Engl J Med **328**(5): 327-35.

Pao, D., M. Fisher, et al. (2005). "Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections." Aids **19**(1): 85-90.

Paraskevis, D., O. Pybus, et al. (2009). "Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach." Retrovirology **6**(1): 49.

Parekh, B. S., M. S. Kennedy, et al. (2002). "Quantitative detection of increasing HIV type 1 antibodies after seroconversion: a simple assay for detecting recent HIV infection and estimating incidence." AIDS Res Hum Retroviruses **18**(4): 295-307.

Parekh, B. S., C. P. Pau, et al. (2001). "Assessment of antibody assays for identifying and distinguishing recent from long-term HIV type 1 infection." AIDS Res Hum Retroviruses **17**(2): 137-46.

Parry, J. V., P. P. Mortimer, et al. (2003). "Towards error-free HIV diagnosis: guidelines on laboratory practice." Commun Dis Public Health **6**(4): 334-50.

Perrin, L., L. Kaiser, et al. (2003). "Travel and the spread of HIV-1 genetic variants." Lancet Infect Dis **3**(1): 22-7.

Perry, K. R., S. Ramskill, et al. (2008). "Improvement in the performance of HIV screening kits." Transfus Med **18**(4): 228-40.

Phillips, A. and P. Pezzotti (2004). "Short-term risk of AIDS according to current CD4 cell count and viral load in antiretroviral drug-naive individuals and those treated in the monotherapy era." Aids **18**(1): 51-8.

Pilcher, C. D., S. A. Fiscus, et al. (2005). "Detection of acute infections during HIV testing in North Carolina." N Engl J Med **352**(18): 1873-83.

Pilcher, C. D., G. Joaki, et al. (2007). "Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection." Aids **21**(13): 1723-30.

Pilcher, C. D., J. T. McPherson, et al. (2002). "Real-time, universal screening for acute HIV infection in a routine HIV counseling and testing population." Jama **288**(2): 216-21.

Pinkerton, S. D. (2007). "How many sexually-acquired HIV infections in the USA are due to acute-phase HIV transmission?" Aids **21**(12): 1625-9.

Porter, K., A. Babiker, et al. (2003). "Determinants of survival following HIV-1 seroconversion after the introduction of HAART." Lancet **362**(9392): 1267-74.

Poulsen, A. G., P. Aaby, et al. (1993). "HIV-2 infection in Bissau, West Africa, 1987-1989: incidence, prevalences, and routes of transmission." J Acquir Immune Defic Syndr **6**(8): 941-8.

Pozniak, A., B. Gazzard, et al. (2003). "British HIV Association (BHIVA) guidelines for the treatment of HIV-infected adults with antiretroviral therapy." HIV Med **4 Suppl 1**: 1-41.

Quinn, T. C., M. J. Wawer, et al. (2000). "Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group." N Engl J Med **342**(13): 921-9.

Reed, C., Branson, B., Janssen, R (2004). Interpreting STARHS Results for Individuals vs Estimation of HIV Incidence. Conference on Retroviruses and Opportunistic Infections San Franc Calif. .

Remis, R. S., M. Alary, et al. (2000). "HIV infection and risk behaviours in young gay and bisexual men." Cmaj **163**(1): 14-5.

Remis, R. S. and R. W. Palmer (2009). "Testing bias in calculating HIV incidence from the Serologic Testing Algorithm for Recent HIV Seroconversion." Aids **23**(4): 493-503.

Rhee, S. Y., M. J. Gonzales, et al. (2003). "Human immunodeficiency virus reverse transcriptase and protease sequence database." Nucleic Acids Res **31**(1): 298-303.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-4.

Schacker, T., A. C. Collier, et al. (1996). "Clinical and epidemiologic features of primary HIV infection." Ann Intern Med **125**(4): 257-64.

Shafer, R. W., S. Y. Rhee, et al. (2007). "HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance." Aids **21**(2): 215-23.

Sharp, P. M., E. Bailes, et al. (2001). "The origins of acquired immune deficiency syndrome viruses: where and when?" Philos Trans R Soc Lond B Biol Sci **356**(1410): 867-76.

Sheth, P. M., C. Kovacs, et al. (2009). "Persistent HIV RNA shedding in semen despite effective antiretroviral therapy." Aids **23**(15): 2050-4.

Simms, I., K. A. Fenton, et al. (2005). "The re-emergence of syphilis in the United Kingdom: the new epidemic phases." Sex Transm Dis **32**(4): 220-6.

Stephenson, J. M., J. Imrie, et al. (2003). "Is use of antiretroviral therapy among homosexual men associated with increased risk of transmission of HIV infection?" Sex Transm Infect **79**(1): 7-10.

Sturmer, M., W. Preiser, et al. (2004). "Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates." Aids **18**(16): 2109-13.

Sudarshi, D., D. Pao, et al. (2008). "Missed opportunities for diagnosing primary HIV infection." Sex Transm Infect **84**(1): 14-6.

Suligoi, B., S. Butto, et al. (2008). "Detection of recent HIV infections in African individuals infected by HIV-1 non-B subtypes using HIV antibody avidity." J Clin Virol **41**(4): 288-92.

Swofford, D. (2001). PAUP 4.0: Phylogenetic Analysis Using Parsimony (and other methods). Sunderland (MA), Sinaur Associates.

Tatt, I. D., K. L. Barlow, et al. (2004). "Surveillance of HIV-1 subtypes among heterosexuals in England and Wales, 1997-2000." J Acquir Immune Defic Syndr **36**(5): 1092-9.

Thompson, J. D., T. J. Gibson, et al. (2002). "Multiple sequence alignment using ClustalW and ClustalX." Curr Protoc Bioinformatics **Chapter 2**: Unit 2 3.

Thompson, J. D., T. J. Gibson, et al. (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res **25**(24): 4876-82.

Tovanabutra, S., V. Robison, et al. (2002). "Male viral load and heterosexual transmission of HIV-1 subtype E in northern Thailand." J Acquir Immune Defic Syndr **29**(3): 275-83.

Turner, B. G. and M. F. Summers (1999). "Structural biology of HIV." J Mol Biol **285**(1): 1-32.

UK-CHIC (2004). "The UK Collaborative HIV cohort Steering Committee The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study." HIV Medicine **5**: 115-24.

UKCollaborativeGroup (2007). "Evidence of a decline in transmitted HIV-1 drug resistance in the United Kingdom." Aids **21**(8): 1035-9.

UNAIDS (2008). 2008 Report on the global AIDS epidemic. J. U. N. P. o. H. A. U. 2008.

UNAIDS. (2008). "Antiretroviral therapy and sexual transmission of HIV." Retrieved 29th June 2009, 2009, from http://data.unaids.org/pub/PressStatement/2008/080201_hivtransmission_en.pdf.

van Griensven, F., J. W. de Lind van Wijngaarden, et al. (2009). "The global epidemic of HIV infection among men who have sex with men." Curr Opin HIV AIDS **4**(4): 300-7.

Van Laethem, K., A. De Luca, et al. (2002). "A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients." Antivir Ther **7**(2): 123-9.

Vernazza, P., B. Hirschel, et al. (2008). "Les personnes séropositives ne souffrant d'aucune autre MST et suivant un traitment antirétroviral efficace ne transmettent pas le VIH par voie sexuelle." Bulletin des médecins suisses **89**(5): 165-169.

Vernazza, P. L., L. Troiani, et al. (2000). "Potent antiretroviral treatment of HIV-infection results in suppression of the seminal shedding of HIV. The Swiss HIV Cohort Study." Aids **14**(2): 117-21.

Wasserheit, J. N. (1992). "Epidemiological synergy. Interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases." Sex Transm Dis **19**(2): 61-77.

Watts, J. M., K. K. Dang, et al. (2009). "Architecture and secondary structure of an entire HIV-1 RNA genome." Nature **460**(7256): 711-6.

Wawer, M. J., R. H. Gray, et al. (2005). "Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda." J Infect Dis **191**(9): 1403-9.

Wensing, A. M., D. A. van de Vijver, et al. (2005). "Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management." J Infect Dis **192**(6): 958-66.

Wiens, J. J. and M. R. Servedio (1998). "Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods." Syst Biol **47**(2): 228-53.

Wilson, D. P., M. G. Law, et al. (2008). "Relation between HIV viral load and infectiousness: a model-based analysis." Lancet **372**(9635): 314-20.

Yang, Z. (1996). "Phylogenetic analysis using parsimony and likelihood methods." J Mol Evol **42**(2): 294-307.

Yerly, S., T. Junier, et al. (2009). "The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection." Aids.

Yerly, S., L. Kaiser, et al. (1999). "Transmission of antiretroviral-drug-resistant HIV-1 variants." Lancet **354**(9180): 729-33.

Zaccarelli, M., V. Tozzi, et al. (2005). "Multiple drug class-wide resistance associated with poorer survival after treatment failure in a cohort of HIV-infected patients." Aids **19**(10): 1081-9.

Zeniga, J. M., Whiteside, A., Ghaziani, A., Bartlett, J. (2008). A Decade of HAART: The Development and Global Impact of Highly Active Antiretroviral Therapy, OUP Oxford;.

Zhang, H., G. Dornadula, et al. (1998). "Human immunodeficiency virus type 1 in the semen of men receiving highly active antiretroviral therapy." N Engl J Med **339**(25): 1803-9.