# Analysis of Serological Surveillance Data: Patterns of Blood Borne Infection and Heterogeneity in People who Inject Drugs

## Thesis

# Analysis of serological surveillance data: patterns of blood-borne infection and heterogeneity in people who inject drugs

Ross J. Harris, MSc
Statistics Unit
Public Health England
London NW9 5EQ
ross.harris@phe.gov.uk

Submitted for the degree of Doctor of Philosophy
School of Mathematics and Statistics

Supported by Public Health England

Submitted for examination September 2018

# Abstract

A large proportion of blood-borne viruses (BBV) are transmitted via injecting drug use. Understanding patterns of risk and monitoring trends over time in people who inject drugs (PWID) is therefore a crucial part of developing public health policy.

Risks of infection with hepatitis C and B (HCV, HBV) and HIV in PWID are investigated using serial cross-sectional surveillance data, in particular via force of infection (FOI) models. Standard models are extended to include a wealth of covariate data. Individual variability is considered via bivariate shared frailty models and correlations between infections. Gamma, inverse Gaussian and time-varying frailty models are fitted to investigate how variability in risk evolves throughout injecting career. Finally, models are extended to the trivariate case and different forms of component frailty models are proposed.

Recent initiates were found to be at high risk of infection, in particular in London and the North West. Subsequently the FOI is broadly constant, and similar across different regions. Frailty models indicate that there is substantial individual variability in risk, although this declines over the course of injecting career. Females have higher overall risks of HCV and HBV, but are less heterogeneous than males. Including covariate data on demographics and a number of risk factors only resulted in modest reductions in frailty variance. Correlations between HCV-HBV and HBV-HIV associations were stronger than for HCV-HIV; trivariate models including additional pairwise components for HCV-HBV and HBV-HIV provided an improvement in model fit compared to a shared frailty model.

Many of the results in this thesis point towards greater variation in risk at initiation, potentially due to the varied circumstances in which individuals start injecting, followed by more comparable risks in those with established injecting behaviour. The relative importance of injecting and sexual risk may explain patterns of risk and correlation in the three infections.

# Acknowledgements

I would like to thank my supervisors, Prof Paddy Farrington, Dr Daniela De Angelis and Dr Heather Whitaker for their support and guidance over the past six years. I am extremely grateful for their insights on this work, their enthusiasm, and their patience. With their encouragement, I have learned to do things that I did not think I was capable of.

I would also like to thank Dr Vivian Hope for many useful discussions on the behaviour of people who inject drugs and the epidemiology of blood-borne viruses in this group. Viv's deep and wide-ranging expertise in this field is unmatched, and for many years he was the driving force behind the UAM study on which this thesis is based. I am also grateful to the UAM study coordinators, and the individuals who took part in the survey, without whom this invaluable public health resource would not exist.

I am deeply indebted to Public Health England for the opportunity to pursue my studies and to my head of department, Dr André Charlett, my line manager, Daniela, and other colleagues for giving me the time and space to work on my thesis when needed. Although studying part-time must always present challenges, I cannot imagine better circumstances in which to have done so.

Thank you to my parents, John and Debbie, for their encouragement over the years. I know they are proud of me and, although I sometimes protest at their extolment, it is a wonderful thing to know that two people have such unwavering belief in me. Thanks also to Pippa for the fantastic job she did proofreading.

Most of all, I must thank my wife, Tanya. She was with me every step of the way, making sure I had time to work when I needed it, encouraging me when I was procrastinating and generally helping me to become the best version of myself that I could be, as she always does.

# Abbreviations

**AAFU** Age at first use

**AIC** Akaike information criteria

**AIDS** Acquired immunodeficiency syndrome

**BBV** Blood-borne virus

**CDF** Cumulative distribution function

**CI** Confidence interval

**CLL** Complementary log-log

**CRF** Cross-ratio function

**DAT** Drug action team

**ELISA** Enzyme-linked immunosorbent assay

**EM** Expectation-maximisation

**ESLD** End-stage liver disease

**FOI** Force of infection

**GLM** Generalised linear model

**GUM** Genitourinary medicine

**HBV** Hepatitis B virus

**HCC** Hepatocellular carcinoma

**HCV** Hepatitis C virus

**HES** Hospital Episode Statistics

**HIV** Human immunodeficiency virus

**HR** Hazard ratio

**LR** Likelihood ratio

**MLE** Maximum likelihood estimate

**MPES** Multi-parameter evidence synthesis

**MSM** Men who have sex with men

**MV** Multivariable

**NESI** Needle Exchange Surveillance Initiative

**NHANES** National Health and Nutrition Examination Survey

**NHS** National Health Service

**OR** Odds ratio

**PCR** Polymerase chain reaction

**PHE** Public Health England

**PVF** Power variance function

**PWID** People who inject drugs

**QQ** Quantile-quantile

**RFV** Relative frailty variance

**RIR** Relative inclusion rate

**RNA** Ribonucleic acid

**SE** Standard error

**TVC** Time-varying covariate

**TVF** Time-varying frailty

**UAM** Unlinked Anonymous Monitoring

**WHO** World Health Organization

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Blood-borne viruses and the burden of disease

The World Health Organization (WHO) estimates that in 2015 around 71 million people were living with chronic hepatitis C virus (HCV) infection worldwide, and 257 million people with chronic hepatitis B virus (HBV) infection (World Health Organization, 2017). Hepatitis infection is a major cause of chronic liver disease, with an estimated 720,000 deaths due to hepatitis-related cirrhosis and 470,000 deaths from liver cancer (hepatocellular carcinoma) in 2015 (World Health Organization, 2017). Human immunodeficiency virus (HIV) remains a persistent problem, with 36.7 million people worldwide estimated to be living with HIV in 2015 (UNAIDS, 2016). Nearly half (17 million) are on antiretroviral therapy, which reduces the risk of developing acquired immunodeficiency syndrome (AIDS) and infectiousness to others. Nevertheless, there were an estimated 2.1 million new infections and 1.1 million AIDS-related deaths globally in 2015 (UNAIDS, 2016).

Blood-borne viruses (BBV) are a major world challenge, and ambitious targets have been set by the WHO for reductions in incidence and prevalence of HCV, HBV and HIV and associated morbidity by 2030 (United Nations, 2015; World Health Organization, 2016). Much of the epidemic is focussed in Asian and African countries (The Polaris Observatory Collaborators, 2017, 2018), but remain persistent problems in Europe, the USA,

Australia and other higher-income countries, where they predominantly (but not exclusively) affect specific high-risk groups.

People who inject drugs (PWID) are a major risk group in higher-income countries. Injecting drug use is associated with various health and social problems such as crime (Stewart et al., 2000; Reuter and Stevens, 2008), drug-related overdoses and mortality (Bargagli et al., 2006; Hickman et al., 2009), bacterial infection and infection from blood-borne viruses, in particular HCV (The Health Protection Agency, 2012*b*).

In England, 203,000 individuals were estimated to have antibodies to HCV in 2005 (Harris et al., 2012b), with over 85% in those that currently or had previously injected drugs (44% in current and 43% in ex-injectors). Around 76% of those infected progress to chronic infection (Micallef et al., 2006), many of whom will go on to develop serious complications of the liver. A back-calculation approach by Sweeting et al. (2007) indicated that over 10,000 individuals would be living with cirrhosis by 2015, and the burden of HCV-related liver disease is likely to increase further unless substantial numbers can be successfully treated (Harris et al., 2014).

HBV prevalence in PWID is also moderately high in England, with around 16% having past or current infection (The Health Protection Agency, 2012*b*). National exercises to estimate HBV prevalence have not been carried out in England, but WHO-commissioned research estimated 441,000 individuals living with current HBV infection in 2015 (The Polaris Observatory Collaborators, 2018), indicating that injecting drug use is a smaller component than it is for HCV. A substantial portion of positive tests for acute HBV occur in Asian, black and other/mixed ethnic groups (compared to white/white British), and of those with a known risk factor, injecting drug use is 7.3% (although this may be under-reported) (Public Health England, 2017).

HIV prevalence in PWID is relatively low in England, at around 1.2% (The Health Protection Agency, 2012*b*,*a*). There were estimated to be 101,200 people living with HIV in the UK in 2015 (Public Health England, 2016). Nearly half (47,000) were estimated to be among gay, bisexual and other men who have sex with men, and only 2,500 PWID. However, HIV-infected PWID have higher mortality and lower proportions of timely linkage to care than other groups, and rapidly spreading outbreaks of HIV still occur

in this population (Public Health England, 2016). HIV is therefore not an insignificant concern in PWID.

In England, evidence suggests that the risk of HCV in PWID was reduced throughout the early 1990s by successful harm reduction programmes (Hope et al., 2001), but there is the suggestion that incidence began to increase again from 2000 onwards (Sweeting et al., 2009b), with a similar pattern for HIV (Hope et al., 2005, 2014). Many of those starting to inject drugs in recent years will be too young to recall the beginning of the epidemic of HIV and blood-borne viruses in the late 1980s and may therefore be less aware of, or concerned with, the consequences of infection. Such a view may be partly due to the dramatic improvement in HIV treatments since the advent of combined antiretroviral therapy (WHO, 2012*c*), and there is the worrying view that HCV infection may be seen as "inevitable" by many PWID (Rhodes et al., 2004).

New treatments offer the potential to markedly reduce HCV prevalence, both by reducing the number of people currently living with chronic infection and reducing transmission, known as *treatment as prevention*. The potential impact of treatment as prevention has been investigated in mathematical models and predicted to dramatically reduce prevalence over 10 to 15-year time-spans, even at modest treatment levels (Martin et al., 2013). In contrast to high endemic levels of HCV in PWID, HBV appears to be well-controlled through vaccination (Public Health England et al., 2014), and HIV transmission is hindered by effective treatment and at a low endemic level. Nevertheless, these infections persist in the PWID population, and show no signs of declining over time. Understanding the risks of blood-borne infection in PWID and the crucial role that new treatments will play in the eventual elimination of HCV will require a thorough understanding of available surveillance data.

## 1.2 Epidemiology and biology

### 1.2.1 Hepatitis C virus (HCV)

HCV is most commonly transmitted through exposure to infectious blood, which can occur through receipt of contaminated blood transfusions, blood

products and organ transplants; injections given with contaminated syringes and needle-stick injuries in health-care settings; injection drug use; and being born to a hepatitis C-infected mother (WHO, 2012*b*). It is thought that transmission via sexual contact is rare (Balogun et al., 2003), although there is increasing concern that certain high-risk sexual practices are leading to transmission in men who have sex with men, in particular those infected with HIV (van de Laar et al., 2007). In the UK, most HCV infection occurs via injecting drug use (De Angelis et al., 2009). Since the introduction of screening for blood donations in 1991, very few infections occur due to blood transfusions (McClelland, 2013) and a substantial proportion of infection in those that have never injected drugs is likely to have been imported from high-prevalence countries (Harris et al., 2012b).

Six known HCV genotypes exist (with various subtypes), although the majority of positive tests in England are genotype 1 (47%) or 3 (44%) (Public Health England, 2015). Outcomes are generally similar for the two types, although genotype 1 infections respond somewhat differently to treatment (European Association for the Study of the Liver, 2015), and for practical purposes the infected population is often dichotomised into genotype 1 and "non-1" types.

Following infection with HCV (and other blood-borne viruses), the virus will begin to reproduce and shortly afterwards the body will produce antibodies in response. The strength of the response is initially weak, but increases over time in most individuals and plateaus at around 6 months (see section 1.2.4). Around 76% of those infected will develop chronic infection within around 6 months (Micallef et al., 2006) and will remain permanently infected, and infectious, unless the virus is successfully cleared by treatment.

Chronic infection results in progressive damage to the liver, which in some individuals will lead to scarring of the liver (cirrhosis). Cirrhosis has two stages, defined as *compensated*, in which there is scarring but while maintaining function; and decompensated cirrhosis, also known as *end-stage liver disease* (ESLD), when liver function is compromised. Individuals with cirrhosis are also at risk of developing hepatocellular carcinoma (HCC) (Mauss et al., 2018). Upon developing either of these diseases, mortality is extremely high, with annual rates of up to 20% and 60% for ESLD and HCC respectively (Hutchinson et al., 2005). At this stage, the only option is liver

transplantation, which is expensive, restricted by a limited supply of donors, and not without risks (Thuluvath et al., 2007).

Interferon-based treatments became available around 20 years ago, but have had limited success: durations are of 24-48 weeks, individuals experience unpleasant side-effects and success rates are low (Thomson et al., 2008). In particular, as individuals age and develop progressive fibrosis of the liver (the formation of excess fibrous tissue that precedes cirrhosis and severe disease) the probability of successfully clearing the virus decreases to as low as 37% in those with genotype non-1 infection, and lower still in those with genotype 1 (Thomson et al., 2008). This means that previous treatment has had virtually no impact on HCV-related morbidity, with low uptake and little possibility of disease prevention in those that need it most urgently (Harris et al., 2014). New treatments became available from around 2014 with markedly improved efficacy and success rates in excess of 90% in all groups, including those with cirrhosis (Afdhal et al., 2014; National Institute for Health and Care Excellence, 2015). It remains to be seen whether clearance of the virus will result in long-term improvements in liver function and life expectancy once substantial damage has already occurred, although studies suggest that risk of further disease progression is greatly reduced (Singal et al., 2010).

### 1.2.2 Hepatitis B virus (HBV)

HBV infection can be acquired through similar routes as HCV, but can also be transmitted sexually and there are risks of vertical transmission from mother to child. HBV infection carries a lower risk of severe disease, as a high proportion will clear infection naturally (WHO, 2012*a*). Chronic infection causes progressive disease of the liver and cirrhosis in much the same way as HCV, eventually leading to ESLD and HCC in some individuals (El-Serag, 2012).

HBV has been a major problem in some parts of the world, with much of the disease burden in children, who have a higher risk of developing chronic infection and disease progression. Fortunately, substantial progress has been made in preventing HBV infection through vaccination programmes (Shepard et al., 2006). In the UK vaccination is only offered to specific risk groups,

which include PWID. A high proportion of PWID in the UK are now vaccinated against infection (Hope et al., 2007).

### 1.2.3 Human immunodeficiency virus (HIV)

The first reported cases of HIV occurred in the 1980s and quickly spread, rising to an estimated peak of 3.7 million new infections in 1997 (Fettig et al., 2016). HIV can be transmitted sexually and through blood-to-blood contact, although HIV is less infectious than HCV via blood (de Vos et al., 2012; Baggaley et al., 2006). Infection causes a fall in CD4 cell count and corresponding immunodeficiency, resulting in the potential to develop a number of conditions that would be rare in the uninfected population. Such conditions are classified as *acquired immunodeficiency syndrome* (AIDS), and are generally associated with high mortality rates (The Antiretroviral Therapy Cohort Collaboration, 2009).

In the UK, the HIV epidemic is focussed in particular risk groups, including men who have sex with men, PWID, and those born in sub-Saharan Africa (Public Health England, 2016). A large proportion of infected individuals are now on antiretroviral drugs, especially in industrialised countries (WHO, 2012*c*), which greatly reduces infectiousness and the risk of AIDS. Those on treatment are now estimated to have life expectancy comparable to the general population (The Antiretroviral Therapy Cohort Collaboration, 2017).

### 1.2.4 Serological testing

Sero-epidemiology refers to the practice of testing for antibodies to determine current or past infection with a virus. Such tests are not always perfect: the test may have imperfect *sensitivity*, whereby those with antibodies test negative. The opposite of this is imperfect *specificity*, where those without antibodies test positive (Altman and Bland, 1994). In general, test status is determined by judging whether some continuous quantity is above a certain threshold, which will be developed to balance sensitivity and specificity in some way; the value selected for the threshold will depend on the expected prevalence in the population of interest, and the potential consequences associated with the two types of error, called false positives and false negatives.

An alternative approach, rather than dichotomising infection status, is to analyse the antibody levels directly in order to answer epidemiological questions (Bollaerts et al., 2012).

In addition to tests for antibodies, active infection may also be determined by directly testing for the presence and quantity of the virus in the blood, usually via Polymerase Chain Reaction (PCR) testing for RNA, the genetic material of the virus. Both antibody and RNA tests are available for HCV, although the latter is only recently becoming more commonplace in surveillance (Public Health England, 2018*a*).

Testing for HBV is more complicated, with three types of test. Hepatitis B surface antigen tests are used to detect the presence of the virus in the blood; hepatitis B surface antibody tests indicate whether the person is protected against HBV, either through past infection or vaccination; and hepatitis B core antibody tests indicate current or past infection, but cannot determine whether this confers protection against HBV. HIV testing also consists of antibody and antigen tests, often using an enzyme-linked immunosorbent assay (ELISA) to detect both HIV antibodies and antigens in the blood. Viral load tests are an integral part of monitoring infected patients in clinical practice once infection is confirmed.

For some infections, it is possible to assess whether an infection occurred recently via the stage of the body's response to the virus. For instance, in the first few weeks, individuals may test positive for the presence of the virus (RNA) but still be antibody negative, as the body has not yet mounted a response. Further, the strength of antibody binding, known as *avidity* may be used as a marker: this is weak shortly after infection occurs and increases over time. However, the process is imperfect: for HCV there is considerable uncertainty regarding the length of the so-called "window" period between initial infection and the development of strongly binding antibodies (Klimashevskaya et al., 2007; Gaudy-Graffin et al., 2010). Sample sizes are also an issue, as very large numbers are required to observe a sufficient number of individuals with markers of recent infection. The period in which an individual is HCV RNA positive and antibody negative is thought to last for less than 2 months; even with a rate of infection as high as 12 per 100 person years this would imply only around 20 such individuals would be expected to be observed in a sample of 1000 antibody negative individuals, and the

confidence intervals for resulting calculations tend to be wide (Public Health England, 2018*a*).

A number of approaches are used to determine recent HIV infection in diagnosis data, including "de-tuned" tests that cannot detect antibodies in samples that have previously been confirmed as antibody positive, proportions of HIV-specific immunoglobulin directed against HIV antigens, and avidity assays (Murphy and Parry, 2008; Rosenberg et al., 2016). An algorithm to determine recent infection is routinely used in diagnostic testing by Public Health England (Public Health England, 2016); as test-seeking often follows an episode of exposure risk, this is more pertinent to testing behaviour than answering epidemiological questions about incidence of infection.

### 1.2.5 Opiate and injecting drug use

The use of opiates administered via injection into the blood, and in particular heroin, has been around since the beginning of the last century when it was first marketed for medicinal use. It was still relatively uncommon in the 1960s but increased markedly in the 1980s, when the epidemic is reported to have begun (De Angelis et al., 2004; BMA Board of Science, 2013, Chapter 5). The population is difficult to enumerate, but there is evidence that injecting drug use in England has fallen in the last decade to below 100,000 people injecting in the last year; however, a far larger number are previous injectors still at risk, or otherwise still recovering from opiate dependency (Hay et al., 2012).

The population considered here are people who currently inject drugs. This population is somewhat difficult to define as the level of usage can vary, with some people only ever occasionally injecting, and longer-term users cycling between multiple periods of abstinence and relapse (Xia et al., 2015). Further, although opiates are the most commonly injected drugs, the group of people who inject drugs may include those injecting crack cocaine, amphetamines and other drugs. These groups may have very different risk patterns in terms of injecting frequency and risk behaviour. Nevertheless, the bulk of injecting drug use is opiate-based, in particular heroin, the addictive properties of which mean that most injectors will be dependent and injecting

on a more or less regular basis. Typically, a "current" injector is taken to mean someone that has injected in the last year, with those that have ceased to inject for one year or more being classed as "ex-injectors" (De Angelis et al., 2009).

Those beginning to inject drugs are often, unsurprisingly, in difficult personal circumstances; challenges in monitoring and treating this population are often described as being due to "chaotic" lifestyles (Mravčík et al., 2013). Typically the age at first use is in the early twenties (Kimber et al., 2010), and the population is characterised by high rates of homelessness and imprisonment (Cullen et al., 2015). Nevertheless, detailed studies of PWID indicate that a large proportion of this population is in contact with some kind of treatment or harm reduction service (i.e., needle exchange) (Hickman et al., 2007), which is promising when considering whether survey data obtained via these services are representative.

Infection with blood-borne viruses will occur due to sharing of needles, syringes, or other injecting paraphernalia that has been contaminated with infected blood. It has been widely reported that risk of infection is far higher in those that have started injecting recently (Crofts et al., 1997; Sutton et al., 2006, 2008). One reason for this is that new initiates are likely to be assisted by someone more experienced and not have their own injecting equipment. Under these circumstances, a significant proportion of initiates may become infected very early on, potentially within a few weeks of initiation. Further, there may be a period of stabilisation before the initiate begins to use needle and syringe exchange services. Once injecting behaviour is established, PWID may actually be at relatively low risk of infection, provided their circumstances remain stable. However, intermittent episodes of upheaval are likely, due to interruptions or changes in drug supply, changing peer groups, periods of homelessness or imprisonment, or any other factor that might cause a change in risk behaviour. Although the timings are not clear, homelessness and imprisonment are strongly associated with the risk of blood-borne infection (Cullen et al., 2015). Patterns of injecting use and mixing may also have a seasonal component, although this would be difficult to assess in practice and no evidence on this is available.

Harm reduction strategies have been in place in England for over two decades (Hope et al., 2001), consisting of opiate substitution therapy and

needle and syringe programmes. The former reduces drug dependence and the frequency of injecting, and the latter aims to reduce unsafe injecting practices. These measures have been shown to be effective in reducing risks of HCV (Turner et al., 2011), although mathematical modelling suggests that other interventions are required to achieve substantial reductions in prevalence, such as HCV treatment (Vickerman et al., 2012).

## 1.3 Surveillance systems for monitoring blood-borne viruses

### 1.3.1 Study designs

Approaches for investigating the incidence or prevalence of infection (or any condition of interest) in a population fall under the three main observational study designs: cohort, case-control and cross-sectional studies. Case-control studies are useful for determining risk factors, but do not provide any information on absolute prevalence or incidence of infection, and are not considered subsequently.

Cohort studies recruit a group of people and follow them up over time, allowing direct observation of the rate at which individuals acquire infection or other event of interest. For "hard" outcomes such as mortality, or conditions that otherwise have a known onset date, this may be known precisely; but incident infection may only be known to have occurred within some time interval between repeated tests. Although this should be accounted for in the analysis of such data, it is not a major concern provided the frequency of testing relative to the typical exposure duration is sufficient. For example, many PWID inject for 10 years or more before permanent cessation (Sweeting et al., 2009a) and serological testing for blood-borne viruses every 6-12 months would provide sufficient information to understand broad patterns of incidence throughout injecting career.

Although cohort studies have the advantage of being able to directly estimate incidence, they are time-consuming and costly to conduct. Further, those agreeing to participate in such studies may not be typical of the population of interest. This may be particularly true of PWID, who are at risk

of homelessness, imprisonment and other instabilities, and are commonly described as being "hard to reach" by treatment or other services (Alliance - CAHR, 2013). PWID who choose to regularly participate in research studies for extended periods may therefore be somewhat different to the norm.

Cross-sectional studies sample from the population of interest and test infection status, providing a snapshot of prevalence at a particular time point. This type of study is often carried out as part of routine surveillance by public health bodies. In many cases the samples are anonymised, and may have little in the way of additional data; frequently, only the age of the individual is considered. Such data are known as age-specific *current status data* (Keiding, 1991). Alternatively, other information may be collected, either based on patient records, or via a questionnaire. Such data will often still be anonymous, which makes obtaining ethical approval for conducting research easier. For example, legislation allows the anonymous testing of residual sera without consent for specific public health purposes. As with cohort studies, there are a number of potential issues to consider when making inferences from such data, as they may be subject to various forms of selection bias, whereby the population sampled is systematically different to the target population.

### 1.3.2 Studies of people who inject drugs

A number of long-term cohort studies have been set up to study PWID. The Amsterdam Cohort Study began in 1985, with the aim of investigating risk factors for blood-borne and sexually transmitted infections, and effects of interventions. Participants complete a standardised questionnaire every 4-6 months when visiting the Health Service of Amsterdam (van den Berg et al., 2007). The Edinburgh Addiction Cohort studies patients sampled from a large primary care facility with a history of injecting drug use, and examines behaviour and outcomes through a combination of interview-based questionnaire data and primary care records (Kimber et al., 2010). Again, there is a strong interest in investigating risk patterns for infections, but also on behavioural aspects of injecting drug use.

Public Health England (and its predecessors) has conducted monitoring programmes in different risk groups under the umbrella of The Un-

linked Anonymous Monitoring programme (UAM) (Kessel and Watts, 2001). This includes monitoring leftover samples from those attending genitourinary medicine (GUM) clinics, pregnant women (via residue samples from newborn infants), and surveys of PWID attending drug treatment centres or needle exchange services. The latter is an annual cross-sectional seroprevalence survey of PWID in England, Wales and Northern Ireland that began in 1990, and is the principal data source used in this thesis (Public Health England, 2014). Briefly, PWID attending treatment and needle exchange surveys are invited to participate in the study. Those that consent complete a questionnaire on risk behaviour and demographics, and provide a serological sample that is tested for HCV, HBV and HIV (see section 1.2.4). Data from the study have been used to monitor prevalence of blood-borne viruses and trends in risk behaviour, and investigate the epidemiology of blood-borne viruses in PWID. The study provides information on prevalence according to the duration of injecting and the correlation between different infections (see section 1.4), which are key themes of this thesis. A more detailed overview of the study and previous research applications is provided in chapter 2.

The Needle Exchange Surveillance Initiative (NESI) in Scotland is another PWID survey programme (Health Protection Scotland, 2017). Similar to the UAM survey, the aim of NESI is to measure and monitor the prevalence of BBVs and injecting risk behaviours among PWID. The initiative was funded by the Scottish Government as part of the Hepatitis C Action Plan, which states that efforts to prevent HCV in Scotland must focus on preventing transmission of the virus among PWID (NHS Scotland, 2005). As with the UAM, the purpose of NESI is to provide information to evaluate and better target interventions aimed at reducing the spread of infection among this population group (Allen et al., 2010). At present, research using NESI data is particularly focussed on monitoring incidence and the potential impact of new HCV treatments on transmission (Palmateer et al., 2014).

Numerous analyses of cross-sectional data in PWID have been conducted. Prior to NESI, a cross-sectional study of PWID in Glasgow was conducted from 1990-1996, finding high HCV prevalence that increased with injecting duration. Age, age started injecting and calendar time were considered, although no force of infection analysis was conducted (Taylor et al., 2000).

Del Fava et al. (2011b) examined cross-sectional data on HCV and HIV infection in PWID in Italy and Spain, noting a strong correlation between the prevalence of infections (odds ratios of 2.56 and 2.42 for Italy and Spain respectively). They used various approaches to jointly model prevalence, including the bivariate Dale model, generalised linear mixed models and shared frailty models. Del Fava et al. (2011a) also examined aggregated cross-sectional data in 20 regions in Italy, finding an association between HCV and HIV prevalence and marked variation between regions. Over-dispersion and other issues are discussed in more detail in the PhD thesis of Del Fava (2012).

### 1.3.3 Other surveillance systems and studies of blood-borne virues

A number of studies have used opportunistic testing to estimate prevalence of HCV and other blood-borne viruses; for instance, national HCV prevalence in Australia was estimated using opportunistic testing of residual sera from pathology laboratories (Amin et al., 2004) and patterns of HCV prevalence in the UK assessed using antenatal screening data (Balogun et al., 2000). HCV prevalence has also been examined in blood donors (The Health Protection Agency, 2009), and those attending GUM clinics (Balogun et al., 2003) and emergency departments (Orkin et al., 2015).

Other studies have used specific sampling frames to assess HCV prevalence, such as The National Health and Nutrition Examination Survey (NHANES) in the USA (Denniston et al., 2014). The survey collects data on the health of the non-institutionalized population using a multistage probability sampling design from approximately 5000 persons annually. By definition, those that are homeless or incarcerated are excluded from the study. A cross-sectional study in France was used to estimate prevalence of HCV and HBV. Again, a complex, stratified, multistage sampling design was required, the sampling frame covered 80% of the population (beneficiaries of the national health insurance system) and response rates were low (11%) (Meffre et al., 2010).

Monitoring HCV in the general population is problematic, due to major difficulties in obtaining an unbiased sampling frame. In particular, PWID

(current or ex) will tend to be under- or over-represented in surveys: they may be less frequently observed in surveys of the general population due to an increased likelihood to be homeless, in temporary accommodation or incarcerated. Conversely they are more likely to use a variety of health services; for example, around 10% of those in the UAM survey of PWID report using a GUM clinic in the past year, and over one quarter using an accident and emergency department. This means that opportunistic testing will, in many cases, result in higher observed prevalence than that of the general population. In England, only around one in 50 adults have ever injected drugs, but the probability of HCV infection in those that have is over 200 times higher than those that have not (De Angelis et al., 2009). Even a small degree of under- or over-sampling of those that have ever injected drugs will therefore have a major impact on observed prevalence.

In England, the lack of a valid sampling frame has led to the development of a multi-parameter evidence synthesis (MPES) approach, which combines multiple sources of data on risk group sizes and risk group-specific prevalence studies (Sweeting et al., 2008; De Angelis et al., 2014). The appeal of this approach is that specific biases can be accounted for, and the consistency of evidence assessed. MPES estimates are the basis of official prevalence figures for HCV and HIV in England. In Scotland, which has better linkage between surveillance systems (for instance, between HCV diagnoses and other sources), a more detailed analysis of diagnosed and undiagnosed prevalence was undertaken in an MPES framework (Prevost et al., 2015).

Further understanding of the HCV epidemic may be inferred from information on the number of people developing severe HCV-related liver disease, namely HCC and ESLD, or HCV-related mortality. Such data are available from Hospital Episode Statistics (HES) data, and the ONS mortality data. Severe liver-related disease is almost certain to result in hospital attendance and is reliably recorded; however, the fact that it is HCV-related may not be recorded, and the potential for under-reporting must be borne in mind (Mann et al., 2009). Examining trends in *disease endpoints* provides information on those with earlier disease stages: compensated cirrhosis is the precursor to severe liver disease, and will roughly follow the same pattern, one step removed. In fact, if rates of disease progression are known (or can be assumed) then earlier disease stages, and the whole course of the historic epi-

demic, can be reconstructed via the back-calculation approach (Brookmeyer and Gail, 1988). This approach can be used to make short-term predictions, and has been much used for HCV to estimate future disease burden and potential impact of treatment (Sweeting et al., 2007; Deuffic-Burban et al., 2012; Harris et al., 2014); and for HIV (Sweeting et al., 2005; Birrell et al., 2013; Brizzi, 2018). However, the approach has limited scope for estimating incidence and overall prevalence, due to a lack of information on recent incidence of infection and the need for accurate population-level rates of disease progression, which are difficult to determine (Sweeting et al., 2006).

## 1.4 Analysis of age-specific prevalence data

This thesis is primarily concerned with the analysis of age-specific prevalence data obtained from cross-sectional seroprevalence surveys, also known as current status data. With "age" interpreted more generally as *time at risk*, such data provide information on the rate at which prevalence increases with time at risk and hence the force of infection, the rate at which uninfected individuals acquire infection. This process was first modelled by Muench (1934) and described as a "catalytic model". The basis of the model is given by

$$S(a) = \exp\left[-\int_0^a \lambda(u)\ du\right],$$

where $S(a)$ is the proportion of uninfected individuals and $\lambda(a)$ the force of infection at age $a$.

A variety of choices are available for defining the functional form of $\lambda(a)$: in its simplest case, this may be constant. Parametric models may be defined based on subject-specific knowledge, incorporating known epidemiological characteristics of the infection. For instance, the force of infection may decrease, increase, or accelerate with age, suggesting the use of Weibull or Gompertz distributions, which are widely used in epidemiology and human survival. Childhood diseases may rise to a peak and fall in adulthood: Farrington (1990) proposed the use of an "exponentially damped" function to capture such patterns.

Piecewise constant models specify constant hazards within a set of age intervals, which may be defined according to known epidemiological charac-

teristics such as infant, school age, adulthood, etc. (Farrington et al., 2001). Alternatively, a large number of small intervals (such as for each year of age) allows for arbitrary age patterns, although there may be insufficient data to obtain precise estimates of the force of infection within each interval.

This has led to the development of different approaches for providing flexible shapes with a comparatively small number of parameters, or providing some smooth function. Keiding (1991) discussed the use of kernel smoothing of the estimated incidence, and later the use of spline-based smoothing (Keiding et al., 1996). A similar, local polynomial approach was employed by Shkedy et al. (2013) to analyse datasets on rubella, mumps and hepatitis A. Marschner (1996) used piecewise constant incidence models along with a moving average as part of an expectation-maximisation (EM) algorithm. Parametric models using fractional polynomials have also been considered (Shkedy et al., 2006), as have penalised splines and generalised linear mixed models (Nagelkerke et al., 1999; Namata et al., 2007).

Flexible, parsimonious approaches are particularly useful when the age-specific pattern is of particular interest, may have an arbitrary shape, and estimation is restricted by comparatively sparse data. Frequently of interest in the study of infectious diseases are contact patterns, in terms of who infects who (Farrington et al., 2001). Detailed examination of age-specific patterns of infection may reveal features of interest: for instance, a secondary peak in infections in parent-age adults may suggest transmission from their school-age children. Such an interpretation was made by Shkedy et al. (2013) with regard to data on mumps, although resting on assumptions about mixing patterns.

### 1.4.1 Models for age and calendar time

A key difficulty with the analysis of cross-sectional data from a single time point is that the effects of calendar time are completely confounded with age, and time homogeneity must be assumed. For a great many infections this is implausible due to improvements in hygiene and awareness, a fact pointed out by a number of authors (Keiding, 1991; Nagelkerke et al., 1999), but often not accounted for in practice. An alternative is to estimate the time-specific force of infection under an assumption of age-homogeneity, as

argued for by Schenzle et al. (1979) in their analysis of hepatitis A in different European countries.

It is preferable to consider that the force of infection may vary by both age and time. Ades and Nokes (1993) proposed an extension to the standard model that may be used where data from multiple serological surveys at different points in time are available, which is the case with the dataset used in this thesis.

Ades and Nokes (1993) treated the effects of age and time as independent (and combining additively or multiplicatively) and only considered more complex functions in terms of testing for the presence of potential age-time interactions in piecewise constant models to verify the adequacy of this assumption. An alternative to the age-time formulation is to specify the model in terms of cohorts (birth years, or year first injected for PWID) plus an age or time dimension; a third effect would be confounded with the other two and not identifiable. Nagelkerke et al. (1999) considered age and cohort effects, which were introduced as covariates, rather than age and calendar time. In any case, interactions between age and time (or cohort effect) can only be estimated within the time frame of serial data collection; prior to this age and time effects are confounded.

Age-time models bring (literally) another dimension to the problem of estimating an underlying process for the force of infection from noisy data with inherently low informational content. Unless age and time effects combine multiplicatively (or additively), specification of two-dimensional parametric forms may be problematic. Smoothing in two dimensions is also not straightforward. One approach of potential use is the highly flexible class of generalised additive models and "thin plate" splines (Hastie and Tibshirani, 1986; Wood, 2003). Such models were used to estimate the underlying incidence rate of HIV in UK men who have sex with men (Brizzi, 2018) and could also be applied to current status data such as those used in this thesis.

### 1.4.2 Mortality and relative inclusion in sample

The effect of mortality in the analysis of current status data can largely be ignored if the focus is on the average age-specific force of infection, or investigation is of an infectious agent that has no, or little, effect on mortality

and mortality is low until advanced age (Farrington et al., 2001). However, difficulties can arise if those infected are subject to a higher mortality rate, and therefore less likely to be observed in the sample, for instance, in the pre-treatment era of HIV. Differential mortality can be considered more generally in terms of the *relative inclusion rate* (RIR), which was considered in the context of HIV-infected women in ante/neonatal surveys by Ades and Medley (1994). Due to the sampling frame, the RIR is relevant here in terms of the relative fertility of those with and without HIV infection. Not all parameters are fully identifiable from the data; age-specific patterns were identified (via serial cross-sectional surveys, as in Ades and Nokes (1993)) but overall incidence could only be estimated to within a constant of proportionality. Alternatively, information on infection-specific mortality can be obtained from external sources and used within the analysis, as in the study by Keiding et al. (1989) on incidence of diabetes. Marschner (1997) discussed the potential for identifying patterns of age- and time-specific force of infection and RIR from cross-sectional data, concluding that an external data source would be preferable, or exploration of the impact of different values for the RIR in sensitivity analyses.

It is not clear whether PWID infected with blood-borne viruses would be more or less likely to take part in surveys, and if so, whether there might be relative differences in the RIR according to injecting duration. HCV, HBV and HIV can all lead to early mortality, although generally at low levels for HBV and HIV in the combined antiretroviral treatment era. Mortality due to HCV is a potential issue, although disease progression at younger ages is generally slow (Sweeting et al., 2006) and unlikely to have a substantial impact on those injecting for less than 20 years, which form the majority of the UAM sample. Nevertheless, this possibility must be borne in mind when considering estimates of the force of infection in those injecting for longer periods.

### 1.4.3 Contact matrices: who infects whom

The force of infection as described so far represents an average age-specific rate. A key parameter of interest for planning public health interventions such as vaccination is $R_0$, the expected number of infections produced by

the introduction of an "average" infected individual in a population of susceptible individuals (Farrington et al., 2001). If $R_0$ is less than 1 (or can be reduced to below 1 via vaccination), the infection will naturally die out within the population, otherwise the infection will remain endemic.

Central to this question are patterns of infectious contact between individuals of different ages (or more generally, between groups). This is commonly defined in terms of a *contact matrix* specifying the rate of contact between individuals of different age classes; again, these may reflect societal structures of pre-school, school age, adult and so on. Unconstrained, the contact matrix will generally have more parameters than the available degrees of freedom and require reductions in the number of parameters: for instance, different within-group rates but the same rate of contact between all groups of different ages. Different assumptions about the structure of the contact matrix may lead to markedly different estimates of $R_0$, with no difference in model fit (Farrington et al., 2001).

An alternative to estimating rates of contact from the data under some assumed structure is to use an external source of information on mixing patterns and relative intensities. The POLYMOD project, for instance, was a large-scale study on the frequency of day-to-day contacts between different ages across several European countries, with the explicit aim of informing mathematical modelling (Mossong et al., 2008).

In terms of PWID and the sharing of injecting equipment that results in effective contact, mathematical transmission models have generally been developed on the basis of homogeneous mixing or an assumed mixing pattern (see, for instance, de Vos et al. (2012) and Fu et al. (2016)). Rolls et al. (2012) based mixing patterns on empirical evidence from a social study of PWID, although, as ever, this population is difficult to study and recall of injecting partners is generally poor (Brewer and Garrett, 2001). These considerations are therefore set aside in this thesis, and the analysis restricted to estimation of the average force of infection, rather than the process of who infects whom. Of note is that in addition to a better understanding of contact patterns in PWID, further development of existing methods would likely be required, which are generally specified for infections with a short infectious period (Farrington et al., 2001), unlike the blood-borne viruses considered here.

### 1.4.4 Individual heterogeneity

Individual heterogeneity can have an important effect when estimating the force of infection. In particular, if some individuals are more likely to be infected than others, then there will be an apparently higher force of infection early on in the period of time at risk than later. This is because those that are at increased risk of infection will, on average, experience the event sooner, whereas the rate of infections will be lower at a later time where those remaining tend to be at lower risk. In this way, heterogeneity has the effect of decreasing the population hazard relative to the individual hazard over time, i.e., a *time gradient* is introduced Aalen et al. (2008, p. 235). Gamma distributions for the frailty are commonly used, although many other functions are possible, notably those from the so-called *power variance* family (Hougaard, 2000).

Heterogeneity in seroprevalence studies is often considered using multivariate data, in which information on frailty comes from the correlation that occurs between infections that share the same transmission route (Farrington et al., 2001, 2013). Various extensions to the shared frailty model have been considered. Hens et al. (2009) describe the use of a *correlated* frailty model for multivariate data, which allows for separate, but correlated frailties for each infection. These models are somewhat limited in that such effects may only be identified if a specific functional form is assigned to the force of infection, in much the same way as shared frailty models in the univariate case.

Frailty may also be allowed to vary by age (or more generally, exposure time), as described in Farrington et al. (2013). In their example, they found greater heterogeneity at younger ages, which diminished in adulthood. In the context of injecting drug use, changes in heterogeneity may be conceivable over the course of injecting career, with a high degree of heterogeneity in risk of blood-borne infection early on, possibly dependent on the context in which initiation occurred, but longer-term users having more homogeneous risk levels.

The selection effect induced by individual heterogeneity means that the hazard in survivors decreases over time, but another question is how heterogeneous these survivors are. This depends on the choice of frailty distribu-

tion, which may result in an increasing or decreasing coefficient of variation over time; in fact, the gamma frailty is the only distribution for which the coefficient of variation is constant (Aalen et al., 2008, p. 234; Farrington et al., 2012).

Note that it is not possible to distinguish between temporal and selection effects (Farrington et al., 2012), i.e., whether increasing or decreasing heterogeneity with age is due to a particular underlying frailty distribution that causes the change via selection, or genuine changes in the frailty distribution over time. Farrington et al. (2012) recommend that subject-specific knowledge or external data are used to guide the choice of frailty distribution, then examine potential time variation given the chosen functional form.

## 1.5 Aims of the thesis

The preceding sections provide some background on blood-borne viruses, epidemiology and sero-surveillance, and the population at risk. The aim of this thesis is to apply methods for estimating incidence and prevalence from current status data to surveillance data on PWID. These data present an opportunity to develop methods for the analysis of cross-sectional data, which in turn may reveal new insights on the epidemiology of blood-borne viruses in PWID. The focus of analyses is on HCV, which of the three infections considered is the greatest public health concern in PWID. However, HBV and HIV are also of interest, and form an integral part of investigation of variability in risk.

The UAM study is a unique data source, with a long series of sequential surveys and a wealth of covariate data not typically found in current status data. This gives the potential to investigate models including risk factors and their effect on the force of infection. The effects of gender, age (in addition to injecting duration), sharing injecting equipment, frequency of injecting, needle exchange use, imprisonment and sexual behaviour are considered. Models are developed that estimate the overall effect of risk factors on the force of infection, and also whether the effect of a covariate is modified according to injecting duration or calendar time (an interaction effect). The data are sampled at a number of different sites, so geographic differences can be examined and in particular regional trends in prevalence and the

underlying force of infection.

Piecewise constant models are employed in order to maintain flexibility in the baseline force of infection while estimating covariate effects and individual variability, which are key points of interest. Parametric models and smoothing approaches, which might more efficiently model the baseline force of infection (especially when considering both age and time effects, such as in Brizzi (2018)) are borne in mind but not used. It is shown later that patterns of risk according to injecting duration ("age" in this context) are relatively simple.

A key aim of this thesis is to understand individual variability in risk. Shared frailty models that make use of bivariate infection data have typically been used to estimate individual variability, most commonly using a gamma frailty distribution. This thesis also examines the inverse Gaussian distribution, which results in a different selection effect, and time-varying frailty (Farrington et al., 2013), a relatively new development that has not been applied to data on PWID before. Such approaches can help to understand how variability in risk evolves throughout injecting career.

Bivariate frailty models are extended to include covariate information. This is seldom considered in practical examples, as frailty models typically use data aggregated by age and all individual variability is assumed to be unmeasured. By including covariates, the frailty component is interpreted as residual variability, which in theory should decrease (compared to a model without covariates) as more information on risk is introduced. This thesis aims to examine this phenomenon and, more practically, whether the marked variability in risk indicated by previous studies can be reduced using risk factor information. This would help to better target harm reduction interventions to those at greatest risk.

The UAM data include infection status for HCV, HBV and HIV, and bivariate models are considered for each infection pair. In this thesis trivariate models are also investigated. Trivariate models extend the shared frailty model to include different frailty components, which can be estimated from the richer correlation structure of the three infections, although there are restrictions on what can be identified. The information available to estimate different aspects of frailty is considered, and different formulations proposed, one of which has a similar form to the correlated frailty model.

This thesis is organised as follows. In the second chapter the UAM study is introduced, giving an overview of its history and purpose, available data, limitations and previous analyses. The third chapter provides an overview of statistical methods for analysing univariate cross-sectional data, covering the basics of generalised linear models for binomial data, the force of infection model, and including covariates. The fourth chapter presents results of fitting the univariate models described previously to the UAM data for HCV, HBV and HIV, with a focus on HCV. The fifth chapter provides an overview of multivariate models, in particular the bivariate shared frailty model, and different frailty distributions are discussed, including measures of association and time-varying frailties. Chapter six then presents results on measures of association in the UAM data and fitting bivariate frailty models, including models with covariates. In the seventh chapter component frailty models are introduced, and different model forms are discussed. Trivariate models are then fitted to the UAM data. The eighth chapter provides overall conclusions on the work conducted in this thesis. R code and other background information is included in the appendices.

# Chapter 2

# The Unlinked Anonymous Prevalence Monitoring Programme

Public Health England (PHE, formerly the Health Protection Agency) has conducted monitoring programmes in various risk groups for the last 20 years or more under the umbrella of Unlinked Anonymous Monitoring (UAM) surveys. This includes monitoring leftover samples from those attending genitourinary medicine (GUM) clinics, pregnant women (via residue samples from newborn infants), and surveys of people who inject drugs (PWID) attending drug treatment centres or needle exchange services. The latter survey on PWID is of principal interest in the following.

The aim of the UAM survey of PWID is to measure the changing prevalence of HIV, HBV and HCV, and monitor levels of risk and protective behaviours among PWID. The data are used to assess and develop appropriate preventative and health education campaigns, evaluate the impact of such interventions, and to assist in the provision of services for PWID in the UK. Survey data have been collected annually since the programme was established in 1990. PHE works in partnership with around 50 of 149 specialist drug agencies in England (data for Wales and Northern Ireland are also collected, but not considered here). The drug action teams (DATs) that are sampled from change from year to year: around 80 DATs have been sampled at some point since the survey was established, and many have samples

for every survey year.

Each year, the agencies are encouraged to ask all eligible clients to participate in the survey, an eligible client being a current or former injecting drug user who has not already participated in the survey in the current calendar year. Each eligible client is asked to complete a short questionnaire, which includes questions on patterns of drug use, including injecting duration, frequency in the last month and sharing of drug-taking paraphernalia, treatment for drug addiction and participation in needle exchange services, and their sexual behaviour. This information is used to assess the association between risky activities (such as needle sharing) and the prevalence of BBVs among PWID. Descriptions of the most relevant variables are given subsequently. A copy of the 2015 questionnaire is shown in the appendices (section 9.1).

Participants also provide a sample for serological testing: in the past, oral fluid samples were used, but the survey moved over to dried blood spot samples during 2009-2010. Identifying information is irreversibly unlinked from all samples before testing, ensuring that both the sample and the questionnaire are completely anonymous. Samples are tested for the presence of antibodies to HIV (signalling current infection), and antibodies to hepatitis C and to the hepatitis B core antigen (which can indicate current or previous infection).

All testing is conducted by the Virus Reference Department at Public Health England Colindale, which has strict policies for quality assurance and maintains all relevant accreditations (Public Health England, 2018*c*). Dried blood spots are assigned a unique identifier and labelled upon receipt and kept at 4°C for short-term storage and prior to testing and −20°C for long-term storage. Samples must be of a sufficient specified size for testing and are prepared according to the manufacturer specifications of the testing equipment and Public Health England's detailed protocols.

## 2.1 Available data

### 2.1.1 Hepatitis C

Hepatitis C testing has been performed regularly on samples since 1998 using the OraSure device, which has an estimated sensitivity of 91.7% (95% confidence interval (CI): 87.5, 94.8) and specificity of 99.2% (95% CI: 97.8, 99.8) (Judd et al., 2003). From 2009 to 2010, there was a gradual switch-over to dried blood spots, which have near-perfect sensitivity and specificity. Both tests were evaluated by the Public Health Laboratory Service (now Public Health England) in those with known infection status, to determine optimal thresholds for maximum sensitivity and specificity (Judd et al., 2003). A smaller number of samples are available for the years 1992, 1994 and 1996. These samples were taken using the Salivette device, which has sensitivity of only 74.1% (95% CI: 68.2, 79.4) and 99.0% (95% CI: 97.4, 99.7) specificity. It is necessary to account for varying sensitivity over time if HCV prevalence prior to 2011 is to be estimated, as in Sweeting et al. (2009b). It is also possible to adjust for the proportion of those infected that naturally clear infection to obtain estimates of the prevalence of chronic HCV infection, for example, as in De Angelis et al. (2009).

Since 2011, HCV positive samples have also been tested for avidity, a measure of how strongly antibodies bind, and a potential marker for recent infection (Klimashevskaya et al., 2007; Coppola et al., 2009; Gaudy-Graffin et al., 2010). Samples with weak avidity (and subject to an RNA positive test) are more likely to be recent infections; although the best value to use as a cut-off, and the likely "window" period between seroconversion and development of strong avidity is not well established.

### 2.1.2 Hepatitis B

Testing for hepatitis B via hepatitis B core antibody (anti-HBc, indicating past or current infection) has been performed in all survey years. The tests used had poor sensitivity prior to the introduction of dried blood spot testing (which has near-100% sensitivity) with 76% sensitivity prior to 1998 and 75% up to 2009-2010; although specificity has always been above 99% (Judd et al., 2007). Hepatitis B surface antigen tests are also conducted, although

numbers are low. The UAM study is currently in the process of carrying out hepatitis B surface antibody tests (see section 1.2.4) on a subset of data, which could provide more insights into patterns of HBV infection and validation of vaccination status (which is otherwise self-reported).

### 2.1.3 HIV

The UAM survey for PWID was originally conceived to monitor HIV and therefore samples are available for all years (Noone et al., 1993; Hope et al., 2014). The Salivette device was used until 1998, followed by the OraSure device before dried blood spots were introduced around 2009-2010; for all years the test has near-perfect sensitivity and specificity. HIV has been studied extensively in terms of CD4 cell counts, RNA levels and treatment with antiretroviral therapy in cohort studies, but only infection status is available in the UAM data. Prevalence of HIV infection in PWID is low compared to HCV and HBV, and follows somewhat different patterns with markedly higher prevalence in London compared to all other regions (The Health Protection Agency, 2012*b*).

### 2.1.4 Injecting behaviour

Participants are asked whether they injected in the last year and in the last month, as well as frequency of recent behaviour. Many studies that have used the UAM data have focussed on those that have injected within the last month, this being seen as the best definition of "current" injectors, who are a group of principal interest. Some participants have not injected for over a year, who might viewed as a potentially unusual group of ex-injectors, as those that have ceased to inject but are still in long-term treatment are likely to have different characteristics to the general population of ex-injectors. Despite this, the UAM data have been used to estimate HCV prevalence in ex-injectors (Sweeting et al., 2008), albeit after careful adjustment for biases that are likely to be present in so-called *snapshot samples* (Kaplan, 1997).

Participants are also asked about sharing needles and other paraphernalia, including whether they have ever received needles or syringes from anyone, and the number of people they have passed on to, and received needles from, in the last 28 days.

Finally, participants are asked for the age at which they first started injecting. Combined with their current age, this provides a key piece of information on understanding the risk of infection: injecting duration. For infection via injecting drug use this corresponds to time at risk. The typical "career" of PWID will often involve multiple periods of stopping and starting, rather than a single, uninterrupted period of injecting prior to cessation, although the simplifying assumption of a continuous period of risk is usually made.

### 2.1.5 Sexual behaviour and health

Participants are asked whether they have had sex in the last 12 months, with how many male and female partners, and whether they have exchanged goods or money for sex in the last year, or ever. They are also asked whether they have attended various types of health services in the last year, in particular GUM or STI clinics. The most recent questionnaire also includes questions on accident and emergency attendance due to overdose.

Participants are also asked about previous HIV and HCV testing and treatment for HCV, whether they have ever been in prison or a young offenders institution (and whether injected drugs while in prison), and whether they have ever been homeless. There is little information on timing and duration of periods of homelessness or imprisonment, although there is a question on when last went to prison.

## 2.2 Previous UAM studies

Data collected from the UAM survey of PWID have been used in various ways to estimate trends in the prevalence of blood-borne infections, behavioural characteristics and demographics of PWID in England and Wales. One of the earliest studies examined prevalence of HIV and HBV in PWID in 1990 and 1991, finding a prevalence of 1.2% and 1.8% respectively for HIV and 33% and 31% for HBV, with high levels of reported sharing of equipment and risky sexual behaviour (Noone et al., 1993).

A later analysis examined HCV prevalence in 1997 and 1998, a time when the effectiveness of harm reduction measures for PWID was being assessed,

and prevalence of BBVs was found to be lower than previously thought (30% for HCV, 21% for HBV, and 0.9% for HIV) (Hope et al., 2001). The association between HCV infection and HBV/HIV status was examined, and found to have odds ratios of 2.3 (95% CI: 1.8, 2.8) and 1.5 (95% CI: 0.6, 3.5) respectively after adjusting for injecting duration, age, sex, area and previous testing for HIV.

Between 1991-2000 an increase in needle sharing was reported, which was thought to be associated with a rise in HBV prevalence (HIV levels were too low for comparison, and HCV testing only recently introduced (Hope et al., 2002)). They speculated that needle sharing may have increased due to changed perceptions of the severity of HIV infection following advances in treatment, or the perception of no longer being at high risk of infection.

HIV trends between 1990-2003 were examined via the UAM and community surveys (Hope et al., 2005). Prevalence was found to have decreased throughout the 1990s, then increased from 2000 onwards, with prevalence far higher in London compared to other areas. They also used a force of infection model with piecewise constant time effects, and found higher incidence between 1998-2002 compared to 1992-1997 in London, but not other areas, and a far higher force of infection in the first year of injecting compared to other years.

Age of starting injecting and subsequent cessation were investigated by Sutton et al. (2005). A gamma distribution for age starting injecting was used, and a piecewise linear function for rates of removal by age chosen after a model selection exercise. They estimated that 50% of injectors start between the ages of 18 to 25, with 15% starting after the age of 30. Annual removal probabilities rose linearly up to a maximum of around 30% at age 30-35. The authors acknowledge potential under-representation of different injecting career lengths and attempted to incorporate this in their model.

Sutton et al. (2006) jointly modelled the force of infection of HCV and HBV between 1998-2003, investigating individual heterogeneity via a shared frailty model with a gamma distribution. They considered trends over calendar time, but eventually assumed a constant force of infection for each virus, which was found to fit the data equally well. For injecting duration, they selected a simple dichotomy of $< 1$ vs. $\geq 1$ years, finding a three-fold reduction in force of infection after the first year. This was assumed the

same for both viruses, again based on goodness of fit of the model. They found substantial individual heterogeneity and concluded that some PWID are at significantly higher risk of blood-borne infections.

The impact of HBV vaccination has been considered, with levels of uptake rising from 27% to 59% from 1998 to 2004 (Hope et al., 2007). However, there was no corresponding decline in prevalence, and the force of infection was estimated to have increased in 1999-2004 vs. 1993-1998 (Judd et al., 2007).

Sweeting et al. (2009b) examined HCV prevalence over time while adjusting for injecting duration and other covariates. Polynomial functions were fitted to the logit prevalence, including non-linear trends for time and injecting duration of up to degree 4, as well as age and region. By doing so, the underlying temporal trend was estimated while controlling for changing demographics in the PWID population. Imperfect sensitivity and specificity, including uncertainty in the estimates used (Judd et al., 2003), was incorporated in a Bayesian framework. Previous studies had acknowledged the limitations of the assays, but the work by Sweeting et al. (2009b) was the first to explicitly include this as part of the model. They found that after accounting for imperfect testing, prevalence had decreased in the early 1990s, but increased from the late 1990s until the mid 2000s.

More recent work using these data estimated HCV prevalence in recent initiates (self-reported injecting duration of three years or less), with this interpreted as a measure of incidence, which was found to be largely stable between 2000-2008 (Hope et al., 2012).

Local level differences in HCV prevalence have also been examined (Harris et al., 2012a). Across 152 Drug Action Team (DAT) areas, prevalence was found to vary substantially, with estimates ranging from 14% to 82%. Spatially correlated random effects models were employed (Besag et al., 1991) in order to identify spatial patterns, and crucially, to derive prevalence estimates for non-sampled areas. The inclusion of area-level covariates, or auxiliary variables in the spatial mapping terminology, further improved estimates in terms of out-of-sample prediction and reducing standard errors of small samples.

Hope et al. (2014) modelled prevalence of HIV between 1992-2012 using polynomial functions, similar to the analysis of Sweeting et al. (2009b). Inci-

dence was also estimated via a joint model for HIV and HCV, using injecting duration- and time-specific contributions to the force of infection. Both time and injecting duration components were estimated via a random walk function in a Bayesian framework, and the effect of injecting duration was shared by the two infections, based on the assumption that risk of either HIV or HCV throughout injecting career is proportional. This approach increased the power to estimate the individual effects, which are difficult to estimate as rates of HIV are relatively low, but does not account for the possibility of sexual transmission of HIV or other differing risk patterns between the two infections. Hope et al. (2014) also constructed a timeline of government policy changes, broadly ascribing periods of increased risk to policies focussing on drug-related crime, rather than treatment and harm reduction.

Markers of recent infection for HCV are currently being investigated, with avidity-based measures and proportions of RNA-positive, antibody-negative individuals being explored to determine patterns of incidence (Cullen et al., 2015). With the advent of new HCV treatments and the potential for reducing incidence via treatment as prevention (Martin et al., 2015), there is great interest in the potential for monitoring progress via such measures. However, estimates suffer from substantial statistical uncertainty and the power to detect reductions over time is low (Public Health England, 2018*a*). Therefore force of infection models still have a crucial role to play in understanding changes in the risk of infection over time.

# Chapter 3

# Methods for the analysis of cross-sectional serological surveillance data

Cross-sectional data on infectious diseases typically consist of binary (0/1) infection status and a set of covariates, the most important generally being age, but potentially including other demographic or risk factor information. The resulting data may be modelled via a generalised linear model (GLM) to determine the relationship between the covariates and infection status. Data consisting of age-specific infection status, or *current status data*, are usually modelled in a particular way: by relating infection status to time at risk, estimates of the age-specific rate of infection may be obtained in a framework that is conceptually similar to survival analysis. In fact, such data are called *interval censored type 1* data in the survival analysis literature.

This chapter is organised as follows: the GLM specification for binary responses is reviewed, and how such models are fitted to observed data by maximising the log-likelihood. Assessment of model fit is then discussed, along with differences between binary data and the aggregated binomial form. The basics of survival analysis are then reviewed, and how this theory is related to the analysis of current status data from cross-sectional surveillance data. The basic model for age-specific data is then extended to allow rates of infection to vary by both age and time, and the inclusion of other covariates.

## 3.1 Models for binary responses

Binary data consist of observations that can take only two possible states, assigning "failure" as zero and "success" as 1 (generally the diseased state or a positive test for infection). The interest lies in estimating the mean probability of success, $p$, and how this varies according to different covariate levels. This is achieved via a generalised linear model (GLM), which is specified by a random component, a systematic component, and a link function (Nelder and Wedderburn, 1972). For binary (or binomial) data, the random component is a set of independent observations with a Bernoulli (or binomial) distribution and mean probability of success $p_i$ for observation $i$. The systematic component relates a vector of parameters to the covariates,

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki},$$

where $\eta_i$ is known as the linear predictor. Finally, the link function relates the linear predictor to the mean of the random component, which in most cases will constrain the values that the transformed linear predictor can take to between zero and 1. The most widely used link function for binary data is the logit function, which gives the logistic regression model

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \tag{3.1}$$

The logit is a symmetric, S-shaped function that is bounded by zero and 1. An alternative to this is the probit model, which uses the cumulative distribution function (CDF) of the standard normal distribution, denoted $\Phi$:

$$p_i = \Phi(\eta_i). \tag{3.2}$$

Finally, the complementary log-log (CLL) link gives the model

$$\log(-\log(1 - p_i)) = \eta_i. \tag{3.3}$$

These links produce different relationships between the linear predictor and the probability of success, which are displayed in Figure 3.1. If the link functions are considered to be cumulative distribution functions, the probability

density function of the logit link has heavier tails than the probit link and the CLL link is negatively skewed.

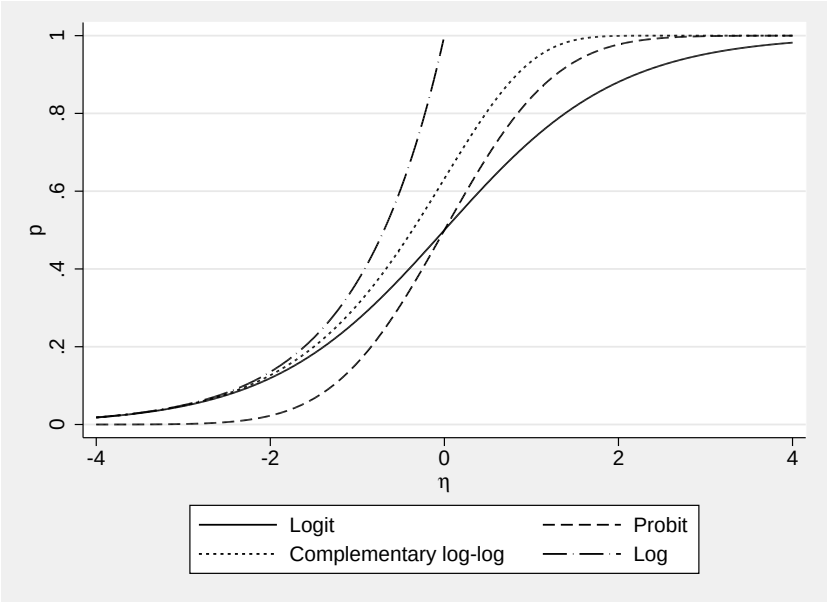


Figure 3.1: Link functions for binary data: linear predictor $\eta$ and resulting probability $p$.

Another possibility is the log link, which is not generally recommended for binary data as a positive value in the linear predictor will result in $p > 1$ (Figure 3.1). Nevertheless, a log link might be used to model binary data when the probabilities of success are low. More pertinently, a GLM with a log link may be used for the analysis of age-specific current status data. This is described in section 3.2.2.

Alternative link functions are of course possible. For instance, Aranda-Ordaz (1981) describes a symmetric family of functions that include the identity (linear) link and logistic as special cases; and an asymmetric family that includes the logistic and CLL as special cases. The univariate prevalence models considered in this thesis use only categorical variables, for which the specification of the relationship between the linear predictor and the mean probability may be of lesser importance: in univariable models fit is identical, although covariate effects may combine in different ways in multivariable models under different link functions. For multivariable models, some care is required to distinguish between the need for an alternative link function and the addition of interaction terms between covariates (Collett, 2002, p. 149).

### 3.1.1 Model fitting

In the following it is assumed that a logistic regression model is to be fitted to binomial data, although the concepts will still hold for binary data and other link functions. The aim is to obtain the maximum likelihood estimates (MLE) of the parameters, $\beta_j$, in the model. Given data where $y_i$ is the number of successes from $n_i$ observations in covariate group $i$, we maximise the binomial log-likelihood:

$$\sum_{i=1}^{n} \left[ \log \binom{n_i}{y_i} + y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) \right]. \tag{3.4}$$

As the term $\log \binom{n_i}{y_i}$ is constant, this does not need to be calculated in order to maximise the log-likelihood, and the function can be simplified to the *kernel* log-likelihood. However, maximising the kernel log-likelihood still requires solving a system of nonlinear equations, so numerical optimisation methods are required to find the set of parameter values that are maximum likelihood estimates. The Newton-Raphson method is typically used for logistic regression, which uses first and second derivatives to obtain local approximations of the function, and converges to the maximum over successive iterations.

### 3.1.2 Accounting for imperfect sensitivity and specificity

When modelling the infection status, or any other binary response that may be imperfectly measured, there is the potential to under- or overestimate the probability of infection $p$ if the test used to determine infection status does not always provide a correct classification. The imperfection of a test, assessed by comparing results against a "gold standard", is generally defined in terms of the proportion of true positives with a positive test status, called *sensitivity* and the proportion of true negatives with a negative test status, called *specificity* (Altman and Bland, 1994) (see section 1.2.4). Imperfect sensitivity (less than 100%) will therefore lead to underestimates of $p$, while imperfect specificity will result in overestimates. Furthermore, if different tests with varying sensitivity and specificity have been used in the sample, and the usage of some tests is associated with a covariate of interest, esti-

mation of the effect of this covariate can be biased due to the systematic under- or overestimation of $p$ at different covariate levels. For instance, if test sensitivity improves over the survey period, $p$ would be underestimated in earlier survey years.

If the sensitivity and specificity of the test are known then the probability of a positive test can be related to the true probability of infection. The table below shows the relationship between infection status and test results given imperfect sensitivity and specificity, where $\pi$ is the true probability of infection and *Sens* and *Spec* are the sensitivity and specificity:

| True disease status | Test result | |
|---|---|---|
| | Negative | Positive |
| Negative | $(1-\pi)Spec$ | $(1-\pi)(1-Spec)$ |
| Positive | $\pi(1-Sens)$ | $\pi Sens$ |

Therefore the observed proportions of a positive test $p$ and a negative test $1-p$ are given by

$$\begin{aligned} p &= \pi Sens + (1-\pi)(1-Spec) \\ 1-p &= (1-\pi)Spec + \pi(1-Sens). \end{aligned} \tag{3.5}$$

This relationship can then be incorporated into model fitting by substituting $p$ with $\pi$ in the link function. The covariates are then estimated according to the true probability of infection $\pi$. Provided that *Sens* and *Spec* are known, the kernel likelihood derived from equation 3.4 can therefore be used to estimate the parameters. Estimates from such a model will then estimate the true probabilities of infection and give unbiased estimates of covariate effects. The disadvantage is that the model can no longer be fitted using standard GLM procedures in most packages and must be implemented using bespoke code.

### 3.1.3 Model assessment

An adequately fitted model should predict the probability of success, or the number of successes for each covariate combination, with a suitable level of

accuracy. Model fit (or lack of) is easiest to interpret for binomial data, that is aggregated successes and failures for each covariate combination. "Covariate combination" here is taken to mean whatever grouping is used for the aggregated data. This might not correspond to the covariates that are used in the model, and instead might be considered as *unique data combinations*. For instance, data might be aggregated by age in years, but analysed in terms of broader age groups. In any case, data in the following are assumed to consist of $y_i$ successes from $n_i$ individuals for each observation. The deviance provides an assessment of how well the model predicts the probability of success, $p$, based on a comparison between the likelihood of the fitted model and the likelihood of the *saturated* model, this being a model with one parameter for each data point and fitting the data perfectly. Predicted probabilities from the fitted model are given by

$$\hat{p}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}), \tag{3.6}$$

where $\hat{\beta}_j$ is the maximum likelihood estimate of parameter $j$. The predicted number of successes for data point $i$ is then $\hat{y}_i = n_i \hat{p}_i$ and the deviance is defined as

$$D = 2 \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right]. \tag{3.7}$$

See, for instance, Collett (2002, p. 65). Given a sufficient sample size and non-sparsity in the aggregate data (see below), the deviance has an approximately $\chi^2$ distribution with $n - k$ degrees of freedom under the null hypothesis that the model is correct, $n$ being the number of data points and $k$ the number of parameters (Collett, 2002, p. 69). This leads to the rule of thumb that the deviance should be roughly equal to the degrees of freedom.

Aggregation of the data is required for the deviance to have any meaningful interpretation. For binary data, the $y_i$ are all zero or 1, and the deviance provides no information on the agreement between the observations and those predicted by the fitted model. This is also the case where aggregated data are so finely cross-classified that many of the cell counts are very small: there should be few data points for which $n_i$ is 1 and most of reasonable size (see, e.g., Collett (2002, p. 69)). The deviance should therefore be examined

at an appropriate level of cross-classification, which may mean aggregating groups that are relatively sparse, especially where there is cross-classification with another variable that contains few observations in certain categories. In fact, where data are purely binary, as is the case when continuous covariates are analysed, tests for goodness of fit need to be based on a categorisation of the data, such as the percentiles (often deciles) of the predicted probabilities for the Hosmer-Lemeshow test (Lemeshow and Hosmer, 1982).

Where data are classified according to a single covariate, a lack of fit may indicate that the relationship between the covariate and the outcome has not been captured by the model (provided that the link function is correctly specified). For instance, a linear relationship between the covariate and the linear predictor may not be appropriate, or for grouped variables the classes may be too broad or have inappropriate cut-points. Discrepancies between the observed and predicted probabilities may be assessed for the individual data points via *deviance residuals*, defined as

$$d_i = \text{sign}(y_i - \hat{y}_i) \left[ 2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right]^{1/2} \tag{3.8}$$

for data point $i$ (Collett, 2002, p. 131). Plotting the residuals against the values of the covariates can help to answer questions such as whether the general relationship of a variable has been correctly specified, or whether there are systematic deviations in the fitted probabilities.

For an adequately fitted model with sufficient observations and successes for each $i$, the $d_i$ can be standardised to have an approximate standard normal distribution (Collett, 2002, p. 131). Histograms of deviance residuals can be used to assess this, or the ordered values can be plotted against the values of a standard normal distribution in a so-called quantile-quantile or Q-Q plot: if the $d_i$ have a standard normal distribution, the points will lie on the line $y = x$. If plots of the residuals against covariate values do not reveal any systematic differences, but the standardised deviance residuals are more variable than would be expected under a standard normal distribution then there is likely overdispersion.

Where data are classified according to multiple covariates, a lack of fit may also be due to the presence of interactions between these variables. This is also known as effect modification, where the effect of one variable

changes according to the values of another variable. Of course, model fit can always be improved on by adding interactions. For categorical covariates, the model including all possible interactions is called the *saturated* model, that is, there is a parameter for every observation, and the model fits perfectly. However, it is generally not desirable to do this. Firstly, it is inefficient and will result in imprecise parameter estimates. Secondly, because the model is then not summarising or simplifying the data. Finally, it is often unlikely the underlying data-generating process really includes high-level interactions, with 2nd order usually being the limit of what is considered.

Whether the data include multiple covariates or a single covariate, it may be that the specification of the relationships between the covariates and the outcome, including interactions, is considered sufficient. Excess variability in the observed values of the outcome would then be considered as *overdispersion*. This simply means that prevalence varies more than would be expected by sampling variability alone, given the structure of the linear predictor in the model. For instance, a linear effect of age might appear broadly correct, but observed proportions with the outcome vary substantially between individual years, which would be an obvious case of overdispersion. More difficult to ascertain is whether the additional structure of interaction terms is required, which may require subject-specific knowledge. However, plots of residuals against covariates, stratified by levels of a second covariate, can still be informative in identifying the need for interaction terms. Formal tests for structure in residuals may be carried out, for instance whether there are serial patterns in positive and negative residuals according to age or calendar year. Such tests are referred to as *runs tests*, as they assess the observed vs. expected number of runs of positive and negative values in the data under the null hypothesis of randomness (Wald and Wolfowitz, 1940; Swed and Eisenhart, 1943).

### 3.1.4 Estimates, confidence intervals and predictions

Having obtained a set of parameter values that maximise the log-likelihood, the uncertainty of these estimates due to sampling variability is also of interest: the less information there is (either in terms of small $n_i$, or a small number of successes/failures) then the less precise the estimates will be. An

indication of the uncertainty in the estimate is given by the standard error, which can then be used to construct confidence intervals. The Hessian matrix (2nd derivatives of the log-likelihood) may be calculated numerically by the numerical optimisation routine, and the observed Fisher information matrix is minus the Hessian matrix. The Hessian matrix is the degree of curvature in the log-likelihood surface and therefore represents the amount of information available to estimate the unknown parameters. Thinking of maximum likelihood estimation as searching for the top of a hill, a sharp peak indicates that the true value of a parameter is likely to be in a small region of the parameter space, whereas gentle curvature indicates greater uncertainty in the location of the MLE. Taking the inverse of the negative Hessian matrix, calculated at the MLE, provides an estimate of the asymptotic covariance matrix, with diagonal elements estimates of the asymptotic variances of the MLEs of the parameters. Hence under the assumption of approximate normality 95% confidence intervals for a parameter $\beta$ may be obtained from the standard errors (SE) as $\text{CI}_{95} = \hat{\beta} \pm 1.96\text{SE}$.

The predicted probability of success for a given covariate pattern is given by equation 3.6, with $\hat{p}_i$ obtained by taking the inverse of the link function $g$. To obtain 95% confidence intervals for $g(\hat{p}_i)$, the variance of the linear predictor is given by

$$\text{var}(\hat{\eta}_i) = \sum_{j=0}^{k} x_{ji}^2 \text{var}(\hat{\beta}_j) + \sum_{h=0}^{k} \sum_{j \neq h} x_{hi} x_{ji} \text{cov}(\hat{\beta}_h, \hat{\beta}_j) \tag{3.9}$$

(Collett, 2002, p. 131). The full asymptotic covariance matrix is also obtained from the inverse of the Hessian matrix, but using the whole matrix rather than just the diagonal elements. Denoting the standard error of the linear predictor as $SE(\hat{\eta}_i)$, 95% confidence intervals for the predicted probability of success are given by $\text{CI}_{95} = g^{-1}(\hat{\eta}_i \pm 1.96\text{SE}(\hat{\eta}_i)$.

Plots may then be constructed that include observed and fitted probabilities $p_i$ and $\hat{p}_i$ and the confidence intervals of the fitted probabilities; if the model fit is satisfactory, then around 95% of the observed probabilities should lie within the 95% confidence intervals, hence this is a useful diagnostic procedure.

Alternatively, predictions may be obtained for a certain set of parameter

values that is not necessarily included in the dataset. Continuing with the example of age, region and sex, predictions may be created with region and sex at fixed levels so that changes in predicted probabilities according to age alone may be assessed, which may be preferable to examining a number of plots for each combination of region and sex. If there are no interactions, the pattern will be similar across different levels of region and sex. However, if interactions between age and the other variables are included in the model, it must be borne in mind that the relationship is conditional on the specified values of these covariates. A set of fixed values may be selected for the largest or most representative group(s), or alternatively averaged over the dataset.

If a model has been fitted accounting for imperfect sensitivity and specificity as in equation 3.5, then the predicted probabilities from equation 3.6 will be for true infection status, rather than observed test results. This may be desirable, but plots of observed and predicted probabilities as described in the above will no longer be interpretable, as the observed probabilities will be systematically higher or lower according to the sensitivity and specificity of the test. In this case, it is preferable to adjust the fitted probabilities for the imperfect sensitivity and specificity according to equation 3.5 so that they revert to fitted probabilities of a positive test result, rather than infection status. This procedure may also be used to produce confidence intervals for a positive test result using equations 3.6 and 3.9, then converting the lower and upper bounds according to equation 3.5.

### 3.1.5 Model comparison

The methods outlined in sections 3.1.3 and 3.1.4 may be used to assess the adequacy of model fit, either overall or in terms of systematic differences according to covariate levels. However, it may be that there are multiple candidate models with different sets of covariates or parameterisations (for instance, inclusion of interactions) that are under consideration, from which it is not clear whether a more complex model should be preferred. Differences in deviance may then be used to assess whether one model provides a substantially better fit than another; for two *nested* models, where the more complex model includes all of the parameters of the less complex one, the

difference in deviance between two models has an approximately $\chi^2$ distribution under the null hypothesis that the simpler model is correct (Collett, 2002, p. 73). An appropriate test statistic is constructed with degrees of freedom equal to the difference in the number of parameters between the two models, and the $p$-value provided by the test gives an indication of whether the more complex model should be preferred.

An alternative is to use model selection measures such as the Akaike Information Criterion (Akaike, 1974). This, and other measures like it, gives a score based on how well a model fits the data, but with a penalty for model complexity. For nested models, the performance of the AIC is very similar to deviance-based tests; however, the AIC may also be used to compare non-nested models. The AIC is defined as:

$$AIC = -2 \log(\mathcal{L}(\hat{\theta}|y)) + 2k, \tag{3.10}$$

which is based on the maximised log-likelihood of the estimated parameters, $\log(\mathcal{L}(\hat{\theta}|y))$, plus a penalty term $2k$, where $k$ is the number of parameters in the model. The measure is simple but underpinned by information theory, being based on minimising the information lost when a given model is used to represent the data compared to the true data-generating process. A variety of alternative measures have been proposed, such as the Bayesian (or Schwartz) Information Criterion (BIC), which applies a heavier penalty for complexity. An overview of various approaches is given in Burnham and Anderson (2002), who recommend the AIC as their measure of choice. In particular, the BIC assumes that the correct model is included in the set of candidate models, which is usually implausible, although the AIC may prefer overly complex models when datasets are large (Kuha, 2004). In practice the choice may come down to the consequences of selecting an overly-complex model (AIC) compared to one that is overly-simple (BIC). Much of the model comparison in chapter 4 concerns whether interactions between injecting duration and time are necessary. In the context of then estimating HRs for risk factors, and subsequent investigation of heterogeneity (chapter 6) a more flexible, if potentially over-complex, parameterisation of the baseline FOI may be preferable.

For the AIC (and similar measures) the absolute value has no meaning,

but differences between scores indicate the relative merits of two or more models, with lower scores being better. A rough rule of thumb is that differences of 2 or less indicate little difference between models, differences of 4-7 indicate a likely difference, and more than 10 a substantial difference.

## 3.2 Analysis of current status data

In this section some basic theory of survival analysis is reviewed before applying this theory to current status data. Survival analysis is concerned with *time to event* data, consisting of observation times until an event occurred, or the time of *censoring* if the event did not occur before observation ended. Data therefore consist of a time $t$ and a binary event indicator. The aim is to estimate the hazard rate, which is the instantaneous event rate in survivors (those who have not yet experienced the event) over time, and differences in the hazard rate according to covariates, which can be expressed as hazard ratios (HR). The mathematical relationship between the hazard rate and the proportion of survivors over time has a direct bearing on the analysis of current status data, where time at risk is related to the binary outcome of infection status. In the infectious disease literature the hazard rate is usually called the *force of infection*, and survivors called *susceptibles* (not yet infected); these terms are exactly equivalent. Models that incorporate this relationship for binary current status data may be of the form of a GLM as in section 3.1 and may also incorporate changes over time and other covariates.

### 3.2.1 Basic theory of survival analysis

The hazard rate and survival function are central to survival theory, and are derived as follows (see, e.g., Aalen et al. (2008, p. 6)). Let $T$ denote survival time, and $f(t)$ be its probability density. The cumulative distribution function of $T$ is then

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du.$$

Hence $F(t)$ is the probability of failure by time $t$. The survival function is defined as

$$S(t) = P(T > t) = 1 - F(t) \tag{3.11}$$

and is the probability of survival beyond time $t$. This is an unconditional probability, whereas the hazard rate, $\lambda(t)$, is related to the probability of failure in an infinitesimally small time period between $t$ and $t+\delta t$, conditional on survival up to time $t$. This is defined as the limit:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(t < T \leq t + \delta t | T > t)}{\delta t}. \tag{3.12}$$

From equation 3.12:

$$\begin{aligned} \lambda(t) &= \lim_{\delta t \to 0} \frac{S(t) - S(t + \delta t)}{\delta t S(t)} \\ &= -\frac{1}{S(t)} \frac{\mathrm{d}S(t)}{\mathrm{d}t} \\ &= -\frac{\mathrm{d}}{\mathrm{d}t} \log[S(t)], \end{aligned}$$

which leads to the equation

$$S(t) = \exp\left[-\int_0^t \lambda(u)\ du\right]. \tag{3.13}$$

### 3.2.2 Application to serological surveillance data

The type of data commonly collected by routine surveillance tends to be cross-sectional and does not include the time of the event, only the duration of exposure and current status (hence *interval censored*). However, the same theory from the survival analysis literature applies: the time-specific proportion susceptible is related to the cumulative force of infection, provided the infection confers life-long immunity and serological tests can reliably determine past infection. In many applications infection may occur from birth, and hence the time at risk is the person's age. Subsequently models are described in terms of age, but the same concepts hold for any measure of time since the individual became at risk.

Given data from a serological survey in which $r_t$ of $n_t$ individuals at age $t$ are infected, the force of infection $\lambda(t)$ may be estimated by maximising

the binomial log-likelihood as in equation 3.4, with $S(t)$ corresponding to the proportion of successes $p_i$ (considering *susceptible* a "success"); see, e.g., Farrington et al. (2001). In the simplest case of a constant force of infection over time ($\lambda(t) = \lambda$), we have:

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda \, du\right] \\ &= \exp(-\lambda t). \end{aligned}$$

The relationship is that of a GLM for binomial data with a log link

$$\log(S(t)) = -\lambda t \tag{3.14}$$

and models of this form can therefore be fitted using standard GLM routines in many statistical software packages (although some will not allow a log link with binomial data). Although a log link is not generally recommended for binomial data, given that $\lambda$ must be positive, the linear predictor will always be negative, so $S(t)$ cannot exceed 1 (and is equal to 1 at $t = 0$). Model fitting routines may be adapted to allow for constraints, which can be used to ensure that the force of infection takes only non-negative values. However, this is often not necessary as the function may be parameterised such that the force of infection is always positive (e.g., by exponentiating). Note that the linear predictor in equation 3.14 does not include a constant, as this would imply non-zero prevalence (proportion susceptible $\neq 1$) at birth. However, the presence of a constant term would give an indication of imperfect test sensitivity or specificity (Ades and Nokes, 1993), given the assumption of constant force of infection.

Equation 3.14 may also be arranged in the form of the complementary log-log (CLL) link function

$$\log(-\log(S(t))) = \log(\lambda) + \log(t). \tag{3.15}$$

This parameterisation may be preferable if a log link is not permitted by the software package to be used, but otherwise gives the same mathematical relationship between the parameters and the data.

The model described in equation 3.14 may easily be extended to a *piece-*

*wise constant* model, where the force of infection is constant within age bands. With cut-points $a_0$ (usually zero), $a_1$, $a_2...a_k$, the duration of time spent within age band $i = 1, 2...k$ is defined as

$$A_i(t) = \begin{cases} 0, & \text{if } t \leq a_{i-1}; \\ t - a_{i-1}, & \text{if } t > a_{i-1} \text{ and } t \leq a_i; \\ a_i - a_{i-1}, & \text{if } t > a_i. \end{cases}$$

For a set of age bands 1, $2...k$ with cut-points $a_0$, $a_1$, $a_2...a_k$ and force of infection $\lambda_i$ in age band $i$, the survivor function is defined as:

$$S(t) = \exp[-(\lambda_1 A_1(t) + \lambda_2 A_2(t) + ... + \lambda_k A_k(t))]. \tag{3.16}$$

This also has the form of a GLM with log link, although unlike the constant model, cannot be arranged to have a CLL link. Piecewise constant models may be useful when risk periods are well-defined, such as pre-school, junior school, secondary school and adulthood; or, if a suitably large number of groups are used, the piecewise constant approach avoids making assumptions of a certain parametric shape. However, the choice of the number and locations of cut-points may be somewhat arbitrary, and modelling the shape of the hazard function may require a higher number of parameters than necessary; in which case, a parametric function for $\lambda(t)$ may be preferred. For example, a simple choice is an exponential decline model, where risk declines asymptotically towards zero. With $\lambda_0$ the force of infection at $t = 0$ and $\rho$ the rate of decline:

$$\lambda(t) = \lambda_0 \exp(-\rho t).$$

Substituting this into the formula for the proportion susceptible gives:

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \lambda_0 \exp(-\rho u)\ du\right] \\ &= \exp\left[\frac{\lambda_0}{\rho}(\exp(-\rho t) - 1)\right]. \end{aligned}$$

This cannot be arranged in the form of a linear model, and hence it is not possible to estimate using standard GLM routines. In this case, the likelihood must be evaluated using non-linear optimisation methods as described

in section 3.1.1. However, some other specific model forms may be fitted within the GLM framework, such as the Weibull model, that allows for a monotonically increasing or decreasing hazard over time and can be fitted with a CLL link.

No matter what the functional form is used to characterise $\lambda(t)$, predictions for the proportion susceptible may be estimated as in section 3.1.3, and the number of susceptibles substituted for the number of successes. Hence all of the apparatus in section 3.1 is available for checking model fit, producing predictions and obtaining confidence intervals.

### 3.2.3 Models for age and time

For many infectious diseases, changes in hygiene, sexual behaviour and other risk factors over time are likely to result in changes in the force of infection. In the presence of such temporal effects, older individuals will have experienced a different force of infection in their earlier years to younger individuals, in addition to any age-specific effects. When data are collected from a single serological survey in the form of age-specific disease status the age-specific risk and any temporal effects are completely confounded. However, when multiple surveys are undertaken at different time points, it is possible to estimate these components separately (Ades and Nokes, 1993). For age $a$ and survey time $t$, an age- and time-specific hazard $\lambda(a,t)$ is defined and

$$S(a,t) = \exp\left[-\int_0^a \lambda(u, t-a+u)\ du\right].$$

In the case of piecewise constant models, the force of infection may consist of age components $\mu_i$ for age band $i$ and time components $\phi_j$ for time band $j$. With $A_1(a)$, $A_2(a)...A_k(a)$ as the durations of time spent in age band 1, 2...$k$, indexed by age $a$, as before, a similar function for calendar time, indexed by $t$ with cut-points $t_0$, $t_1$, $t_2...t_m$ is defined as

$$T_j(a,t) = \begin{cases} 0, & \text{if } t-a > t_j \text{ or } t < t_{j-1}; \\ min(t,t_j) - max(t-a, t_{j-1}) & \text{if } t-a \leq t_j \text{ and } t \geq t_{j-1}, \end{cases}$$

this being the duration of time spent in time band $j$. In the framework of Ades and Nokes (1993) the effects of age and time combine additively to produce the force of infection as $\lambda(a,t) = \mu_i + \phi_j$ for age band $i$ corresponding to age $a$ and time band $j$ for time $t$. The cumulative hazard is then related to the survivor function according to the age- and time-specific contributions of the two functions. For a set of age bands $i = 1, 2...k$ and time bands $j = 1, 2...m$,

$$S(a,t) = \exp\left[-\left(\sum_{i=1}^{k} \mu_i A_i(a) + \sum_{j=1}^{m} \phi_j T_j(a,t)\right)\right]. \tag{3.17}$$

This model assumes that the time and age effects combine additively and are independent, i.e., there is no interaction corresponding to a change in age-specific force of infection over time. Such an assumption could be relaxed by the incorporation of interaction terms, or equivalently, by specifying age-time specific bandings. In either case, the model can be fitted using standard GLM routines. Alternatively, a model with multiplicative age and time effects for the force of infection would require $\lambda(a,t) = \mu_i \phi_j$, which can no longer be fitted as a standard GLM. In general such models will require bespoke code.

To illustrate the information contained in age-specific current status collected at different time points, a Lexis diagram is displayed in Figure 3.2. Clearly, if observations were at a single time point (e.g., the year 2000) there would be no information to differentiate between age and time effects.

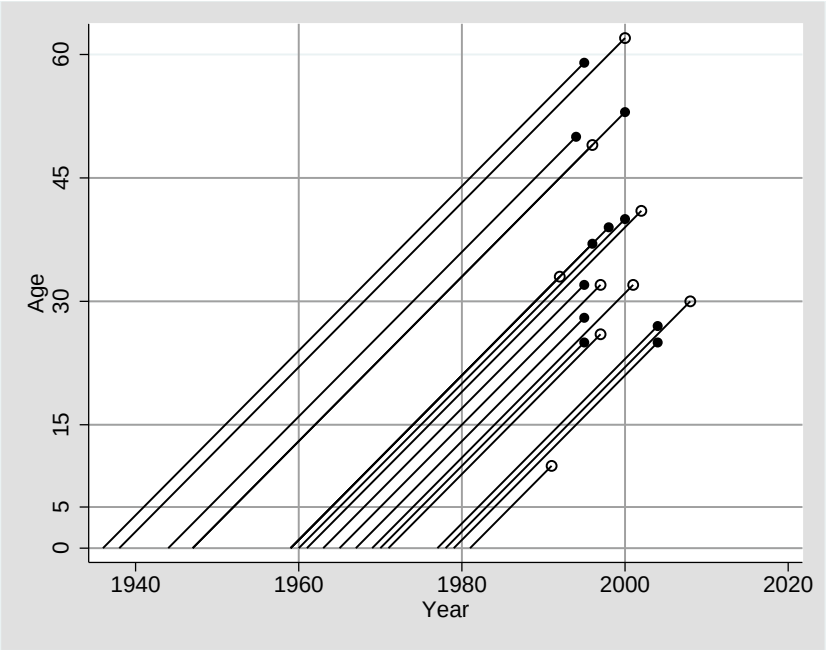


Figure 3.2: Lexis diagram of individual exposure by age and time. Circles denote time of observation (between 1991 and 2008), solid are infected, hollow uninfected. Age bands with cut-points 0, 5, 15, 30 and 45, and time bands with cut-points 1960, 1980 and 2000 are shown on the figure.

Using the age and time bands in Figure 3.2, an individual born at the start of 1958 and observed at the end of 2006 would have the following piecewise constant age-time contributions (in years):

| Time | Age | | | | | |
|---|---|---|---|---|---|---|
| | 0-5 | 5-15 | 15-30 | 30-45 | 45+ | Total |
| Pre-1960 | 2 | 0 | 0 | 0 | 0 | 2 |
| 1960-1980 | 3 | 10 | 7 | 0 | 0 | 20 |
| 1980-2000 | 0 | 0 | 8 | 12 | 0 | 20 |
| 2000 onwards | 0 | 0 | 0 | 3 | 4 | 7 |
| Total | 5 | 10 | 15 | 15 | 4 | 49 |

Using a parameter for each individual cell corresponds to an age and time-specific force of infection, whereas using row and column totals would imply independence. Note that there is only information available to estimate age

and time interactions where data are available; for instance, in the example above no individuals were of age 30-45 or 45+ prior to 1960, nor of age 45+ in 1960-1980; therefore the force of infection in these bandings would have to be extrapolated from other time periods or age bands. Note that although no information on the interaction is provided outside the range of the data, there is however a restriction on the upper bound that the FOI can take, as the FOI in the pre-survey period must be consistent with the observed prevalence, which should not decrease with increasing exposure time.

Some thought is required as to whether the temporal effect or the relative effect of age is assumed constant outside the range of the data. In the following, models are considered in terms of a time-specific FOI $\lambda_0(t)$ at a baseline age $a_0$ and a function for age-specific modifications to the baseline hazard, which may be independent of time, or allow for interactions within the survey period. Within this framework the age-specific FOI can be estimated along with any changes in this pattern during the survey, and subject to the assumption of constant age effects outside the survey period, historical changes in the FOI can be considered. Therefore $\lambda(a,t) = \lambda_0(t) + D(a,t)$, where $D(a,t)$ is the time-specific additive difference in the FOI at age $a$ vs. age $a_0$ (and $D(a_0,t) = 0$), giving

$$S(a,t) = \exp\left[-\int_0^a (\lambda_0(t-a+u) + D(u,t-a+u))\ du\right], \tag{3.18}$$

which may be fitted as a piecewise constant model using standard GLM routines. The general form for multiplicative age- and time-specific differences in the FOI, with $R(a,t)$ the ratio in the FOI at age $a$ vs. age $a_0$ (and $R(a_0,t) = 1$) is

$$S(a,t) = \exp\left[-\int_0^a \lambda_0(t-a+u)R(u,t-a+u)\ du\right], \tag{3.19}$$

the piecewise constant form of which cannot be fitted using standard GLM packages.

The components of these models may be split further into main effects for time and age and a separate age-time interaction. This would be mathematically equivalent to the models in equations 3.18 and 3.19 (any set of estimates from one model can be transformed to another) but with a different

interpretation of the parameters. More importantly, this specification allows for the necessary constraint that there is no age-time interaction outside the range of the data.

Careful consideration is needed as to whether age and time effects are independent, and whether they combine additively or multiplicatively, as decisions on the latter will produce different estimates for the independent models. This can of course be assessed by comparing the model fit under additive and multiplicative assumptions. However, models that include the full set of estimable age-time interactions are saturated in a sense, as there is a parameter for every combination of the age and time bands. Estimates of the proportion susceptible will then be the same for both the additive and multiplicative model, as will the FOI within the range of the survey.

Another consideration is that the multiplicative model, in which age-time contributions are exponentiated to obtain the FOI, is constrained to have a positive FOI, but the additive model is not. Additive models therefore permit negative values for the FOI where prevalence decreases with time at risk within an age or age-time banding. The likelihood of this occurring increases with model complexity, as with a large number of age/time bands or interactions it becomes more likely that observed prevalence is not strictly increasing for all age-time combinations. Piecewise constant models for age alone can be constrained so that the force of infection for each age band $i$ is non-negative; however, constraining the *sum* of age-time contribution parameters in the framework of Ades and Nokes (1993) is not supported by most statistical software packages or optimisation routines. It can of course simply be taken that negative estimates of the FOI are equivalent to zero, but this issue also affects model fit statistics. In particular, allowing violations of monotonically increasing prevalence with time at risk in additive models but not multiplicative ones could skew comparisons of the two model types in favour of additive models.

### 3.2.4 Parametric models

Parametric forms for the force of infection can be useful in that they require fewer parameters to model a particular shape compared to the step-wise changes of a piecewise constant model, particularly if expert knowledge in-

dicates that a particular pattern in the age- or time-specific FOI is likely. Some choices of parametric function can provide a certain degree of flexibility in addition to easily interpretable quantities: the exponentially damped line rises to a peak before tailing off to some constant level (Farrington, 1990). This is a likely pattern for childhood infections, and the parameter values determine the size of the peak, how long this period of elevated risk extends for, and the level of risk that remains in adulthood. A selection of parametric shapes are shown in Table 3.1 and Figure 3.3.

Table 3.1: A selection of parametric hazard functions discussed in Farrington (1990) and Ades and Nokes (1993).

| Name | Function | Values in figure 3.3 |
|---|---|---|
| Exponential polynomial (1) | $\exp(-a - bt)$ | $a = 0.8, b = 0.2$ |
| Exponential polynomial (2) | $\exp(-a - bt - ct^2)$ | $a = 0.9, b = 0.1,$ $c = 0.05$ |
| Exponentially damped line | $(at - c)\exp(-bt) + d,$ $b \geq 0$ | $a = 0.5, b = 0.45,$ $c = 0.025, d = 0.025$ |
| Gompertz function | $a\exp[-\exp(-b(t - c))] + d,$ $a, d > 0$ | $a = 0.15, b = 0.5,$ $c = 10, d = 0.01$ |
| Symmetric logistic | $a/[1 + \exp(-b(t - c))] + d,$ $a, d > 0$ | $a = 0.15, b = 0.5,$ $c = 10, d = 0.01$ |

These functions may be deemed epidemiologically plausible for certain applications, but in the absence of such knowledge, or more complex patterns of risk according to age or time, might be too restrictive to capture the true shape of underlying hazard. Changes over time may be particularly unpredictable: more than one peak may occur, or a peak, then decline, followed by a later, steady rise, or any other shape. It will then become increasingly difficult to find a smooth function that will fit the data well and still have any sensible interpretation. Polynomial or fractional polynomial models may provide a good fit to the data (Royston, 2000), but the interpretation of parameters beyond a linear trend is difficult. There is also the difficulty of specifying interactions between age and time; Ades and Nokes (1993) assume independent parametric functions for age and time, but the

data may indicate that they do not combine independently, and defining a suitable parametric function in two dimensions may be challenging.

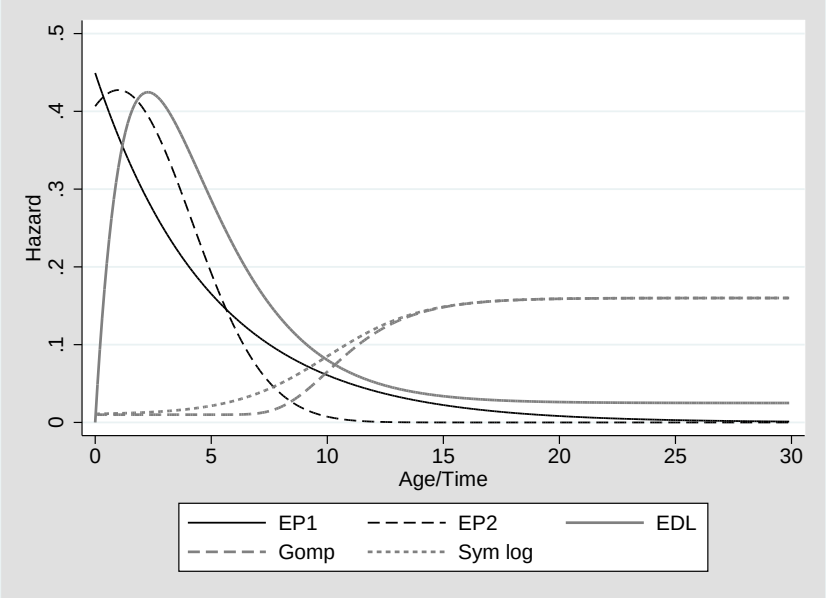


Figure 3.3: A selection of parametric hazard functions discussed in Farrington (1990) and Ades and Nokes (1993). *EP1* and *EP2*: exponential polynomials of order 1 and 2 (no constant), *EDL*: exponentially damped line, *Gomp*: Gompertz, *Sym log*: symmetric logistic. The latter two may rise or fall between two asymptotes.

An alternative is the use of spline functions, piecewise polynomial functions (usually cubic) that can provide any level of flexibility, often including a smoothing component to avoid over-complexity. Such models have been applied to current status data, providing a smooth function for the hazard or cumulative hazard (Namata et al., 2007; Nagelkerke et al., 1999). The flexibility of these models is appealing, but again can be challenging in two dimensions.

### 3.2.5 Covariates for force of infection models

Section 3.2.3 shows how the basic model for age-specific force of infection may be extended to include changes over time, with the combination of age- and calendar time-specific time at risk acting as a special type of covariate. Additional covariates may also be included in the model to allow changes in the FOI due to other factors. The effects may be assumed to modify

the FOI multiplicatively at all ages and/or times, which would be described as *proportional hazards* in survival analysis terminology. Alternatively, the covariate may have an interaction with age, time, or both, such that the effect of the covariate on the FOI can vary. Finally, if the covariate itself changes over time, then the FOI may be modified according to the *time-varying* nature of the covariate. An example of this would be the age of vaccination or uptake of some other intervention or behaviour. In this case the FOI can change between the pre- and post-intervention periods defined by the time-varying covariate (TVC). A TVC may also act proportionally at all times and ages, or have a varying effect.

For a covariate with proportional hazards, the FOI at age $a$, time $t$ and covariate level $x$ is defined as

$$\lambda(a, t, x) = \lambda_0(a, t) \exp(\beta x), \tag{3.20}$$

where $\lambda_0(a, t)$ is the force of infection at the baseline level of $x$, which should be set to zero for the baseline FOI to have a sensible interpretation. This will mean centering continuous variables by subtracting the desired baseline value. For categorical variables standard practice is to treat the baseline as zero and other groups via 0/1 indicator variables.

As with the age-time models in section 3.2.3, the multiplicative form of the covariate model cannot be fitted in the GLM framework, but covariates with additive effects can. In this case

$$\lambda(a, t, x) = \lambda_0(a, t) + \beta x, \tag{3.21}$$

again, with the baseline value of $x$ being zero. Considering a model with piecewise constant age-specific FOI and a covariate $x$ with constant, additive differences, yields an extension to the survivor function in equation 3.16

$$S(a, x) = \exp[-(\lambda_1 A_1(a) + \lambda_2 A_2(a) + ... + \lambda_k A_k(a) + \beta x a)]. \tag{3.22}$$

Therefore the addition of the covariate is achieved by including a variable that multiplies the individual's age, $a$ by the covariate $x$ in the GLM, rather than simply including $x$ in the list of variables for inclusion as one would for a logistic regression model.

Interactions between age and the covariate, or age-specific covariate effects, in which $\lambda(a,x) = \lambda_0(a) + \beta(a)x$ can also be included in a GLM form by multiplying the covariate by the time spent in age band $i$ rather than the total time at risk. The survivor function for this model would therefore be

$$S(a) = \exp[-((\lambda_1+\beta(1)x)A_1(a)+(\lambda_2+\beta(2)x)A_2(a)+...+(\lambda_k+\beta(k)x)A_k(a))]. \tag{3.23}$$

Inclusion of covariates in age-time models follows readily from this within the additive framework, and may be parameterised with fixed differences in the FOI at all times and ages, or with interactions with age, time or both by relating covariates to specific ages and times and the duration spent within each band. The most general form of covariate model with FOI dependent on age, time and covariate effect $\lambda(a,t,x)$ is therefore

$$S(a,t,x) = \exp\left[-\left(\int_0^a \lambda_0(u, t-a+u, x)du\right)\right], \tag{3.24}$$

with any simpler model obtained by specifying independent contributions of age and time effects in the FOI, either multiplicatively or additively. As noted in section 3.2.3, increasing complexity of the model can result in negative FOI estimates in additive models, and this will become increasingly likely if complex interactions are considered.

Time-varying covariates (TVC) are a further extension of the above framework in which the covariate itself changes over time. Current status data are by definition collected at a single time point and would not usually include detailed time-varying information in the way that a cohort study might, where individuals are followed up over time. TVCs in current status data are therefore likely to consist of dichotomous changes at a single age/time point, rather than detailed histories. An age-specific FOI with multiplicative effect of a covariate $x$ which changes from 0 to 1 at time $a_x$ can be expressed as

$$\lambda(a, a_x) = \begin{cases} \lambda_0(a), & \text{if } a < a_x; \\ \lambda_0(a)\exp(\beta), & \text{if } a \geq a_x. \end{cases} \tag{3.25}$$

The additive model is defined similarly with the parameter $\beta$ being added to

the baseline hazard. Again, in a piecewise constant framework the model can be fitted as a GLM by subdividing the age bands into pre- and post-TVC, and similarly for age and time models.

### 3.2.6 Concluding remarks

This chapter has outlined the basic tools for fitting models to binary data, assessing their adequacy and between-model comparisons. These tools may be applied to current status data via generalised linear models (GLM) where age and time effects for the force of infection are modelled as piecewise constant and combine additively. Some parametric forms are also available in the GLM framework (such as the Weibull model with complementary log-log link) but are not considered further. There are a number of reasons for preferring piecewise constant models here: firstly, parametric models place restrictions on the assumed shape of the FOI, which would require a priori knowledge to justify. Secondly, the effect of covariates or age-time interactions become difficult to parameterise. Thirdly, the focus of later chapters shifts to heterogeneity; estimates of the FOI are somewhat less important, but functions for the baseline FOI must be sufficiently flexible so that they do not distort estimates of the frailty parameters.

It must be borne in mind however that age-time interactions may result in large numbers of model parameters and model instability where data are sparse. Resulting standard errors should therefore be checked carefully, and simpler models considered as required. Another option is Bayesian methods, under which semi-informative priors could be placed on model parameters to ensure baseline FOIs or age/time-specific HRs are within a plausible range. For instance, baseline rates may be low but not effectively zero, HRs within bounds of 0.1 to 10, and so on; see, for instance Greenland (2001). Of course, where there is little information to estimate a parameter, results may then be sensitive to the choice of prior.

GLMs for the force of infection can also include covariates in the piecewise constant framework, again assuming additivity. Implementation in standard statistical packages could be advantageous for exploring changes in risk according to a number of possible factors, as the inclusion of a candidate covariate in a model is generally quick to implement. In contrast, multi-

plicative effects are not possible to parameterise in the form of a GLM, and must therefore be fitted by maximising the likelihood using bespoke code. This is not only simply a software limitation: the non-linear combination of parameters means that a general-purpose specification is difficult to achieve. Therefore the GLM framework may be used for exploratory purposes, guiding the choice of which covariates to use and how to parameterise them, before building a multiplicative model if required.

# Chapter 4

# Blood-borne viruses in people who inject drugs: trends and risk patterns

In this chapter the methods described in chapter 3 for analysing cross-sectional surveillance data are applied to data collected by the Unlinked Anonymous Prevalence Monitoring Programme (UAM). These data include current infection status, time at risk, and a large number of self-reported risk factors. Careful analysis of the UAM data may help to gain insights into patterns of risk behaviour, informing public health policy and preventative interventions. This section also aims to identify key covariates for inclusion in models of heterogeneity in subsequent chapters.

Prevalence of blood-borne viruses (BBV) according to injecting duration, the time at risk (analogous to age), calendar time and reported risk factors is examined via generalised linear models (GLM). Models are then fitted that incorporate the relationship between prevalence and time at risk via the force of infection (FOI), as in section 3.2, to understand how the FOI changes according to injecting duration and calendar time. Finally, models that incorporate changes in the FOI according to different covariates are developed.

## 4.1 Prevalence and risk patterns of blood-borne viruses

Between 1990 and 2014, 43,002 participants were sampled with information on injecting duration and a test result for HCV, HBV or HIV. Of those sampled, 13,383 (31.1%) had participated in the survey before. 32,784 (76.2%) participants were male. The median age of those taking part in the survey was 30 (IQR: 25-36, range: 16-68) and the median age at first injecting was 20 (IQR: 17-25, range: 12-63). The median length of injecting duration was 8 years (IQR: 3-14, range: 1-53). The overall mean prevalence of infection with HCV, HBV and HIV was 42.1%, 20.1% and 1.1% respectively. Observed prevalence was relatively high in those injecting for one year for all infections, rising from 17.2% for those injecting for 1 year or less to 47.7% after 10 years for HCV. For HBV, prevalence in first-year injectors was 6.0%, rising to 22.4% after 10 years; and for HIV prevalence in first-year injectors was 0.5%, rising to 1.4% after 10 years. 10-year results are for average prevalence in those injecting 10-12 years. In the following, all participants were assumed to have been injecting for at least 1 year, rounding the injecting duration upwards where the current age is the same as age first injected.

Figure 4.1 shows the prevalence of HCV, HBV and HIV according to injecting duration in different survey periods. In all periods for HCV and HBV and more recent periods for HIV, there is a relatively abrupt increase in prevalence from 0 to 1 year of injecting duration, after which prevalence increases more slowly. For HCV, prevalence generally decreased over time from 1990 to 2000 and increased again recently (as noted by Sweeting et al. (2009b)). After the first year of injecting, HCV prevalence according to injecting duration has a pattern that is consistent with a constant FOI, i.e., $S(t) = \exp(-ct)$. Conversely, HBV prevalence decreases over the survey period, and appears less likely to have a constant force of infection, especially in recent survey years: the flatter trajectory does not seem to suggest a constant FOI. These observations point to temporal changes in the FOI and potential interactions between time and injecting duration.

For HIV, patterns of prevalence are more irregular as there are far fewer infections, although there are some visible patterns. In the earliest 1990-1994 survey period, prevalence was low in those with injecting durations of up to

7-8 years but increased markedly for longer durations. These individuals would have been at risk during the peak of the 1980s epidemic, before harm reduction campaigns were introduced. This peak in infections is evident in later survey years at longer injecting durations, with prevalence generally remaining low in those that had not started injecting prior to 1990. However, prevalence does decrease in this cohort over time, which could potentially be due to a selection effect from higher mortality rates in those infected early in the epidemic (see, e.g., Ades and Medley (1994)). Of note is that prevalence at shorter injecting durations rises somewhat in the surveys from 2005 onwards, indicating an increased risk of infection in newer injectors (as noted by Hope et al. (2014)).

### 4.1.1 Generalised linear models for HCV by age and time

The relationship between injecting duration and calendar time and HCV prevalence was modelled via a GLM with a logistic link, producing estimated odds ratios (OR) for different covariate levels in comparison to a baseline category. Injecting duration was categorised with cut-points of 1, 2, 3, 5, 7, 9, 12, 15, 20, 25, 30 and 35. As prevalence is expected to increase continuously with exposure time, a relatively large number of groups are required to accurately capture the relationship between injecting duration and prevalence. The chosen cut-points result in approximately equal-sized groups up to 15 years (ranging from 3262-5699) but numbers injecting for longer durations fall rapidly beyond this. Fewer categories were used for survey year, with a four-year group for the earliest survey years, 1990-1993, then three-year groups subsequently: 1994-1996, 1997-1999 ... 2012-2014; again, each group is of roughly equal size (4500-6359).

Table 4.1 shows resulting model coefficients for a GLM with logistic link fitted to the HCV prevalence data, with main effects for injecting duration (baseline 5-6 years) and survey period (baseline 2006-2008). The constant term for log odds in the baseline category is -0.79 (95% CI -0.87, -0.70) indicating HCV prevalence of around 31% in those injecting for 5-6 years in survey years 2006-2008. As expected, the ORs show a monotonic increase in prevalence with injecting duration. Risk in the first year of injecting is

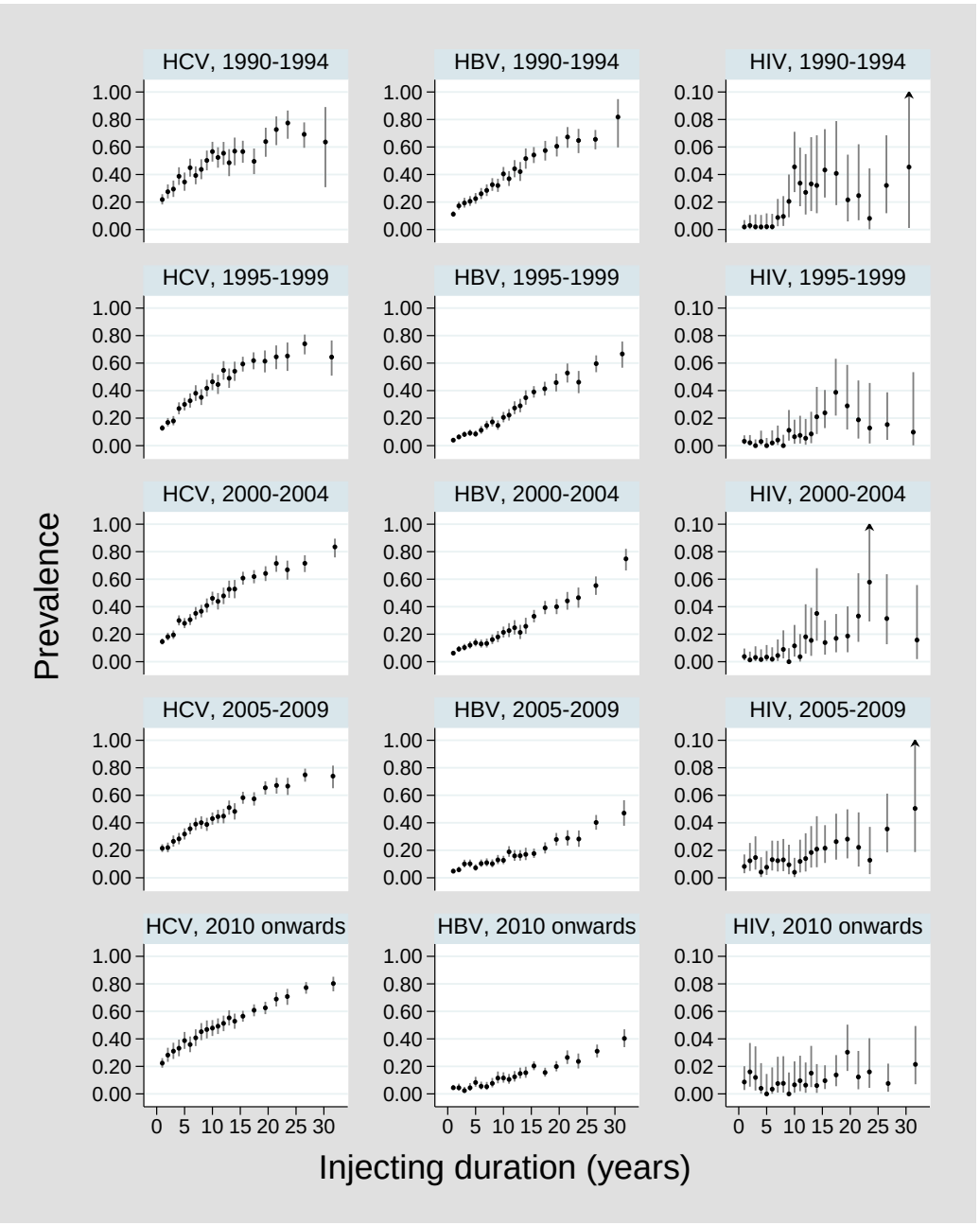


Figure 4.1: Prevalence of HCV, HBV and HIV by injecting duration in different survey periods, point estimates and 95% binomial confidence intervals. Injecting durations are grouped at longer durations to maintain adequate size groups, and durations longer than 35 years are omitted for clarity.

obviously high, with the OR vs. 5-6 years indicating only around half the odds (0.44) of infection at 1 year. The effect for time is more interesting, with higher prevalence in 1990-1993 vs. 2006-2008 (OR=1.57), followed by a possible decline (OR=0.93 for 2000-2002); then a further rise in recent years (OR=1.27 for 2012-2014). This pattern was noted by Sweeting et al. (2009b), who found a similar pattern after adjusting for additional covariates and the imperfect sensitivity of earlier tests. The true prevalence in early survey years is likely to be higher than indicated by the ORs in Table 4.1 due to imperfect sensitivity.

Table 4.1: Model coefficients for injecting duration and survey period, main effects model fitted to HCV data, exponentiated to give odds ratios (ORs) (except constant).

| | OR | 95% CI | $Z$-val | $p$-val |
|---|---|---|---|---|
| Injecting duration | | | | |
| 1 | 0.44 | 0.39, 0.49 | -15.39 | <0.001 |
| 2 | 0.54 | 0.48, 0.61 | -10.25 | <0.001 |
| 3-4 | 0.74 | 0.67, 0.82 | -6.10 | <0.001 |
| 5-6 | 1 (ref) | | | |
| 7-8 | 1.28 | 1.16, 1.41 | 5.00 | <0.001 |
| 9-11 | 1.66 | 1.52, 1.81 | 11.22 | <0.001 |
| 12-14 | 2.10 | 1.92, 2.31 | 15.71 | <0.001 |
| 15-19 | 2.89 | 2.64, 3.16 | 23.10 | <0.001 |
| 20-24 | 4.22 | 3.79, 4.69 | 26.40 | <0.001 |
| 25-29 | 5.72 | 4.94, 6.62 | 23.47 | <0.001 |
| 30-34 | 6.76 | 5.47, 8.36 | 17.66 | <0.001 |
| 35+ | 9.74 | 6.83, 13.90 | 12.56 | <0.001 |
| Survey period | | | | |
| 1990-1993 | 1.57 | 1.40, 1.77 | 7.56 | <0.001 |
| 1994-1996 | 1.04 | 0.95, 1.14 | 0.89 | 0.376 |
| 1997-1999 | 0.99 | 0.91, 1.08 | -0.20 | 0.845 |
| 2000-2002 | 0.93 | 0.86, 1.00 | -1.86 | 0.063 |
| 2003-2005 | 1.18 | 1.09, 1.28 | 3.94 | <0.001 |
| 2006-2008 | 1 (ref) | | | |
| 2009-2011 | 1.19 | 1.09, 1.29 | 4.01 | <0.001 |
| 2012-2014 | 1.27 | 1.16, 1.38 | 5.49 | <0.001 |
| Constant | -0.79 | -0.87, -0.70 | -18.21 | <0.001 |

The resulting deviance from the main effects model presented in Table 4.1 was 962 on 768 degrees of freedom, which indicates that the model fit is not quite satisfactory; one would hope for the deviance to be roughly equal to the degrees of freedom (section 3.1.3). Including interaction terms for injecting duration category and survey period provides an indication of whether the lack of fit is due to systematic changes in the relationship between prevalence and injecting duration according to survey period. The interaction model gave a deviance of 730 on 692 degrees of freedom, which is an acceptable fit. However, this model uses 117 parameters and only a handful of interaction terms are significant, so it is rather inefficient. Figure 4.2 shows the observed data by year and injecting duration and the model fit of the main effects and interaction models. In some places the interaction model is clearly performing better, particularly between 1992-1998. However, the interaction model does not look as if it provides a better fit to the data for later years. This is further borne out by plots of the residual deviances, shown in Figure 4.3. There appear to be some patterns to the residuals between 1990-1998, with something of a downward trend for 1992 for injecting duration between 1-18 years vs. a sharp upward trend in 1998. A runs test indicated non-randomness of residuals across injecting duration for the main effects model ($p = 0.007$), although not for the interaction model ($p = 0.260$). In subsequent years both models have a fairly random scatter. This is of note, as many key covariates are only available from 2000, so the interaction may not be necessary for covariate models that only use more recent data. Finally, Figure 4.4 shows histograms and quantile-quantile (QQ) plots of the residuals from main effect and interaction models. Both have an approximately normal distribution, but the main effects model has variance greater than 1 and both models exhibit divergence from the theoretical distribution at the tails.

Re-fitting the models to data from 2000 onwards, the deviance for the main effects model is 659 on 601 degrees of freedom and the interaction model 555 on 557 degrees of freedom; a moderately good fit for the main effects and very good for the interaction model. However, both models still show some visible patterns in the residuals, particularly for those injecting for less than 10 years in the year 2000, and some extreme outliers. Although the overall fit is better for the interaction model, the most extreme outliers generally

occur in both models. Figures 4.5, 4.6 and 4.7 show observed/predicted HCV prevalence, deviance residuals by injecting duration and calendar time and the distributions of residuals respectively under the main effects and interaction model for data from 2000 onwards.

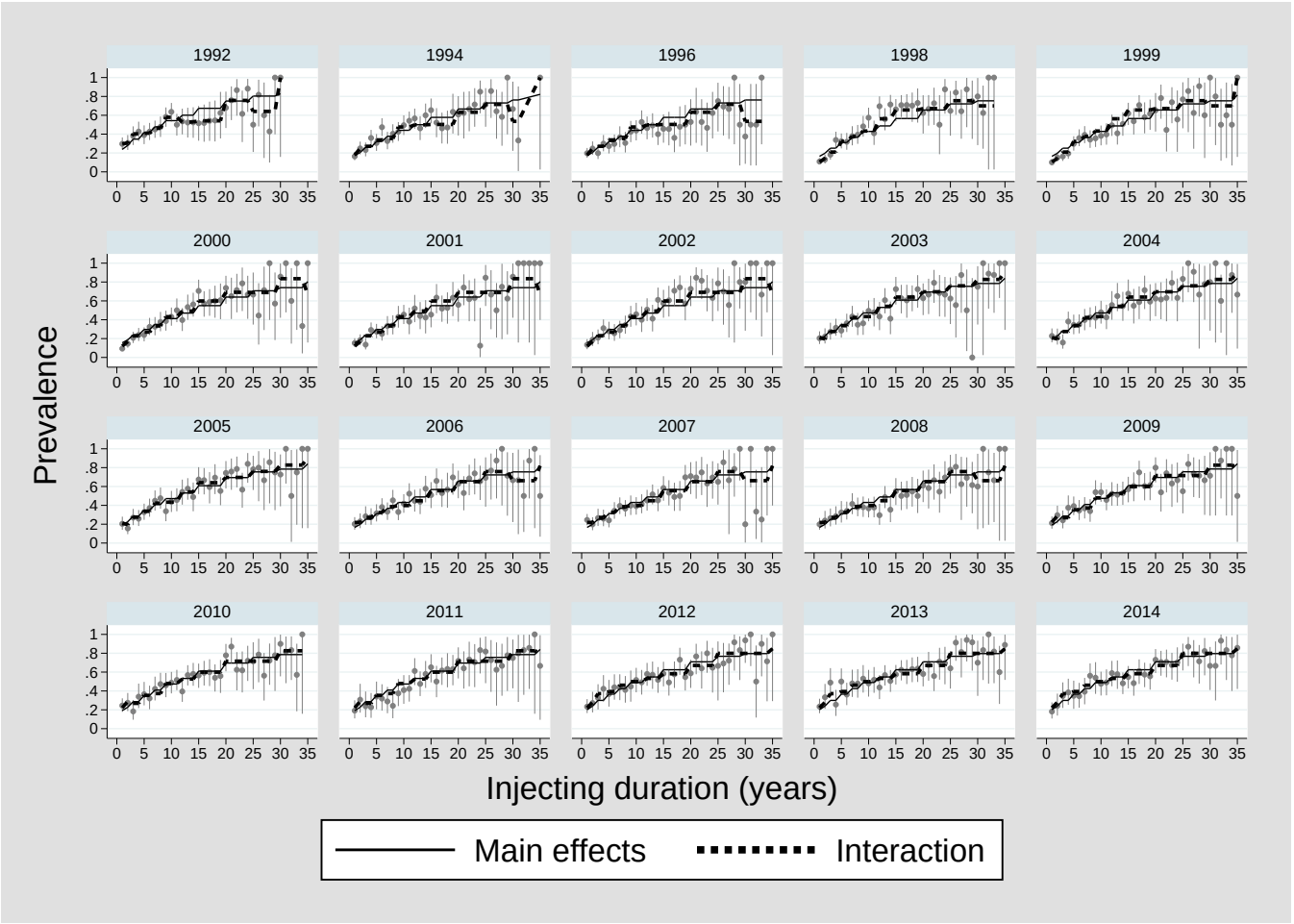


Figure 4.2: Observed and predicted HCV prevalence by injecting duration and survey year; main effects of injecting duration and calendar time, and interaction model. 95% binomial confidence intervals are displayed around the observed prevalence. Note that testing for HCV was only conducted for a subset of participants in 1992, 1994 and 1996 prior to 1998.

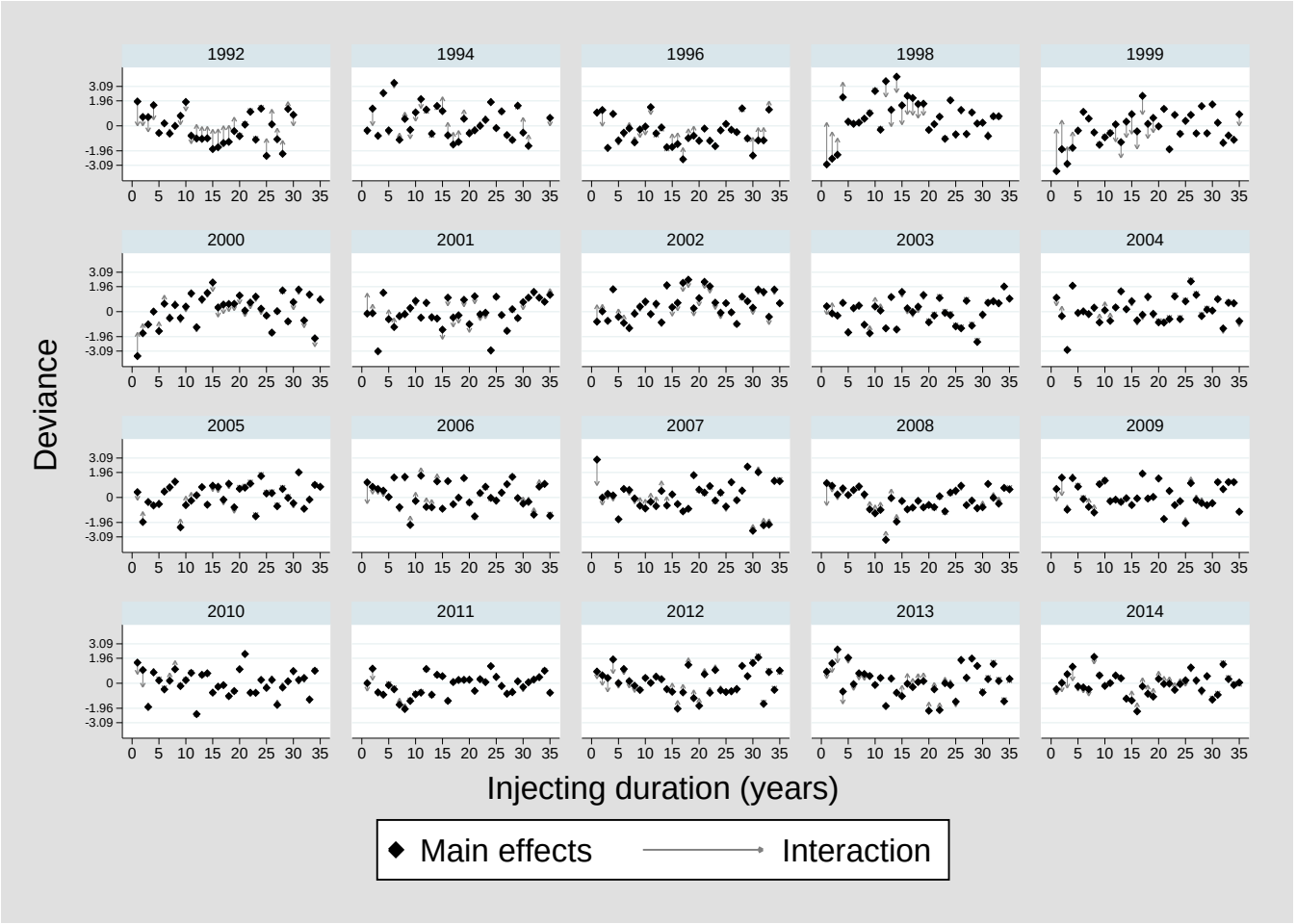


Figure 4.3: Deviance residuals by injecting duration and survey year from main effects and interaction models fitted to HCV prevalence data. Reference lines are the 2.5/97.5th and 0.1/99.9th percentiles of the standard normal distribution. Note that testing for HCV was only conducted for a subset of participants in 1992, 1994 and 1996 prior to 1998.

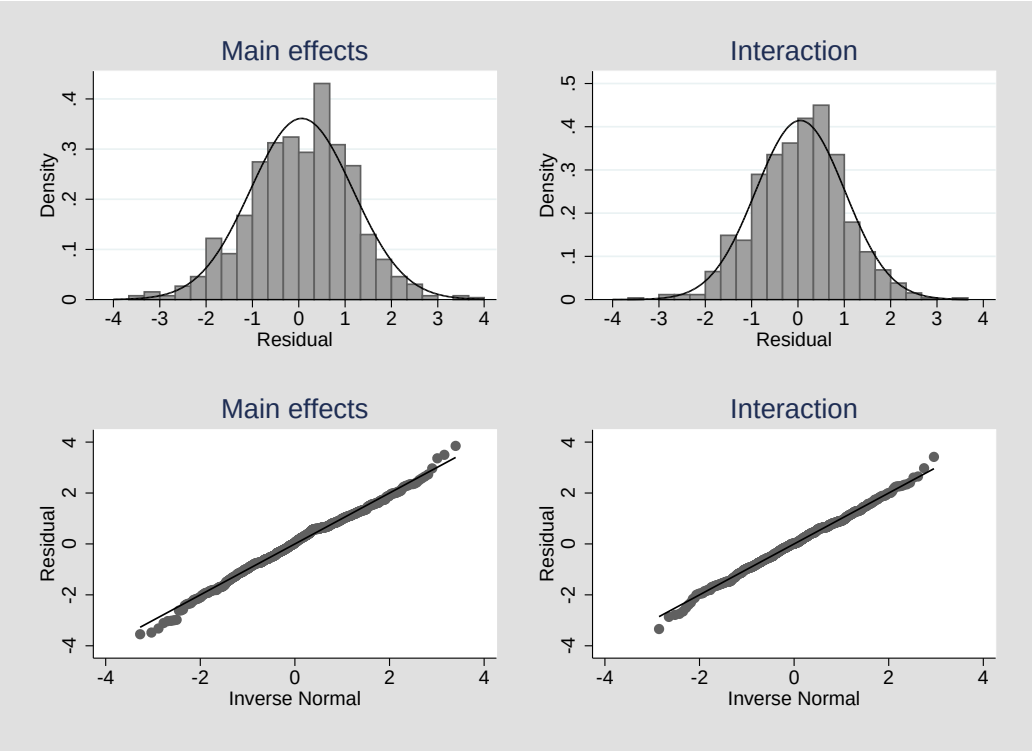


Figure 4.4: Distributions of deviance residuals from main effects and interaction models fitted to HCV prevalence data. Histograms with normal density overlaid and quantile-quantile plots.

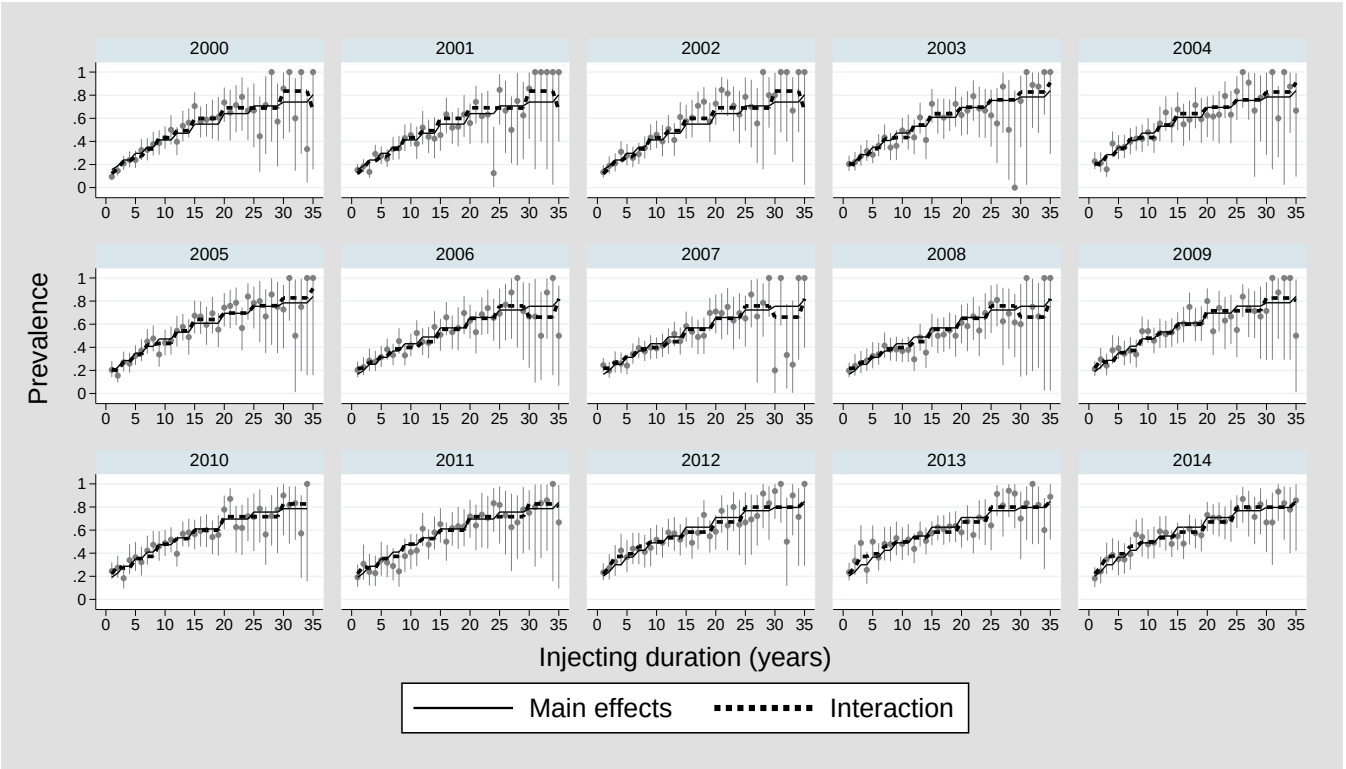


Figure 4.5: Observed and predicted HCV prevalence from 2000 onwards by injecting duration and survey year; main effects of injecting duration and calendar time, and interaction model. 95% binomial confidence intervals are displayed around the observed prevalence.

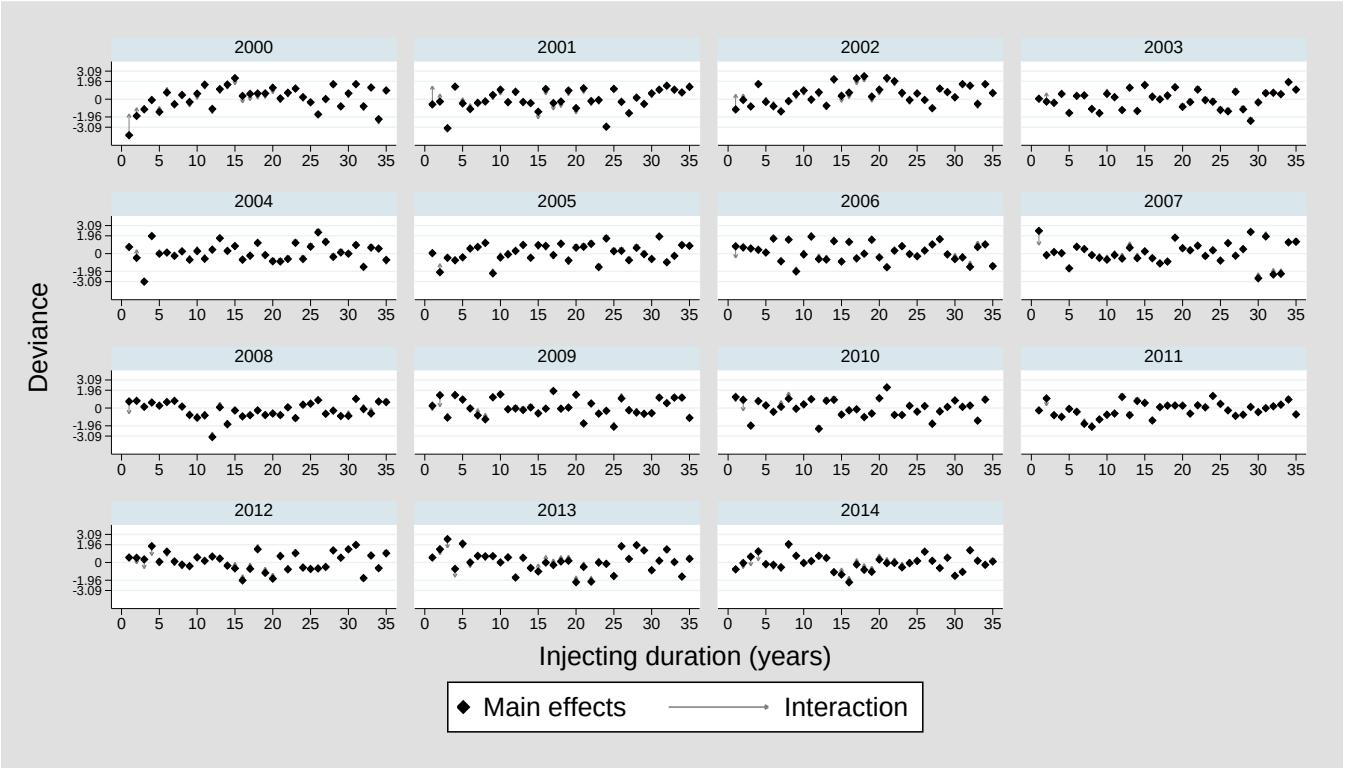


Figure 4.6: Deviance residuals by injecting duration and survey year from main effects and interaction models fitted to HCV prevalence data from 2000 onwards. Reference lines are the 2.5/97.5th and 0.1/99.9th percentiles of the standard normal distribution.

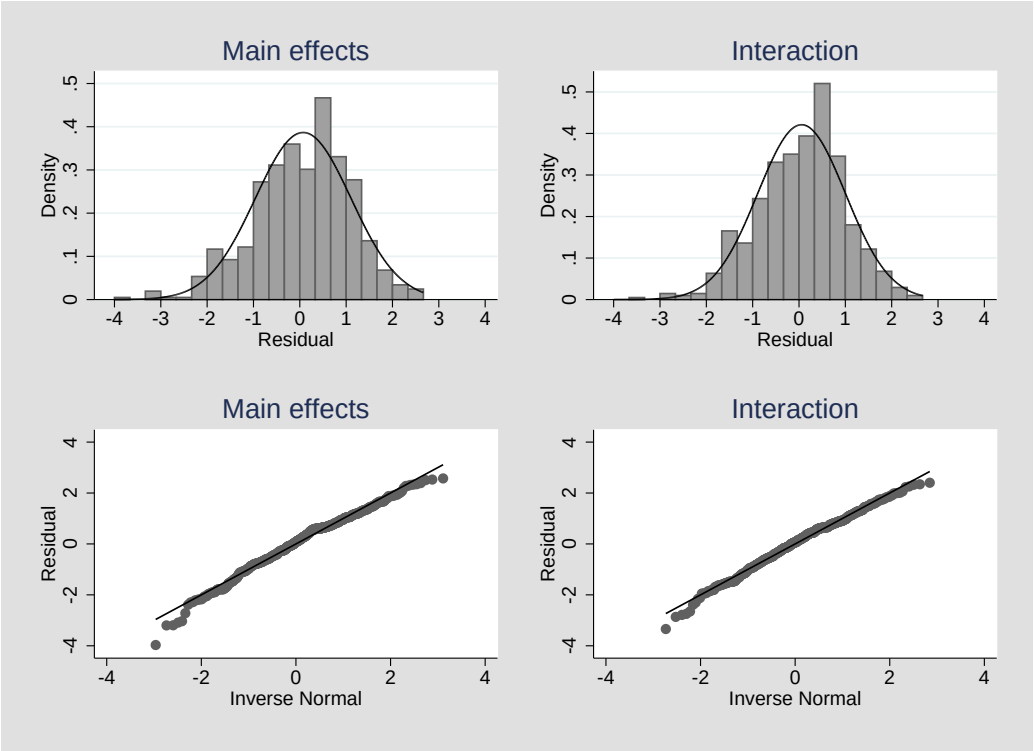


Figure 4.7: Distributions of deviance residuals from main effects and interaction models fitted to HCV prevalence data from 2000 onwards. Histograms with normal density overlaid and quantile-quantile plots.

### 4.1.2 Age and time patterns for HBV infection

The age and time models from section 4.1.1 were fitted to the data on HBV infection. The constant term (-2.44) indicated HBV prevalence of 8% in those injecting for 5-6 years in survey years 2006-2008. The ORs for injecting duration from the main effects model follow a broadly similar pattern to HCV, but are more extreme for longer injecting durations, although the constant (logit prevalence at baseline) is lower (Table 4.2. The temporal effect is very different though, with a much greater risk in 1990-1993 compared to 2006-2008 (OR=3.94) and subsequently decreasing; instead of the ORs increasing again for more recent survey years, the ORs continue to decline (0.63 in 2012-2014). Figure 4.8 shows observed and predicted prevalence from the main effects and age/time interaction models.

The deviance of the main effects model was 1295 on 919 degrees of freedom, which is somewhat inadequate. The interaction model provided a sig-

nificantly better overall fit than the main effects model (likelihood ratio test $\chi^2 = 182.4$ on 77 degrees of freedom, $p < 0.0001$), but in terms of absolute fit was also inadequate, with a deviance of 1113 on 842 degrees of freedom. Note that the number of observations is larger for the HBV data, which has been collected every year since the inception of the survey. Similarly to the HCV data, much of the lack of fit occurs in data from earlier survey periods, with both models fitting particularly badly for 1990 and 1991, with a downward trend in residuals according to injecting duration for 1990 and an upward trend for 1991 (Figure 4.9). Runs tests of deviance residuals across injecting duration indicated non-randomness in the years 1990 ($p < 0.001$), 1999 ($p = 0.015$) and 1999 ($p = 0.025$) for the main effects model. From 2000 onwards there was less evidence of non-randomness, with an overall $p$-value of 0.076 for the main effects model. Figure 4.10 shows histograms and quantile-quantile (QQ) plots of the residuals from main effect and interaction models. Both have an approximately normal distribution, although there is an apparent spike in the distribution of residuals for the main effects model that is not present in the interaction model. Further investigation of the distribution showed no specific pattern in the binomial data (for instance, small counts of 1 or 2) or covariates associated with the spike: the distribution is simply somewhat lumpy.

Marked differences between one year and the next cannot be captured by broad categories for temporal effects whether an interaction is included or not. The only alternative would be to include individual year effects, although it seems unlikely that underlying prevalence would change so markedly in successive years; more likely, there are substantial differences in the sampling frame and survey methodology in the earlier years of the survey.

Restricting to 2000 onwards as before results in a deviance of 785 on 600 degrees of freedom for the main effects model and 710 on 556 degrees of freedom for the interaction model; both models providing a somewhat inadequate fit to the data. Examination of the residuals reveals some patterns in specific years of the survey and some extreme outliers, rather than systematic deviations that persist across different years. These outliers occur both in the main effects and interaction models, and therefore correspond to variability in prevalence within a particular time or injecting duration band.

As the categories are already relatively small, it seems reasonable to consider this excess variability as overdispersion rather than a failure to adequately describe the structural part of the model. Such factors could of course be due to changes in risk factors in those recruited in particular years, and therefore could be captured if other risk information from the questionnaire data is included in the model.

Table 4.2: Model coefficients for injecting duration and survey period, main effects model fitted to HBV data, exponentiated to give odds ratios (ORs) (except constant)

| | OR | 95% CI | $Z$-val | $p$-val |
|---|---|---|---|---|
| Injecting duration | | | | |
| 1 | 0.42 | 0.37, 0.49 | -11.71 | <0.001 |
| 2 | 0.63 | 0.55, 0.74 | -6.01 | <0.001 |
| 3-4 | 0.83 | 0.73, 0.94 | -2.99 | 0.003 |
| 5-6 | 1 (ref) | | | |
| 7-8 | 1.33 | 1.18, 1.50 | 4.65 | <0.001 |
| 9-11 | 1.85 | 1.66, 2.06 | 11.18 | <0.001 |
| 12-14 | 2.64 | 2.37, 2.95 | 17.19 | <0.001 |
| 15-19 | 3.99 | 3.59, 4.44 | 25.58 | <0.001 |
| 20-24 | 5.62 | 5.00, 6.31 | 29.15 | <0.001 |
| 25-29 | 8.36 | 7.27, 9.62 | 29.74 | <0.001 |
| 30-34 | 13.31 | 11.03, 16.07 | 26.97 | <0.001 |
| 35+ | 19.62 | 14.78, 26.06 | 20.56 | <0.001 |
| Survey period | | | | |
| 1990-1993 | 3.94 | 3.57, 4.35 | 27.14 | <0.001 |
| 1994-1996 | 2.06 | 1.86, 2.27 | 14.27 | <0.001 |
| 1997-1999 | 1.56 | 1.41, 1.73 | 8.53 | <0.001 |
| 2000-2002 | 1.66 | 1.50, 1.84 | 9.79 | <0.001 |
| 2003-2005 | 1.43 | 1.28, 1.59 | 6.55 | <0.001 |
| 2006-2008 | 1 (ref) | | | |
| 2009-2011 | 0.69 | 0.61, 0.77 | -6.21 | <0.001 |
| 2012-2014 | 0.63 | 0.56, 0.70 | -7.91 | <0.001 |
| Constant | -2.44 | -2.55, -2.32 | -42.26 | <0.001 |

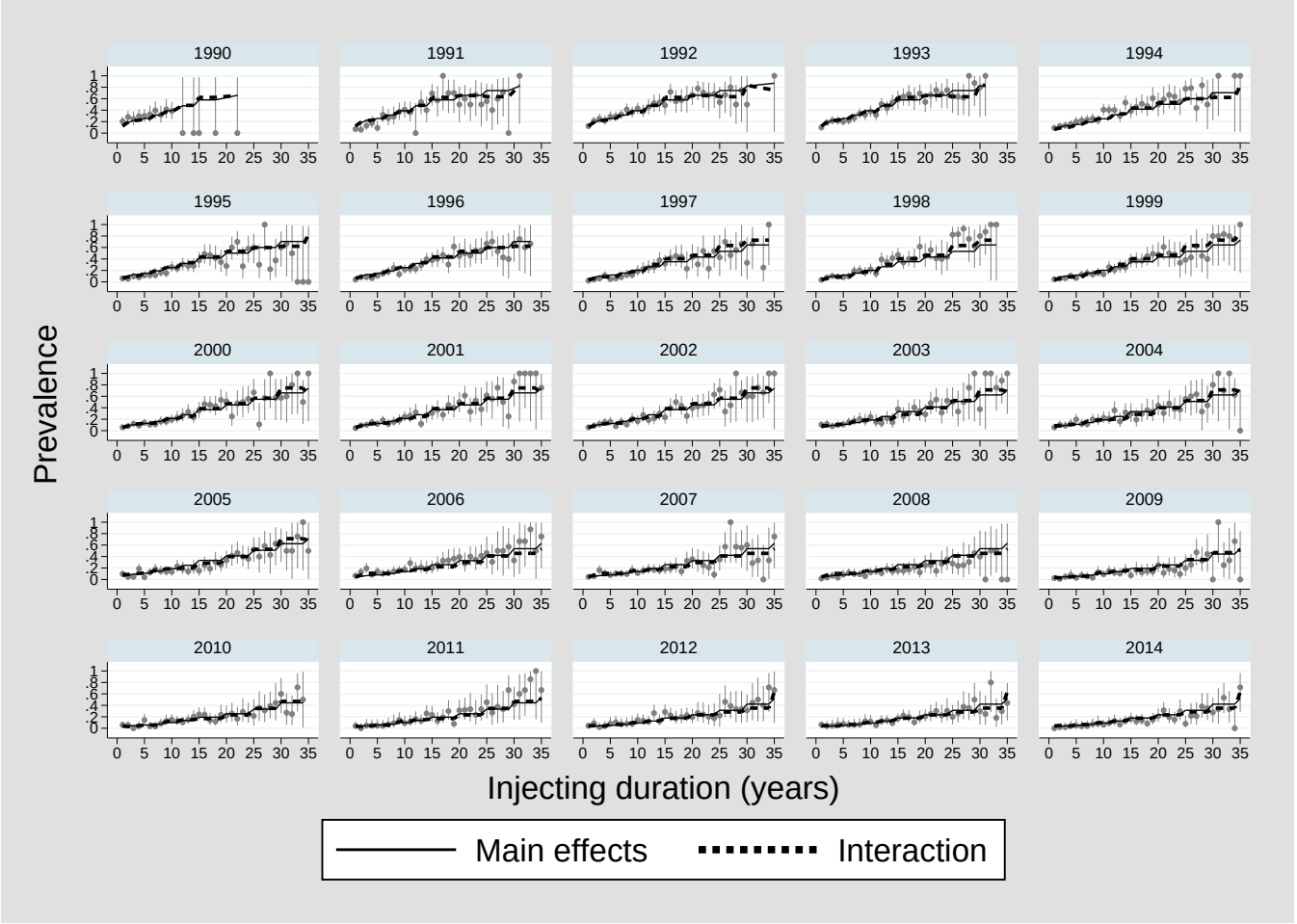


Figure 4.8: Observed and predicted HBV prevalence by injecting duration and survey year; main effects of injecting duration and calendar time, and interaction model. 95% binomial confidence intervals are displayed around the observed prevalence.

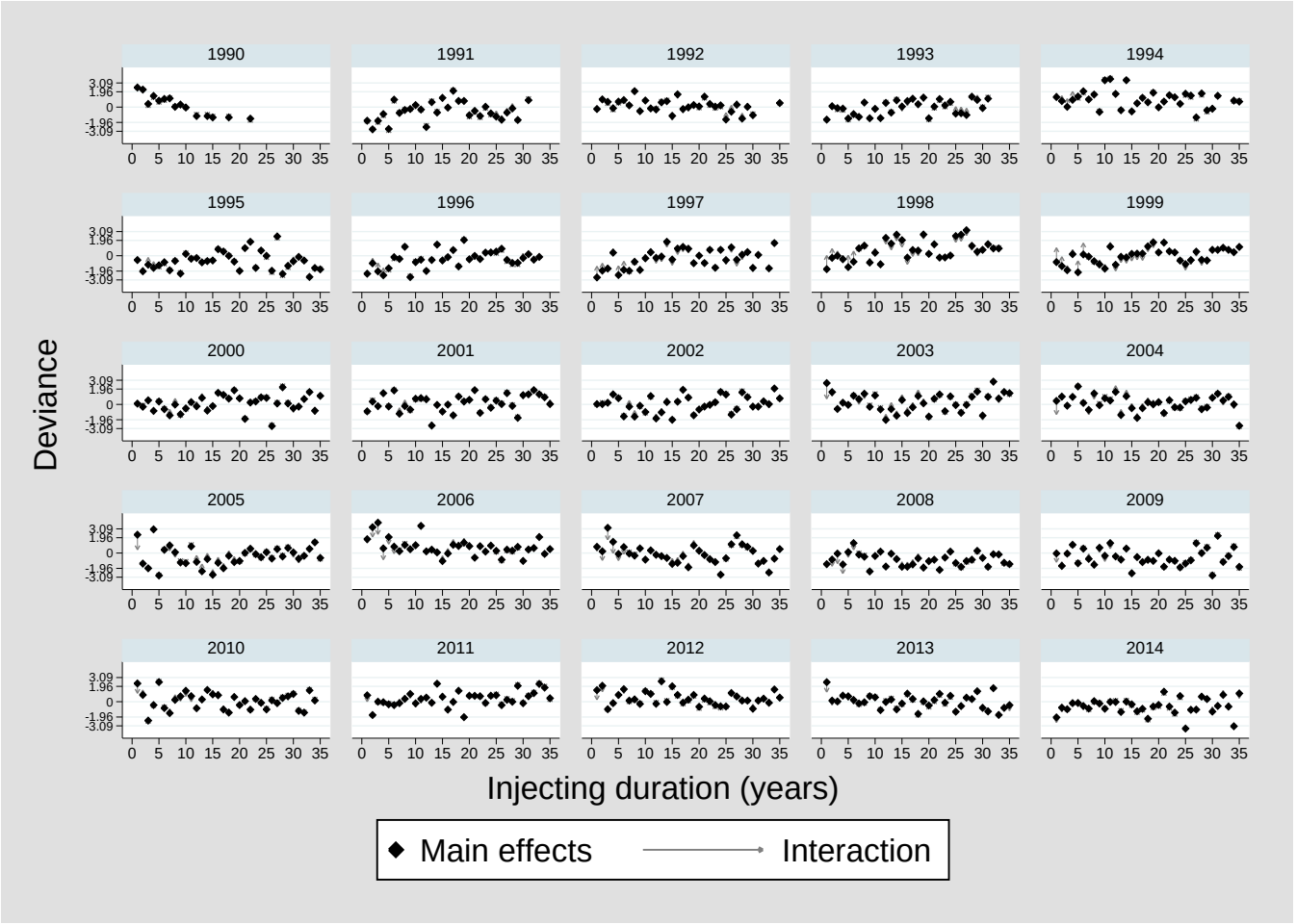


Figure 4.9: Deviance residuals by injecting duration and survey year from main effects and interaction models fitted to HBV prevalence data. Reference lines are the 2.5/97.5th and 0.1/99.9th percentiles of the standard normal distribution.

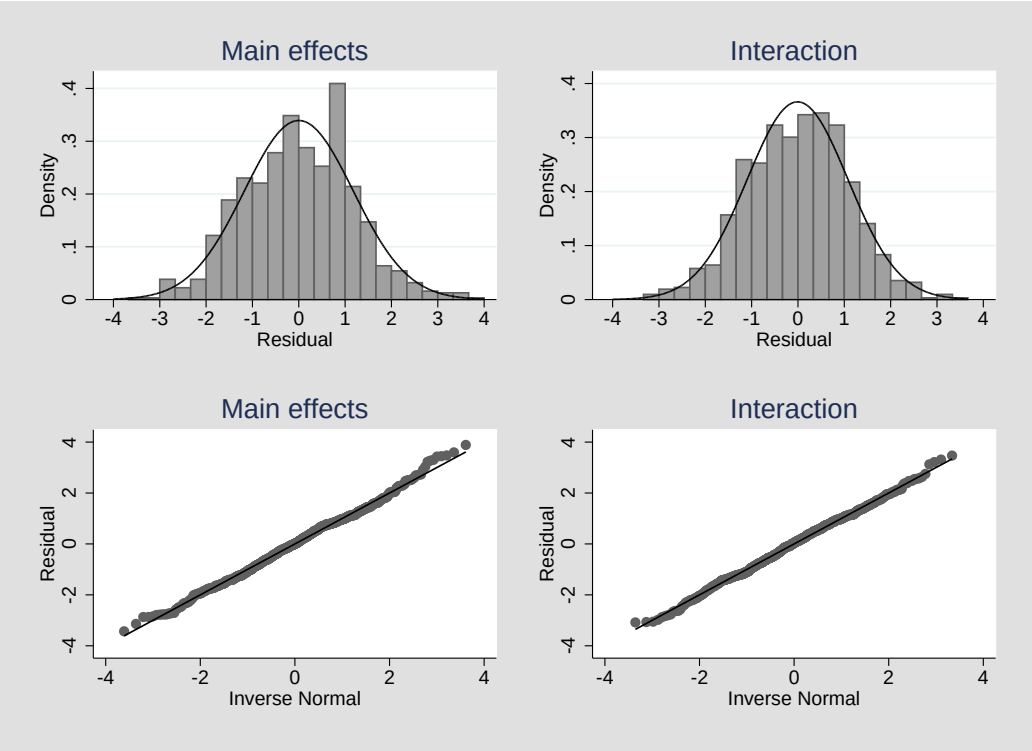


Figure 4.10: Distributions of deviance residuals from main effects and interaction models fitted to HBV prevalence data. Histograms with normal density overlaid and quantile-quantile plots.

### 4.1.3 Age and time patterns for HIV infection

Finally, GLMs for age and time were fitted to the HIV data. Prevalence is far lower than the other two infections, so results were far less precise; further, the categorisation used previously leads to over-fitting (data not shown). Broader categories were therefore used, with survey period grouped into 5-year intervals (1990-1994 etc.) and injecting duration as 1, 2-4, 5-9, 10-14, 15-19, 20-34 and 35+, with 2005-2009 and 5-9 as the baseline groups. Table 4.3 shows that the prevalence in the baseline category of 5-9 years injecting in survey years 2005-2009 was around 0.7%. The odds of infection with HIV at 0-1 and 2-4 years injecting were not significantly less than 5-9 years; and the increase in odds with injecting duration was not quite monotonic, even with the broader categories.

As with the other infections the OR for the earliest period, 1990-1994, was highest at 1.21 (although the $p$-value was 0.139) followed by a decline in

Table 4.3: Model coefficients for injecting duration and survey period, main effects model fitted to HIV data, exponentiated to give odds ratios (ORs) (except constant). Categories are broader than HCV/HBV due to sparse data.

| | OR | 95% CI | $Z$-val | $p$-val |
|---|---|---|---|---|
| Injecting duration | | | | |
| 1 | 0.75 | 0.47, 1.21 | -1.18 | 0.240 |
| 2-4 | 0.72 | 0.48, 1.08 | -1.58 | 0.113 |
| 5-9 | 1 (ref) | | | |
| 10-14 | 2.62 | 1.95, 3.53 | 6.36 | <0.001 |
| 15-19 | 4.20 | 3.10, 5.69 | 9.30 | <0.001 |
| 20-34 | 3.95 | 2.91, 5.37 | 8.77 | <0.001 |
| 35+ | 4.61 | 1.97, 10.78 | 3.52 | <0.001 |
| Survey period | | | | |
| 1990-1994 | 1.21 | 0.94, 1.57 | 1.48 | 0.139 |
| 1995-1999 | 0.62 | 0.47, 0.82 | -3.40 | 0.001 |
| 2000-2004 | 0.72 | 0.55, 0.95 | -2.32 | 0.020 |
| 2005-2009 | 1 (ref) | | | |
| 2010-2014 | 0.61 | 0.46, 0.81 | -3.44 | 0.001 |
| Constant | -4.92 | -5.20, -4.63 | -34.06 | <0.001 |

prevalence before increasing again in 2005-2009; although the odds of HIV were then lower for 2010-2014 vs. 2005-2009 (OR=0.61). The deviance of the main effects model was 875 on 928 degrees of freedom, indicating that the model may be over-fitting somewhat. Examination of the deviance residuals (Figure 4.12) indicates that both the main effects and the interaction models tend to have small, negative deviance residuals for most data points, with a smaller number of extreme positive outliers. This is due to the sparsity of the data; most HIV counts are zero and the model produced overestimates of prevalence, but for non-zero counts the model produced marked underestimates. Neither a main effects or interaction model can capture this (short of specifying a parameter for every survey year/injecting duration combination in the saturated model); therefore although the interaction model apparently produces a better fit to the data (likelihood ratio test $\chi^2 = 93.5$ on 24 degrees of freedom, $p < 0.0001$) the extra parameters do not materially improve the model.

Figures 4.11, 4.12 and 4.13 show observed and predicted prevalence by injecting duration and time, deviance residuals and their distributions respectively. Runs tests of deviance residuals across injecting duration strongly indicated non-randomness, with $p$-values of less than 0.001 for both main effects and interaction models. Clearly assumptions of asymptotic normality do not apply due to the sparsity of the data, and there is little that can be done about this, short of specifying a far simpler model; the information provided in terms of differences in HIV prevalence according to injecting duration and time is far lower than for HCV or HBV.

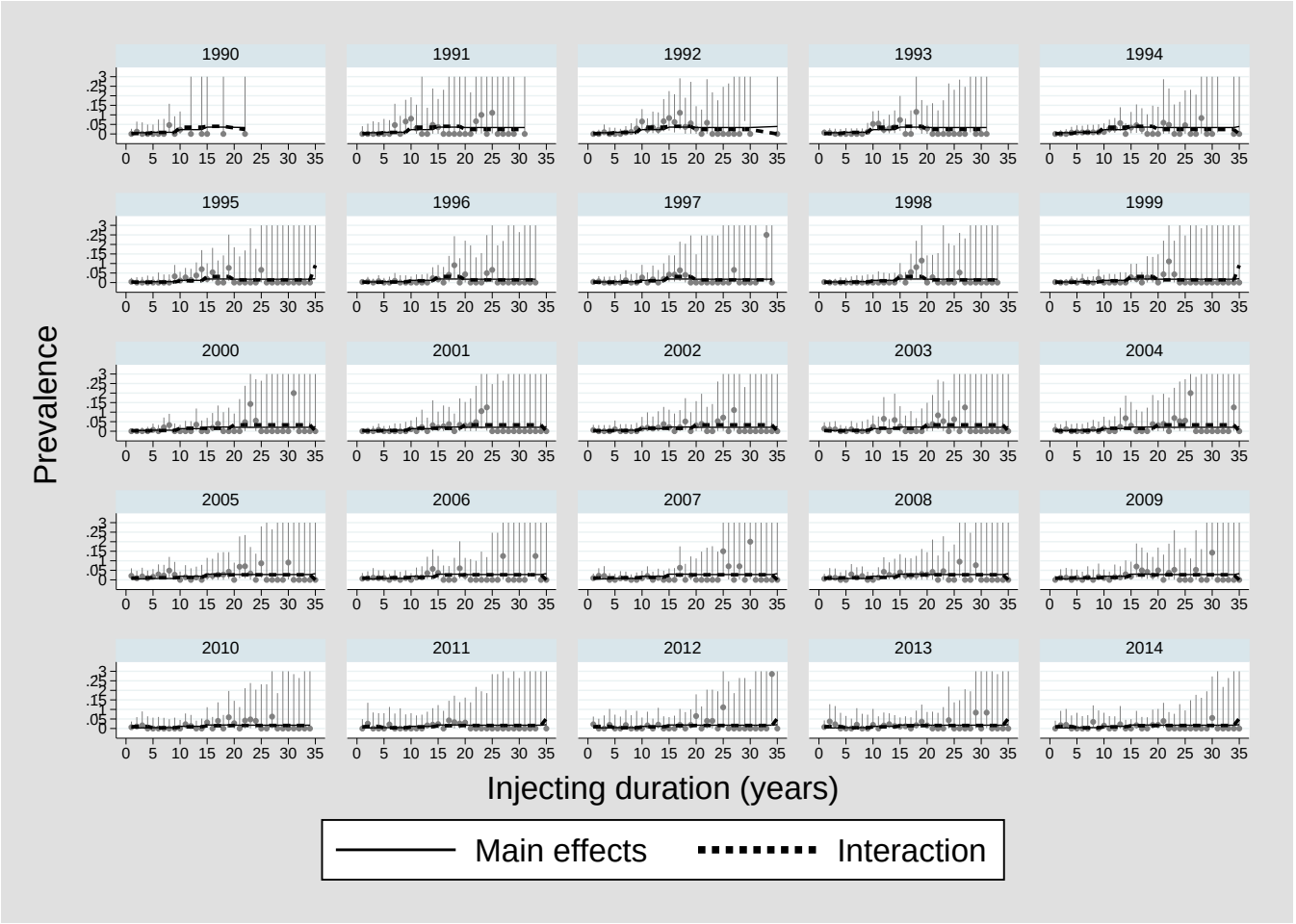


Figure 4.11: Observed and predicted HIV prevalence by injecting duration and survey year; main effects of injecting duration and calendar time, and interaction model. 95% binomial confidence intervals are displayed around the observed prevalence.

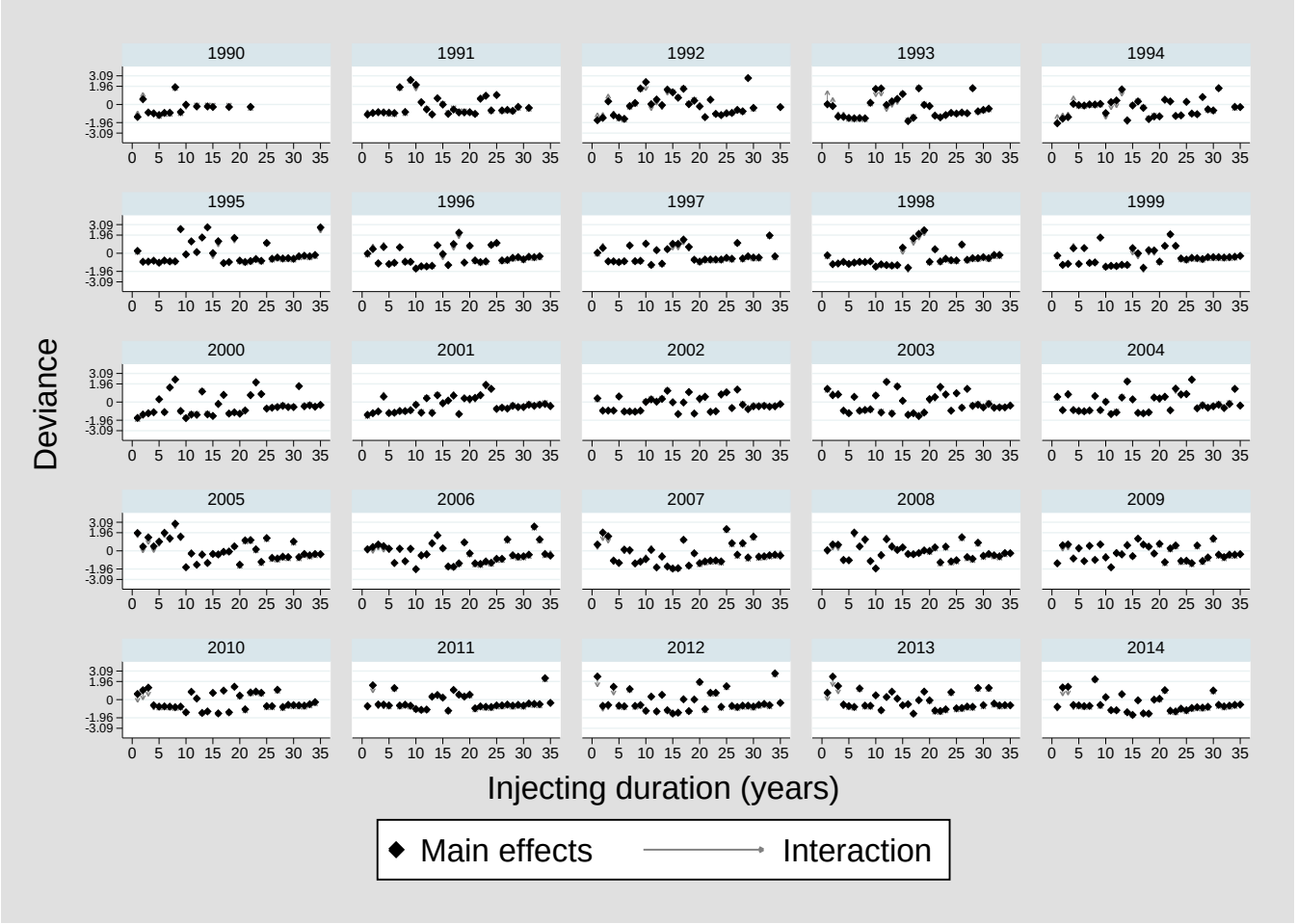


Figure 4.12: Deviance residuals by injecting duration and survey year from main effects and interaction models fitted to HIV prevalence data. Reference lines are the 2.5/97.5th and 0.1/99.9th percentiles of the standard normal distribution.

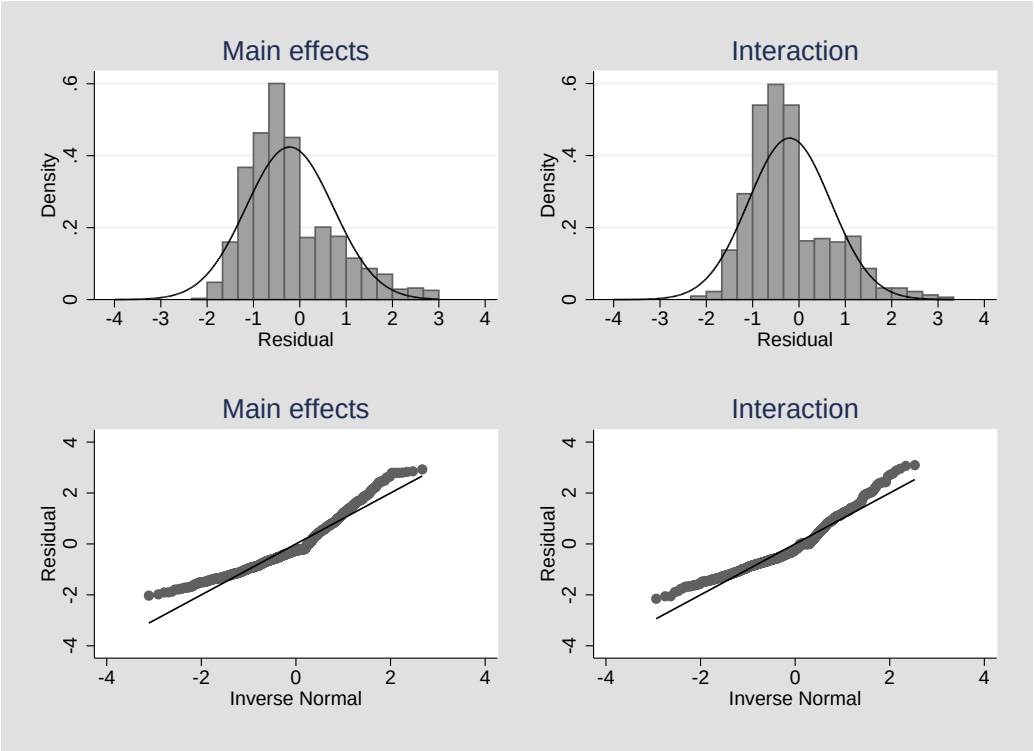


Figure 4.13: Distributions of deviance residuals from main effects and interaction models fitted to HIV prevalence data. Histograms with normal density overlaid and quantile-quantile plots.

### 4.1.4 Generalised linear models with covariates

In this section the models for injecting duration and survey year in section 4.1.1 are extended to include additional covariates, based on the information provided in the UAM questionnaire. Including data on participant characteristics and injecting behaviour can provide insights into risk factors for infection and may also help to account for overdispersion in prevalence according to injecting duration and time, which was observed in sections 4.1.1 to 4.1.3. If survey participants vary in characteristics over time, which is quite possible in the UAM data as the service providers sampled vary from year to year, then accounting for these changes will reduce variation in the temporal component of the model. This will then provide a better indication of underlying trends in the overall risk of infection.

Available covariates include age, gender and government office region, the latter consisting of East of England (EE), London, South East (SE), South

West (SW), West Midlands (W Mids), North West (NW), Yorkshire and the Humber, East Midlands (E Mids), North East (NE) and Wales. SE was taken as the baseline category for region; London and NW had the largest sample sizes but also had higher prevalence of HCV and HIV, with lower prevalence in more rural regions, and SE somewhere in between. Including age as a covariate requires some caution as it is highly correlated with injecting duration, with age at first use generally being in early adulthood. Age was therefore parameterised in terms of age first injected, which results in a model mathematically equivalent to including current age (if injecting duration and time effects are included), but provides parameter estimates that are more easily interpreted. In order to continue model assessment in terms of categorical variables, age at first use (AAFU) was categorised as $< 18$, 18-24 and 25+. 93% of participants reported AAFU as 30 or below, so there is little scope for further investigation of differences at older ages.

In addition to the demographic variables above, information on risk behaviour was considered. These variables include ever received works, i.e., used drug equipment and paraphernalia, principally needles or syringes, from another drug user; ever used a needle exchange and age first used; ever been in prison and age first imprisoned; number of days injected in the last 28 days; and sexual activity/risk in the last year including number of male and female partners and condom use (always, sometimes and never).

The variables used in the analysis were coded as follows. Ever received works is binary. Use of needle exchange is categorised as used from the first year of injecting, used at some point following first year of injecting, and never. Imprisonment is categorised as never imprisoned, first been to prison before started injecting, and first been to prison after started injecting. Days injecting is categorised as injecting for 14 or more days of the last 28, or fewer. Variables for sexual risk are defined as two or more partners vs. none or one, incomplete condom use (sometimes/never vs. always), and men who have sex with men (MSM), defined as one or more male partners for males.

Many of these variables are only available from the year 2000 onwards, as the UAM questionnaire has evolved and expanded over time, and therefore analyses are restricted to these years, leaving 25280 observations available for analysis. There are also missing data for a number of responses. Table 4.4 shows the available data for each of the covariates, and overall. The

amount of missing data is less than 5% for most of the variables, but over 20% are missing for number of partners and condom use, meaning that over a quarter of the observations would be discarded in a complete case analysis. Age, sex and region are complete for all observations.

Table 4.4: Available data for key covariates in the UAM data from 2000 onwards, percentage of complete data from 25280 observations.

| Variable | % complete |
|---|---|
| Ever received works | 98.8% |
| Needle exchange | 99.3% |
| Prison | 97.5% |
| Number of partners | 79.9% |
| Condom use | 78.0% |
| MSM | 96.9% |
| Complete cases | 74.4% |

Generalised linear models (GLM) with a logit link were then fitted to the data from 2000 onwards with complete data for all of the variables above. Of the three infections, 4 observations were missing HBV test results, so these were excluded also to give 17116 observations with no missing data. Variables were considered without interactions in a 3-stage approach: the unadjusted effect of each variable on prevalence; an intermediate stage where injecting duration and survey period are adjusted for; and a full multivariable model including all covariates. The intermediate stage was performed as a check to see if each covariate in isolation was confounded with injecting duration and/or survey year, although these results are only discussed where of special interest and the main focus is on the univariable and full multivariable results in the following.

### 4.1.5 Risk factors for HCV infection

Results for HCV from are shown in tables 4.5 and 4.6. Odds ratios (OR) for injecting duration and survey year do not change much between univariable and multivariable (MV) analyses, although there is slight attenuation in the injecting duration-specific ORs. There was a strong effect of region, with East of England, SW, W Mids, E Mids, NE and Wales having lower odds

of HCV compared to SE England; but higher risks in London (OR=1.34, MV) and NW (OR=1.77, MV). Results were generally similar for univariable and multivariable analyses. Age at first use shows an increased risk for <18 (OR=1.35) and a slight decrease for 25+ (OR=0.95) compared to 18-24 in univariable analyses; however, after adjusting for other covariates the relationship is almost exactly reversed, with higher risk for 25+ and a slight decrease for <18. The difference is largely due to adjusting for injecting duration, as in the univariable analysis those that began injecting at a later age will, on average, have shorter injecting durations. Females had no difference in risk for univariable analysis but higher risk in multivariable analyses; this is again due to adjusting for injecting duration, as females tend to have shorter injecting duration (median 6 vs. 9 for males) but comparable prevalence levels.

Ever receiving works from another PWID was associated with a higher risk of HCV infection, with an OR of 1.73, which persisted after adjusting for other variables. Needle exchange use showed some interesting univariable results, with an increased risk of HCV (OR=1.65) for those beginning to use a needle exchange after 1 year vs. starting in their first year of injecting, but a decreased risk in those that had never used a needle exchange. In the multivariable model, there was no difference in using a needle exchange before or after one year of injecting. Some care is required in the interpretation of these results however, as 1st year injectors must by definition either start using needle exchange in their first year or be classed in the survey as "never". The OR for never using a needle exchange is rather counterintuitive; a possible explanation is that this group are less frequent injectors or otherwise have lower levels of opiate dependence. However, the result did not change in the multivariable analysis, which includes information of frequency of injecting and other measures of the potential "riskiness" of the individual. In fact, the number of days injecting in the last month showed little association in multivariable analyses.

Imprisonment was a significant risk factor, with ORs of 1.82 and 2.13 for having been to prison before or after started injecting in the multivariable analysis. The similarity of the ORs indicates that there may be little difference in the timing of first prison sentence. Sexual behaviour variables were difficult to interpret: there was no significant difference in the number

of partners but, strangely, a protective effect for incomplete condom use. Although sexual transmission is thought to be rare (Balogun et al., 2003), more risky sexual behaviour might be thought to be correlated with more risky injecting behaviour, rather than the reverse. Being a MSM was associated with an increase in risk; although sexual transmission is uncommon in general, the MSM population does have a higher risk of HCV infection due to certain high-risk sexual practices and HIV co-infection (van de Laar et al., 2007). Model fit for the multivariable model cannot be assessed in the same way as section 4.1.1 as the deviance relies on a sufficient number of binomial observations for each data point. The aggregated data for the analysis above consist of 14623 combinations of the explanatory variables with 87% consisting of a single observation, and are therefore practically binary. The Hosmer-Lemeshow statistic based on deciles of the linear predictor gives a $p$-value of 0.308, indicating no evidence of a lack of fit; however, the test is quite sensitive to the chosen percentiles; with 15 equal-size groups the $p$-value is 0.033.

Having added covariates to the main effects model for injecting duration and survey period in section 4.1.1, the presence of interactions was assessed, starting with the interaction between injecting duration and survey period to determine whether the interaction observed in section 4.1.1 was still necessary after accounting for other risk factors. The likelihood ratio (LR) test gave a $p$-value of 0.055, although model selection scores indicated that the model may be over-parameterised, with an AIC score of 19574.7 for the main effects model vs. 19602.7 with the addition of interactions.

Interactions between the other covariates and injecting duration and survey year were also assessed. Such interactions could arise from a variety of causes, such as changes in prevalence over time in a particular region, differences in risk pattern according to injecting duration in males and females due to different risk behaviours, and so on. For injecting duration, there were significant interactions with age at first use and prison, and possibly ever receiving works. For survey year, there were significant interactions with region, gender, ever receiving works and use of needle exchange. To summarise the key findings, first-year injectors that began injecting before the age of 18 had an increased risk of infection (OR=2.04, 95%CI 1.19-3.50), as well as first-year injectors that had been to prison before starting injecting

Table 4.5: Univariable and multivariable results from logistic regression model for HCV and reported risk factors. Odds ratios and 95% confidence intervals

| Variable | | Univariable | Multivariable |
|---|---|---|---|
| Injecting duration | 0-1 | 0.48 (0.41, 0.57) | 0.50 (0.42, 0.59) |
| | 2 | 0.57 (0.48, 0.67) | 0.60 (0.50, 0.71) |
| | 3-4 | 0.82 (0.71, 0.94) | 0.82 (0.71, 0.95) |
| | 5-6 | 1 (ref) | 1 (ref) |
| | 7-8 | 1.34 (1.16, 1.54) | 1.29 (1.11, 1.49) |
| | 9-11 | 1.75 (1.54, 1.99) | 1.60 (1.40, 1.83) |
| | 12-14 | 2.21 (1.93, 2.52) | 1.93 (1.68, 2.23) |
| | 15-19 | 3.18 (2.79, 3.61) | 2.67 (2.31, 3.07) |
| | 20-24 | 4.94 (4.24, 5.75) | 4.16 (3.52, 4.91) |
| | 25-29 | 6.46 (5.22, 8.00) | 5.47 (4.33, 6.89) |
| | 30-34 | 8.87 (6.46, 12.18) | 7.62 (5.45, 10.65) |
| | 35+ | 9.29 (5.57, 15.49) | 8.02 (4.68, 13.74) |
| Survey period | 2000-2002 | 0.81 (0.74, 0.89) | 0.93 (0.84, 1.03) |
| | 2003-2005 | 1.15 (1.04, 1.26) | 1.15 (1.03, 1.28) |
| | 2006-2008 | 1 (ref) | 1 (ref) |
| | 2009-2011 | 1.29 (1.17, 1.42) | 1.23 (1.10, 1.37) |
| | 2012-2014 | 1.55 (1.40, 1.71) | 1.53 (1.37, 1.72) |
| Region | East of England | 0.54 (0.46, 0.64) | 0.56 (0.47, 0.67) |
| | London | 1.47 (1.30, 1.65) | 1.34 (1.17, 1.53) |
| | South East | 1 (ref) | 1 (ref) |
| | South West | 0.51 (0.45, 0.57) | 0.54 (0.47, 0.62) |
| | West Midlands | 0.45 (0.38, 0.52) | 0.49 (0.41, 0.58) |
| | North West | 1.71 (1.52, 1.93) | 1.77 (1.55, 2.02) |
| | Yorkshire and H | 1.40 (1.18, 1.66) | 1.09 (0.91, 1.32) |
| | East Midlands | 0.76 (0.67, 0.87) | 0.83 (0.72, 0.96) |
| | North East | 0.34 (0.30, 0.38) | 0.48 (0.42, 0.56) |
| | Wales | 0.44 (0.37, 0.52) | 0.43 (0.36, 0.51) |
| Age at first use | <18 | 1.35 (1.25, 1.45) | 0.93 (0.85, 1.02) |
| | 18-24 | 1 (ref) | 1 (ref) |
| | 25+ | 0.95 (0.88, 1.03) | 1.29 (1.18, 1.41) |
| Gender | Male | 1 (ref) | 1 (ref) |
| | Female | 1.00 (0.93, 1.07) | 1.50 (1.38, 1.64) |

Table 4.6: *Continued from 4.5*: Logistic regression model results for HCV.

| Variable | | Univariable | Multivariable |
|---|---|---|---|
| Ever rec'd works | No | 1 (ref) | 1 (ref) |
| | Yes | 1.73 (1.63, 1.84) | 1.67 (1.56, 1.80) |
| Needle exchange | Started 1st year | 1 (ref) | 1 (ref) |
| | Started >1 yr | 1.65 (1.55, 1.76) | 1.00 (0.93, 1.08) |
| | Never | 0.61 (0.49, 0.77) | 0.71 (0.56, 0.92) |
| Days injecting per month | <14 days/mo | 1 (ref) | 1 (ref) |
| | 14+ days/mo | 0.88 (0.83, 0.94) | 1.03 (0.96, 1.11) |
| Prison | Never imprisoned | 1 (ref) | 1 (ref) |
| | Prison before injecting | 2.61 (2.40, 2.83) | 1.82 (1.65, 2.01) |
| | Prison after injecting | 2.18 (2.01, 2.35) | 2.13 (1.95, 2.34) |
| Number of partners | 0 or 1 partner | 1 (ref) | 1 (ref) |
| | 2+ partners | 0.86 (0.80, 0.91) | 0.96 (0.90, 1.04) |
| Condom use | Always | 1 (ref) | 1 (ref) |
| | Sometimes/never | 0.78 (0.73, 0.84) | 0.79 (0.73, 0.86) |
| MSM | No | 1 (ref) | 1 (ref) |
| | Yes | 1.26 (1.03, 1.54) | 1.25 (0.99, 1.57) |

(OR=3.14, 95% CI 1.82-5.42). Estimates for interactions with other injecting durations were generally non-significant and had no particular pattern according to injecting duration.

The interaction between survey period and region is shown in Figure 4.14 by way of predicted probabilities of HCV infection. The general trend in SE England is of increasing prevalence over time. Significant interactions therefore occur for East England, which has less of a trend, and Wales, which has a higher prevalence in 2000-2002 and increases sharply in 2012-2014 (OR=2.90, 95% CI 1.70-4.93). A number of other regions show significant differences in individual years, with relatively higher prevalence in W Mids in 2000-2002, Yorkshire and the Humber in 2003-2005, and NE in 2000-2002; and lower prevalence in the NW in 2009-2011.

The interaction between gender and survey period showed a further increase in risk of infection in females in 2012-2014 (OR=1.45, 95% CI 1.13-1.87), but no difference in other periods. Ever receiving works was associated with increased risk in 2003-2005 (OR=1.26, 95% CI 1.02-1.56) and 2012-2014

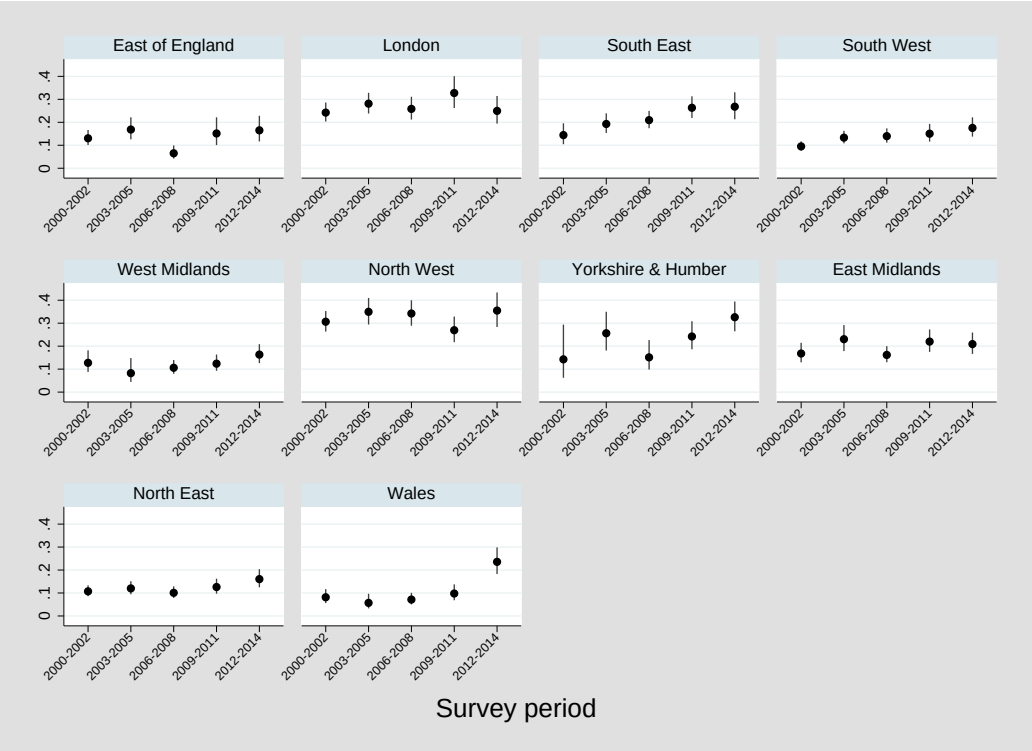


Figure 4.14: Predicted HCV prevalence by region and survey period with 95% confidence intervals, multivariable logistic model adjusting for main effects of all risk factors and region/survey period interaction. Predictions are made at baseline levels of other risk factors; e.g., injecting duration 5-6 years.

(OR=1.36, 95% CI 1.08-1.71), with no difference in other periods. Finally, never using a needle exchange was associated with increased risk in 2000-2002 (OR=2.03, 95% CI 0.97-4.23), with no difference in other periods.

At the risk of over-complicating the model and its interpretation, there were also significant interactions between a number of risk factors in addition to the interactions described above. A base model was specified with main effects and the survey period/region interaction, which was by far the strongest, and the addition of interactions for other variables compared to this model via likelihood ratio tests. Interactions were identified between region and age at first use ($p$=0.002), gender ($p$=0.001), needle exchange ($p$=0.004) and prison ($p$=0.006); and possibly days injecting ($p$=0.066). All $p$-values are for LR tests. $p$-values for all possible two-way interactions are shown in Table 4.7.

Table 4.7: Interactions between risk factors for HCV, likelihood ratio test $p$-values for generalised linear model with main effects, region/survey year interaction and interaction of interest. MSM/female interaction is not defined.

| | region | aafu | female | erec | exch | d14 | pris | part2 | cond |
|---|---|---|---|---|---|---|---|---|---|
| aafu | 0.002 | | | | | | | | |
| female | 0.001 | 0.597 | | | | | | | |
| erec | 0.121 | 0.942 | 0.251 | | | | | | |
| exch | 0.004 | 0.889 | 0.249 | 0.045 | | | | | |
| d14 | 0.066 | 0.993 | 0.965 | 0.359 | 0.400 | | | | |
| pris | 0.006 | 0.838 | 0.064 | 0.215 | 0.101 | 0.825 | | | |
| part2 | 0.343 | 0.457 | 0.866 | 0.059 | 0.054 | 0.006 | 0.721 | | |
| cond | 0.524 | 0.005 | 0.936 | 0.100 | 0.636 | 0.582 | 0.536 | 0.914 | |
| msm | 0.937 | 0.906 | . | 0.977 | 0.506 | 0.142 | 0.222 | 0.454 | 0.762 |

**aafu**: age at first use, **erec**: ever received works, **exch**: needle exchange use, **d14**: injected 14 or more days in last month, **pris**: imprisonment, **part2**: two or more sexual partners, **cond**: condom use, **msm**: men who have sex with men.

### 4.1.6 Risk factors for HBV infection

The models described in section 4.1.4 were fitted to the HBV data and are reported more succinctly in the following. Covariate effects for the main effects models for injecting duration, survey period and reported risk factors are displayed in tables 4.8 and 4.9. Region showed a similar pattern to HCV with somewhat higher prevalence in London (OR=1.24, MV) and higher in the NW (OR=1.98, MV). In general, the regional differences where less extreme. Age at first use showed the same pattern as HCV in univariable and multivariable results, with higher risk for those starting age 25+ in multivariable analyses. Risk was slightly higher in females (OR=1.26) and for those that had ever received works (OR=1.24). First use of needle exchange after the first year of injecting was associated with a small increase in risk compared to uptake in the first year (OR=1.12) and also never used (OR=1.23) although the latter was not significant. Days injecting per month and number of partners showed no effect. Ever being imprisoned was associated with increased risk, but not as high as HCV, with ORs of 1.29 and 1.45 for imprisonment before and after started injecting respectively. Imperfect condom use showed a slight protective effect compared to always

using a condom; and MSM had an increase in risk, with an identical OR to HCV (1.25). It is somewhat surprising that the sexual risk factors did not play more of a role for HBV, which is much more easily transmitted via the sexual route than HCV. The goodness of fit for the multivariable model was unclear; the Hosmer-Lemeshow $p$-value was 0.071 for deciles of the predicted probabilities, but again sensitive to the number of groups used; with 15 equal-width groups of the predicted probabilities the $p$-value was 0.145. However, neither indicate a severe lack of fit.

After adjusting for other risk factors, the LR test for an interaction between injecting duration and survey year gave a $p$-value of 0.061, with AIC scores preferring the simpler model (13064.4 vs. 13093.0). There were no significant interactions between injecting duration and other risk factors, but a number of interactions with survey period, including region ($p < 0.001$), gender ($p < 0.001$), needle exchange ($p$=0.040) and injecting 14 or more days per month ($p$=0.011). Females had a relatively lower risk of HBV infection in earlier survey periods, with ORs of 0.53 (95% CI 0.40-0.72) for 2000-2002 and 0.61 (95% CI 0.45-0.83) for 2003-2005 vs. 2006-2008; and those injecting for 14 or more days per month had a higher risk in 2000-2002 (OR=1.55, 95% CI 1.19-2.02) and 2012-2014 (OR=1.40, 95% CI 1.02-1.91). Region and time again showed the strongest interaction, and is summarised by way of predicted HBV prevalence in Figure 4.15. SE England, the baseline group, has a relatively stable trend but somewhat lower prevalence to other regions in 2000-2002, therefore many regions have a significant interaction for this period. NW England, East Midlands and West Midlands all have a decline in prevalence of around two-thirds, and most areas show a general decline; the trend for the NW appears particularly striking due to the high observed prevalence in 2000-2002. Some temporal patterns are harder to interpret: Yorkshire and the Humber for instance had very low prevalence in 2006-2008 and a sharp increase subsequently.

Table 4.10 shows $p$-values for interactions between risk factors, adjusted for all risk factors and the survey period/region interaction. Interactions with region were again the strongest, with generally little effect for the interaction between other variables. The strongest interaction was for gender and region, which showed lower prevalence in females for almost all regions vs. SE England, and markedly so for London (OR=0.54, $p = 0.001$), NW

Table 4.8: Univariable and multivariable results from logistic regression model for HBV and reported risk factors. Odds ratios and 95% confidence intervals.

| Variable | | Univariable | Multivariable |
|---|---|---|---|
| Injecting duration | 0-1 | 0.51 (0.39, 0.66) | 0.48 (0.37, 0.64) |
| | 2 | 0.65 (0.49, 0.86) | 0.63 (0.47, 0.84) |
| | 3-4 | 0.98 (0.79, 1.22) | 0.94 (0.75, 1.17) |
| | 5-6 | 1 (ref) | 1 (ref) |
| | 7-8 | 1.16 (0.93, 1.45) | 1.17 (0.93, 1.46) |
| | 9-11 | 1.78 (1.47, 2.16) | 1.81 (1.49, 2.21) |
| | 12-14 | 2.18 (1.80, 2.65) | 2.37 (1.93, 2.90) |
| | 15-19 | 2.99 (2.49, 3.59) | 3.26 (2.68, 3.97) |
| | 20-24 | 4.75 (3.91, 5.76) | 5.26 (4.26, 6.50) |
| | 25-29 | 7.45 (5.93, 9.38) | 9.39 (7.28, 12.11) |
| | 30-34 | 11.61 (8.69, 15.51) | 15.33 (11.17, 21.05) |
| | 35+ | 12.78 (8.30, 19.68) | 15.81 (9.98, 25.05) |
| Survey period | 2000-2002 | 1.37 (1.22, 1.55) | 1.57 (1.38, 1.79) |
| | 2003-2005 | 1.36 (1.20, 1.54) | 1.37 (1.19, 1.56) |
| | 2006-2008 | 1 (ref) | 1 (ref) |
| | 2009-2011 | 0.86 (0.74, 0.99) | 0.76 (0.65, 0.88) |
| | 2012-2014 | 0.94 (0.81, 1.08) | 0.79 (0.67, 0.93) |
| Region | East of England | 0.97 (0.78, 1.20) | 0.88 (0.70, 1.11) |
| | London | 1.80 (1.54, 2.09) | 1.24 (1.05, 1.47) |
| | South East | 1 (ref) | 1 (ref) |
| | South West | 0.85 (0.72, 1.00) | 0.84 (0.70, 1.00) |
| | West Midlands | 0.38 (0.29, 0.49) | 0.46 (0.34, 0.61) |
| | North West | 2.30 (1.98, 2.67) | 1.98 (1.69, 2.33) |
| | Yorkshire and H | 0.70 (0.54, 0.91) | 0.70 (0.53, 0.92) |
| | East Midlands | 0.72 (0.60, 0.88) | 0.84 (0.68, 1.03) |
| | North East | 0.53 (0.44, 0.64) | 0.81 (0.67, 0.99) |
| | Wales | 0.57 (0.45, 0.73) | 0.60 (0.46, 0.78) |
| Age at first use | <18 | 1.30 (1.19, 1.44) | 0.84 (0.75, 0.94) |
| | 18-24 | 1 (ref) | 1 (ref) |
| | 25+ | 0.92 (0.83, 1.02) | 1.46 (1.29, 1.64) |
| Gender | Male | 1 (ref) | 1 (ref) |
| | Female | 0.88 (0.80, 0.96) | 1.26 (1.13, 1.40) |

Table 4.9: *Continued from 4.8*: Logistic regression model results for HBV.

| Variable | | Univariable | Multivariable |
|---|---|---|---|
| Ever rec'd works | No | 1 (ref) | 1 (ref) |
| | Yes | 1.52 (1.40, 1.65) | 1.24 (1.13, 1.36) |
| Needle exchange | Started 1st year | 1 (ref) | 1 (ref) |
| | Started >1 yr | 1.98 (1.82, 2.15) | 1.12 (1.01, 1.23) |
| | Never | 1.19 (0.90, 1.58) | 1.23 (0.90, 1.67) |
| Days injecting per month | <14 days/mo | 1 (ref) | 1 (ref) |
| | 14+ days/mo | 0.94 (0.86, 1.02) | 1.03 (0.94, 1.13) |
| Prison | Never imprisoned | 1 (ref) | 1 (ref) |
| | Prison before injecting | 1.93 (1.72, 2.15) | 1.29 (1.14, 1.46) |
| | Prison after injecting | 1.60 (1.44, 1.78) | 1.45 (1.29, 1.64) |
| Number of partners | 0 or 1 partner | 1 (ref) | 1 (ref) |
| | 2+ partners | 0.89 (0.81, 0.96) | 1.07 (0.97, 1.17) |
| Condom use | Always | 1 (ref) | 1 (ref) |
| | Sometimes/never | 0.70 (0.63, 0.76) | 0.83 (0.75, 0.92) |
| MSM | No | 1 (ref) | 1 (ref) |
| | Yes | 1.34 (1.04, 1.71) | 1.25 (0.95, 1.64) |

(OR=0.46, $p < 0.001$) and E Mids (OR=0.50, $p = 0.008$).

### 4.1.7 Risk factors for HIV infection

The models described in section 4.1.4 were fitted to the HIV data, but due to the low prevalence, have less scope for investigating risk patterns. Injecting duration and survey year were grouped into broader categories as in section 4.1.3. Model results for univariable and multivariable main effects logistic models are shown in tables 4.11 and 4.12; all estimates presented subsequently are for the multivariable model.

Interestingly, after adjusting for the covariates on demographics and risk factors the ORs for injecting durations of 20-34 and 35+ years vs. 5-9 years show no increase in prevalence ($p$-values 0.101 and 0.978 respectively). The attenuation compared to the model in section 4.1.3 largely occurs when region is adjusted for, with substantial variation in injecting duration by region. London had the longest median injecting duration at 12 years, compared to an overall median of 9 years and as low as 5 in the NE. These are

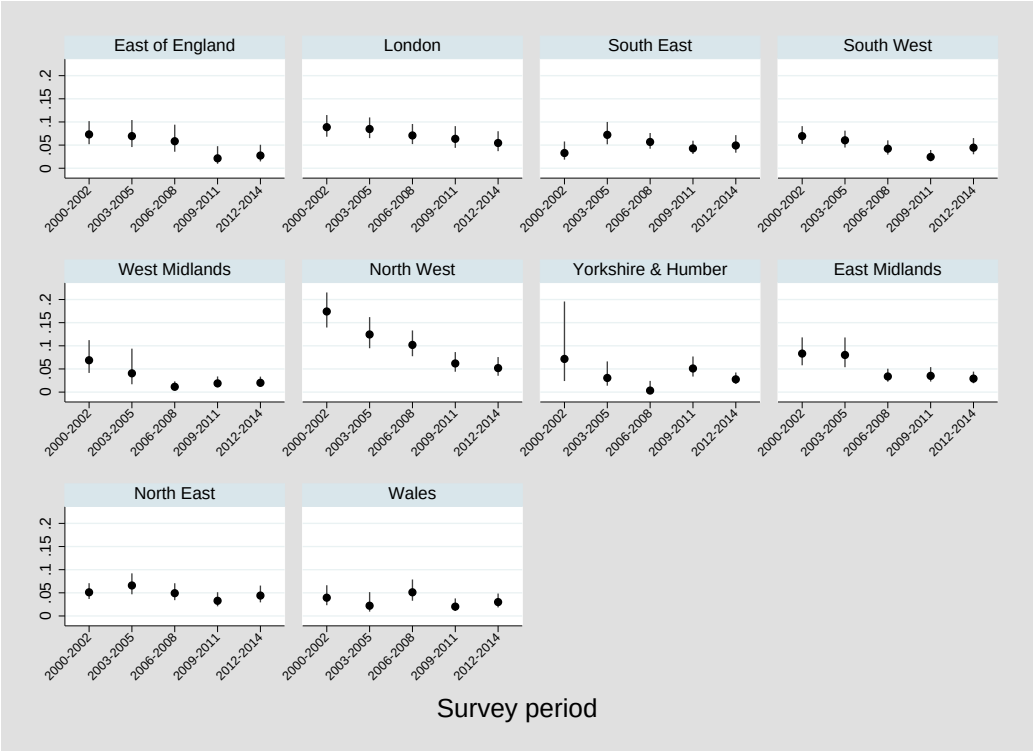


Figure 4.15: Predicted HBV prevalence by region and survey period with 95% confidence intervals, multivariable logistic model adjusting for main effects of all risk factors and region/survey period interaction. Predictions are made at baseline levels of other risk factors; e.g., injecting duration 5-6 years.

the regions with the highest and lowest prevalences respectively, so it appears that the more extreme unadjusted differences in injecting duration are due, in part, to confounding with region. The only regions with significantly different HIV prevalence were London, which was far higher than SE England (OR=5.31, 95% CI 3.03-9.32) and NE, which was far lower (OR=0.15, 95% CI 0.04-0.68). However, data are too sparse in most regions to estimate differences with any confidence.

Those that began injecting below the age of 18 had an increased risk (OR=1.85, 95% CI 1.25-2.73), in contrast to the results for HCV and HBV; and ever receiving works was associated with higher risk (OR=1.88, 95% CI 1.32-2.68). The only other significant effects were MSM, which was associated with a greatly increased risk of HIV infection (OR=4.91 95% CI 2.73-8.85) and condom use, with incomplete use (vs. always used) showing

Table 4.10: Interactions between risk factors for HBV, likelihood ratio test $p$-values for generalised linear model with main effects, region/survey year interaction and interaction of interest. MSM/female interaction is not defined.

| | region | aafu | female | erec | exch | d14 | pris | part2 | cond |
|---|---|---|---|---|---|---|---|---|---|
| aafu | 0.005 | | | | | | | | |
| female | 0.000 | 0.649 | | | | | | | |
| erec | 0.414 | 0.962 | 0.084 | | | | | | |
| exch | 0.130 | 0.470 | 0.298 | 0.059 | | | | | |
| d14 | 0.145 | 0.117 | 0.836 | 0.212 | 0.693 | | | | |
| pris | 0.032 | 0.096 | 0.250 | 0.750 | 0.024 | 0.740 | | | |
| part2 | 0.046 | 0.374 | 0.791 | 0.574 | 0.256 | 0.640 | 0.942 | | |
| cond | 0.037 | 0.976 | 0.734 | 0.348 | 0.882 | 0.761 | 0.014 | 0.818 | |
| msm | 0.780 | 0.129 | . | 0.623 | 0.293 | 0.311 | 0.024 | 0.523 | 0.094 |

**aafu**: age at first use, **erec**: ever received works, **exch**: needle exchange use, **d14**: injected 14 or more days in last month, **pris**: imprisonment, **part2**: two or more sexual partners, **cond**: condom use, **msm**: men who have sex with men.

an unexpected and substantial protective effect (OR=0.27 95% 0.19-0.37), as for HCV and HBV.

After adjusting for other risk factors, the interaction between injecting duration and survey year was not significant, with a LR test $p$-value of 0.152. There were no significant associations for injecting duration and survey year with risk factor variables, with the lowest $p$-value being 0.208; Wald tests were employed as many interactions resulted in zero cells and a differing number of observations between models. There were just a few significant interactions between risk factors, with the strongest being between gender and prison ($p$=0.012); in males, the baseline group, ORs for being imprisoned before/after starting injecting were below 1 at 0.62 (95% CI 0.39-0.98) and 0.50 (95% CI 0.31-0.80) respectively; the OR for female vs. males changed to become protective (OR=0.43, 95% CI 0.21-0.85), but females that had been to prison had a higher risk, with OR of 2.33 (95% CI 0.93-5.88) and 4.38 (95% CI 1.65-11.62) for imprisonment before or after starting injecting respectively. This could potentially be explained by prostitution, which would likely increase the risk of HIV infection and possibly imprisonment; although the number of sexual partners had little influence, nor was there an interaction with gender ($p$=0.114).

Table 4.11: Univariable and multivariable results from logistic regression model for HIV and reported risk factors. Odds ratios and 95% confidence intervals.

| Variable | | Univariable | Multivariable |
|---|---|---|---|
| Injecting duration | 0-1 | 1.11 (0.48, 2.55) | 1.26 (0.53, 3.00) |
| | 2-4 | 1.30 (0.67, 2.53) | 1.54 (0.78, 3.03) |
| | 5-9 | 1 (ref) | 1 (ref) |
| | 10-14 | 2.48 (1.40, 4.39) | 2.10 (1.17, 3.77) |
| | 15-19 | 4.15 (2.37, 7.30) | 2.64 (1.46, 4.80) |
| | 20-34 | 4.43 (2.52, 7.79) | 1.68 (0.90, 3.14) |
| | 35+ | 5.14 (1.18, 22.49) | 1.02 (0.22, 4.82) |
| Survey period | 2000-2004 | 0.68 (0.47, 0.98) | 0.53 (0.36, 0.78) |
| | 2005-2009 | 1 (ref) | 1 (ref) |
| | 2010-2014 | 0.73 (0.49, 1.09) | 0.97 (0.63, 1.48) |
| Region | East of England | 0.93 (0.38, 2.27) | 1.18 (0.47, 2.93) |
| | London | 4.94 (2.89, 8.45) | 5.31 (3.03, 9.32) |
| | South East | 1 (ref) | 1 (ref) |
| | South West | 0.50 (0.23, 1.11) | 0.60 (0.27, 1.33) |
| | West Midlands | 0.36 (0.10, 1.23) | 0.43 (0.12, 1.50) |
| | North West | 0.92 (0.46, 1.82) | 1.12 (0.55, 2.27) |
| | Yorkshire and H | 0.18 (0.02, 1.34) | 0.19 (0.02, 1.45) |
| | East Midlands | 0.72 (0.32, 1.64) | 0.83 (0.36, 1.92) |
| | North East | 0.11 (0.02, 0.46) | 0.15 (0.04, 0.68) |
| | Wales | 0.67 (0.24, 1.83) | 0.85 (0.31, 2.35) |
| Age at first use | <18 | 1.96 (1.37, 2.79) | 1.85 (1.25, 2.73) |
| | 18-24 | 1 (ref) | 1 (ref) |
| | 25+ | 1.00 (0.65, 1.53) | 0.95 (0.60, 1.52) |
| Gender | Male | 1 (ref) | 1 (ref) |
| | Female | 0.89 (0.62, 1.30) | 0.91 (0.61, 1.36) |

Table 4.12: *Continued from 4.11*: Logistic regression model results for HIV.

| Variable | | Univariable | Multivariable |
|---|---|---|---|
| Ever rec'd works | No | 1 (ref) | 1 (ref) |
| | Yes | 2.22 (1.59, 3.09) | 1.88 (1.32, 2.68) |
| Needle exchange | Started 1st year | 1 (ref) | 1 (ref) |
| | Started >1 yr | 1.89 (1.36, 2.62) | 1.10 (0.76, 1.58) |
| | Never | 1.09 (0.34, 3.50) | 0.80 (0.24, 2.64) |
| Days injecting per month | <14 days/mo | 1 (ref) | 1 (ref) |
| | 14+ days/mo | 0.72 (0.53, 1.00) | 0.94 (0.68, 1.31) |
| Prison | Never imprisoned | 1 (ref) | 1 (ref) |
| | Prison before injecting | 1.10 (0.76, 1.61) | 0.78 (0.52, 1.18) |
| | Prison after injecting | 0.65 (0.44, 0.96) | 0.72 (0.47, 1.10) |
| Number of partners | 0 or 1 partner | 1 (ref) | 1 (ref) |
| | 2+ partners | 1.19 (0.86, 1.63) | 1.16 (0.82, 1.64) |
| Condom use | Always | 1 (ref) | 1 (ref) |
| | Sometimes/never | 0.24 (0.18, 0.34) | 0.27 (0.19, 0.37) |
| MSM | No | 1 (ref) | 1 (ref) |
| | Yes | 5.39 (3.22, 9.01) | 4.91 (2.73, 8.85) |

### 4.1.8 Summary of risk factors

The results from generalised linear models for HCV, HBV and HIV indicate that there were significant changes in prevalence according to survey period. Prevalence of all three blood-borne viruses increases with injecting duration, which is taken to be the time at risk, as would be expected for infections with long-lasting antibodies. Likelihood ratio tests indicated interactions between injecting duration and survey period, although there was less evidence of interactions in the data from 2000 onwards. In general, interaction terms were estimated imprecisely and could be modelled more efficiently (for instance, using parametric functions, smoothing, or within a Bayesian framework), although the main aim here was to determine whether main effects models were sufficient.

There were also important risk factors for infection, with region having the strongest effect of any covariate. Prevalence of HCV is markedly higher in London and the NW, HBV is very high in the NW, and there is far higher HIV prevalence in London compared to all other regions. Trends in

prevalence over the survey period also varied substantially by region. The decrease in HBV prevalence over time is likely attributable to vaccination, but varies across regions. Injecting epidemics were particularly severe in the NW and London, and mixing with large migrant populations from high-prevalence countries may have accelerated the spread of BBVs in these areas. In this case, the subsequent effect of vaccination may be greater in these areas, reducing both within- and between-group transmission.

Ever being imprisoned and ever receiving works were both associated with around a two-fold increased risk of infection for HCV and HIV, and to a lesser extent HBV. Some results are counter-intuitive: incomplete condom use (compared to always using a condom) appeared to have a protective effect for all infections, and never having used a needle exchange also appeared to confer a lower risk. MSM were found to be at somewhat greater risk of HCV and HBV infection, and markedly so for HIV. Starting injecting at an older age (25+) was associated with an increased risk of HCV and HBV infection, although those starting below the age of 18 were at higher risk of HIV infection.

There is a possibility of bias in the results for condom use and number of partners due to systematic missingness. Both are relatively sensitive questions, and prevalence of all BBVs is somewhat higher in those with missing information. Techniques such as multiple imputation may be used to handle missing data issues, and can reduce potential biases under the *missing at random* assumption (Rubin, 1987). However, systematic differences in missing data that are not accounted for by relationships between the observed covariates (*missing not at random*) are always a possibility, and there is no guarantee of obtaining unbiased estimates (Sterne et al., 2009). Further, sensitivity analyses were undertaken in which all missing responses were assumed to have, or not have, the risk factor, with little change in estimated odds ratios: the apparent protective effect of imperfect condom use persisted for all infections, and results on numbers of partners were still inconclusive for HCV and HBV.

Subsequently, condom use and number of partners are no longer considered, due to the unusual effect estimates and because answers are frequently omitted from questionnaires, resulting in nearly a quarter of the post-2000 data being omitted. Main effects models for each of the infections were re-

estimated with these variables excluded to ensure that multivariable results from section 4.1.4 did not change markedly.

An important factor that has not been considered here is vaccination for HBV. This is self-reported in the UAM data and could in theory be used for analysis; however, interpretation is somewhat difficult as vaccination is often provided regardless of infection status. Therefore although the decrease in HBV prevalence over time is likely to be attributable to increased vaccination coverage, the direct effect of HBV vaccination on prevalence cannot be tested within this modelling framework.

## 4.2 Models for the force of infection

This chapter has so far investigated the relationship between injecting duration and calendar time with prevalence of HCV, HBV and HIV, including the effect of additional covariates and the various possible interactions that may be considered. A large number of parameters are required to adequately model the effect of injecting duration. This will hold generally for age-specific current status data, in order for the piecewise constant bands to adequately capture the relationship between exposure time and risk described by equation 3.13. By incorporating this relationship in a model for the force of infection (FOI), fewer parameters are needed, as only changes to the infection rate according to time at risk need be parameterised. If the FOI is constant, only one parameter is required.

In this section FOI models are fitted that allow for changes in the risk of infection by injecting duration and calendar time. So far the imperfect sensitivity of pre-dried blood spot tests for HCV and HBV has not been accounted for, which would result in apparently lower prevalence in earlier survey years. While this can be taken into consideration for effect estimates of survey period, the effect of imperfect sensitivity may be more subtle on a FOI model, potentially inducing an interaction between injecting duration and survey period where there may be none. Therefore, models are fitted that account for the sensitivity of tests used to recover the true prevalence (given the assumed sensitivity), and therefore true FOI, by including the relationship specified in equation 3.5 into the log-likelihood. This requires that models are fitted using bespoke code, and are optimised using the BFGS

method (Nocedal et al., 2006, p. 194) in the R routine `optim`.

Piecewise constant models for the force of infection (FOI) were fitted to the HCV, HBV and HIV data, including effects for injecting duration and calendar time as in section 3.2.3, where injecting duration is taken to be the "age" or time at risk for people who inject drugs (PWID). Calendar time is split into pre-1980, 1980-1985, 1985-1990, 1995-2000, 2000-2005, 2005-2010 and 2010 onwards; and injecting duration as $\leq 1$, 1-3, 3-5, 5-10, 10-15, 15-25 and $> 25$ years. Injecting duration is assumed to be at least 1 year in all participants. Four models are tested for the effects of injecting duration and time for each infection: with and without interactions between injecting duration and time, and additive (equation 3.18) versus multiplicative (equation 3.19) injecting duration and time effects. Interaction models are specified via main effects for injecting duration and calendar time, with deviations from the main effects estimated via time- and injecting duration-specific interaction terms, as described in section 3.2.3. These interactions are only estimable within the survey period, and the earliest period in the survey is taken as the baseline, such that the relative injecting duration-specific risks are assumed to be the same prior to and at the beginning of the survey.

Having assessed different model forms for age and time, covariates are included in the model as described in section 3.2.5. The additive model turns out to provide an adequate fit, and since full covariate data are only available from 2000 onwards and sensitivity to HCV antibodies is good as of 1998 and does not need to be accounted for, models can be fitted in the GLM framework as specified in equation 3.22, without needing to resort to bespoke code for maximising the likelihood. The GLM approach with additive effects is therefore used to further explore and refine the relationships between the covariates and infection status, with HCV being the main outcome of interest.

### 4.2.1 Force of infection models for HCV by injecting duration and time

The deviances for the different parameterisations of the FOI for HCV are shown in Table 4.13. For all data, there was a large improvement in model fit for the interaction models compared to main effects for injecting dura-

tion and time only (multiplicative model 859.7 on 711 d.f. vs. 979.4 on 735 d.f.). There were smaller, but still substantial differences between additive and multiplicative models, with the multiplicative model providing a better fit without interactions, but the best fit from the additive model with interactions. Differences in fit are expected for the main effects model, which could provide a quite different fit to the data if injecting duration and time effects are combined independently. However, the difference in fit appears unexpected for the interaction models, as both additive and multiplicative models contain a parameter for each injecting duration/survey period. The difference is likely due to the constraint that the FOI is strictly positive in the multiplicative model, but can be negative in the additive model, allowing a better fit where observed prevalence does not increase monotonically with time at risk. In addition, although no direct information is available on interaction terms prior to the survey period, the constraint that the FOI in the pre-survey period has constant hazard ratios/differences for injecting duration over time could lead to a conflict with the observed prevalence in the earliest survey period, leading to a difference in fit.

In the models fitted to data from 2000 onwards, the difference in deviances between the main effects and interaction models is less extreme (multiplicative model 615.8 on 560 d.f. vs. 643.8 on 572 d.f.). This is similar to the logistic model for HCV prevalence in section 4.1.1, which also showed a smaller improvement with the addition of interaction terms for the restricted data. The multiplicative model fits better for the main effects only model, but the difference in deviances is only 4.4 for the interaction models.

Table 4.13: Deviances and degrees of freedom for force of infection models for HCV according to injecting duration and time, additive vs. multiplicative and main effects vs. interaction models. Results are presented for all available data, and data from 2000 onwards. The degrees of freedom is the number of unique data points (injecting duration/survey year combinations) minus the number of parameters in the model.

| | | All data | | 2000 onwards | |
|---|---|---|---|---|---|
| Model | Parameterisation | Deviance | d.f. | Deviance | d.f. |
| Main effects | Additive | 995.7 | 735 | 661.1 | 572 |
| Main effects | Multiplicative | 979.4 | 735 | 643.8 | 572 |
| Interaction | Additive | 836.3 | 711 | 620.2 | 560 |
| Interaction | Multiplicative | 859.7 | 711 | 615.8 | 560 |

Figure 4.16 shows predicted HCV prevalence from the models above fitted to data from all years; the additive and multiplicative main effects models give surprisingly similar predictions (given they might be expected to diverge more than the interaction results) and the multiplicative main effects-only model results are therefore omitted for clarity. All models performed poorly at the beginning period of the survey data, years 1992 and 1994, and especially 1992. This is likely due to the constraints imposed on the FOI in pre-survey years. There is little to distinguish between model predictions in later survey years, and virtually no difference between the additive and multiplicative interaction models. This supports the argument that the constraint of constant injecting duration hazard ratios/differences in pre-survey years can influence model fit, despite there being no direct information to estimate differences over time in the pre-survey period. All of the models appear to provide a reasonable fit visually, except for some systematic divergence in particular years. For instance, prevalence in 2005 for those injecting for more than 12 years is systematically underestimated, but not for 2006. This could point to a different composition in the UAM sample for certain years.

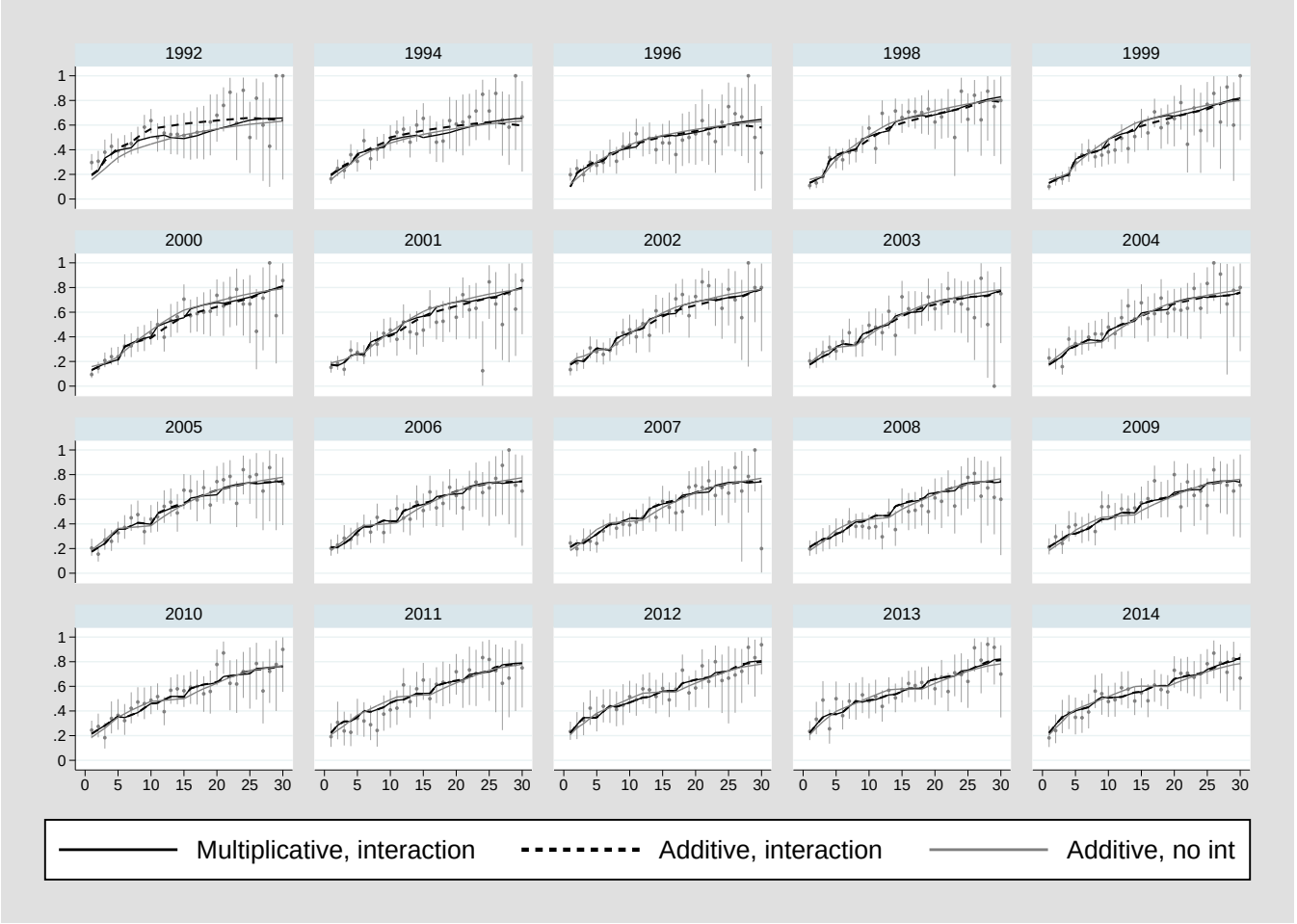


Figure 4.16: Predicted HCV prevalence by injecting duration and survey year from the force of infection models. Additive vs. multiplicative and main effects vs. interaction models are shown, although the main effects multiplicative model is omitted for clarity (see text).

Table 4.14 shows the estimated FOI by injecting duration and survey period from the multiplicative and additive interaction model. Both models indicate a high FOI in the first year of injecting, which decreases by around 5-fold subsequently. The estimates are very similar within the survey period, except that for some injecting durations/times the estimate from the multiplicative model is practically zero, but slightly negative for the additive model. This is because the multiplicative FOI is additive on the exponential scale and constrained to be positive, whereas the additive model components can combine to produce negative estimates if the injecting duration and time-specific FOI does not increase monotonically. Outside the range of the survey, there are slight differences in the estimated FOI due to the constant injecting duration effects being either additive or multiplicative.

Table 4.14: Force of infection for HCV by injecting duration and time, multiplicative and additive interaction models.

| Multiplicative | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.546 | 0.135 | 0.085 | 0.014 | 0.001 | 0.108 | 0.000 |
| 1980-1985 | 0.633 | 0.156 | 0.098 | 0.016 | 0.001 | 0.125 | 0.000 |
| 1985-1990 | 0.449 | 0.111 | 0.070 | 0.012 | 0.000 | 0.089 | 0.000 |
| 1990-1995 | 0.296 | 0.073 | 0.046 | 0.008 | 0.000 | 0.058 | 0.000 |
| 1995-2000 | 0.144 | 0.032 | 0.021 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2000-2005 | 0.198 | 0.044 | 0.086 | 0.055 | 0.062 | 0.044 | 0.024 |
| 2005-2010 | 0.246 | 0.045 | 0.054 | 0.034 | 0.043 | 0.034 | 0.013 |
| 2010-2015 | 0.253 | 0.090 | 0.045 | 0.071 | 0.044 | 0.048 | 0.106 |
| Additive | | | | | | | |
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.483 | 0.252 | 0.230 | 0.171 | 0.056 | 0.046 | -0.043 |
| 1980-1985 | 0.477 | 0.246 | 0.224 | 0.165 | 0.051 | 0.041 | -0.049 |
| 1985-1990 | 0.389 | 0.157 | 0.136 | 0.076 | -0.038 | -0.048 | -0.137 |
| 1990-1995 | 0.309 | 0.077 | 0.056 | -0.004 | -0.118 | -0.128 | -0.217 |
| 1995-2000 | 0.145 | 0.031 | 0.020 | -0.012 | -0.017 | -0.025 | 0.025 |
| 2000-2005 | 0.198 | 0.043 | 0.087 | 0.056 | 0.074 | 0.063 | 0.027 |
| 2005-2010 | 0.246 | 0.045 | 0.055 | 0.034 | 0.041 | 0.029 | 0.018 |
| 2010-2015 | 0.253 | 0.090 | 0.045 | 0.070 | 0.045 | 0.050 | 0.100 |

Table 4.15 shows the estimated FOI from models fitted to data from 2000 onwards. Data within the survey period, except for 2000-2005, show very similar estimates to those from the complete data. However, estimates for

the pre-survey period are very different: whereas the FOI was estimated to be generally high at shorter injecting durations prior to the year 2000 and highest in 1980-1985, estimates from the restricted dataset indicate little difference in FOI for this group pre-2000. Conversely, the estimated FOI in longer-term injectors (> 25 years) is estimated to be higher in the pre-2000 period. The data from 2000 appear to be more generally consistent, with fewer negative estimates of the FOI and less reliance on interactions to capture the variability of the data. There are questions as to the reliability of HCV data in earlier years (Vivian Hope, *personal communication*) so using only more recent data might give more reliable estimates. However, it may be that important changes in injecting-duration specific risks were occurring in the 1990s, which are not captured without the earlier data.

Table 4.15: Force of infection for HCV by injecting duration and time, multiplicative and additive interaction models, fitted to data from 2000 onwards.

| Multiplicative | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.057 | 0.015 | 0.026 | 0.014 | 0.019 | 0.017 | 0.034 |
| 1980-1985 | 0.200 | 0.053 | 0.091 | 0.050 | 0.067 | 0.058 | 0.120 |
| 1985-1990 | 0.177 | 0.047 | 0.081 | 0.044 | 0.059 | 0.051 | 0.106 |
| 1990-1995 | 0.171 | 0.045 | 0.078 | 0.043 | 0.057 | 0.049 | 0.102 |
| 1995-2000 | 0.130 | 0.035 | 0.059 | 0.032 | 0.044 | 0.038 | 0.078 |
| 2000-2005 | 0.192 | 0.051 | 0.087 | 0.048 | 0.064 | 0.055 | 0.115 |
| 2005-2010 | 0.245 | 0.049 | 0.051 | 0.035 | 0.046 | 0.032 | 0.000 |
| 2010-2015 | 0.253 | 0.091 | 0.043 | 0.071 | 0.044 | 0.049 | 0.108 |
| Additive | | | | | | | |
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.113 | -0.016 | 0.015 | -0.016 | -0.008 | -0.017 | 0.063 |
| 1980-1985 | 0.192 | 0.063 | 0.093 | 0.062 | 0.071 | 0.062 | 0.142 |
| 1985-1990 | 0.176 | 0.047 | 0.078 | 0.047 | 0.056 | 0.047 | 0.127 |
| 1990-1995 | 0.188 | 0.059 | 0.089 | 0.058 | 0.067 | 0.058 | 0.138 |
| 1995-2000 | 0.151 | 0.022 | 0.053 | 0.022 | 0.031 | 0.022 | 0.102 |
| 2000-2005 | 0.182 | 0.053 | 0.084 | 0.053 | 0.062 | 0.053 | 0.133 |
| 2005-2010 | 0.243 | 0.050 | 0.055 | 0.033 | 0.044 | 0.034 | -0.008 |
| 2010-2015 | 0.253 | 0.092 | 0.042 | 0.070 | 0.045 | 0.049 | 0.114 |

### 4.2.2 Force of infection models for HBV by injecting duration and time

The same models were then fitted to the HBV data, and deviances shown in Table 4.16. The interaction model has substantially lower deviance for both the full data and that restricted to 2000 onwards. None of the models fit the data well for the full dataset, although the additive interaction model is preferred in all comparisons and provides a reasonable fit to the data from 2000 onwards. As before however, the additive model will tend to overstate the goodness of fit due to the allowance of negative FOI estimates.

Table 4.16: Deviances and degrees of freedom for force of infection models for HBV according to injecting duration and time, additive vs. multiplicative and main effects vs. interaction models. Results are presented for all available data, and data from 2000 onwards.

| | | All data | | 2000 onwards | |
|---|---|---|---|---|---|
| Model | Parameterisation | Deviance | d.f. | Deviance | d.f. |
| Main effects | Additive | 1247.4 | 881 | 697.2 | 572 |
| Main effects | Multiplicative | 1324.6 | 881 | 793.6 | 572 |
| Interaction | Additive | 1199.6 | 857 | 659.1 | 560 |
| Interaction | Multiplicative | 1283.7 | 857 | 760.2 | 560 |

Figure 4.18 shows predicted HBV prevalence from the models above fitted to data from all years. As with HCV, the predicted prevalences were similar for the additive and multiplicative main effects models, so the latter is not shown. The main effects model performs poorly in earlier survey years, with a systematic underestimation of prevalence. Again, the additive and multiplicative models produce similar predictions for all but the first few survey years.

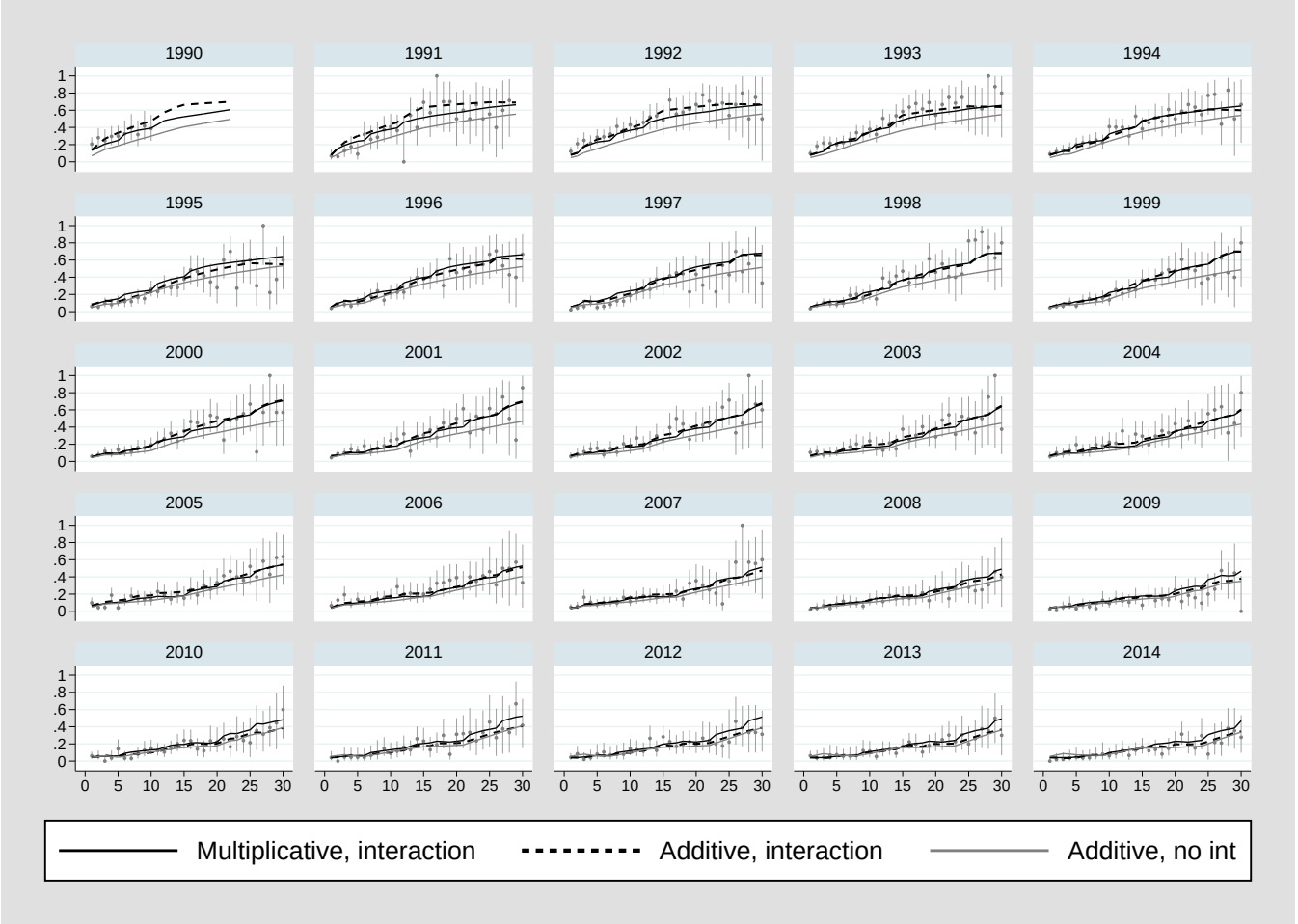


Figure 4.17: Predicted HBV prevalence by injecting duration and survey year from the force of infection models. Additive vs. multiplicative and main effects vs. interaction models are shown, although the main effects multiplicative model is omitted for clarity (see text).

Tables 4.17 and 4.18 show the force of infection from multiplicative and additive interaction models for all data and data from 2000 onwards. The two model forms produce very different patterns for the injecting-duration specific risk in the pre-survey period, with relatively high risk in 1st year injectors and those injecting for $> 25$ years in the multiplicative model. Both models indicate a general decline in the FOI over calendar time, which is expected due to increasing levels of HBV vaccination in the PWID population. However, the FOI did not decline to the same relative degree in 1st year injectors compared to longer durations, whose risk approached zero from 2000 onwards. In contrast to the HCV data, there were smaller differences in the prediction of pre-survey risk levels when the data were restricted to 2000 onwards.

Table 4.17: Force of infection for HBV by injecting duration and time, multiplicative and additive interaction models.

| Multiplicative | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.531 | 0.160 | 0.068 | 0.099 | 0.018 | 0.000 | 0.309 |
| 1980-1985 | 0.303 | 0.092 | 0.039 | 0.056 | 0.010 | 0.000 | 0.176 |
| 1985-1990 | 0.192 | 0.058 | 0.024 | 0.036 | 0.006 | 0.000 | 0.111 |
| 1990-1995 | 0.104 | 0.031 | 0.013 | 0.019 | 0.003 | 0.000 | 0.061 |
| 1995-2000 | 0.082 | 0.025 | 0.010 | 0.015 | 0.003 | 0.000 | 0.047 |
| 2000-2005 | 0.083 | 0.025 | 0.011 | 0.015 | 0.003 | 0.000 | 0.048 |
| 2005-2010 | 0.049 | 0.009 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2010-2015 | 0.042 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Additive | | | | | | | |
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.177 | 0.119 | 0.089 | 0.104 | 0.102 | 0.080 | 0.135 |
| 1980-1985 | 0.153 | 0.094 | 0.065 | 0.079 | 0.078 | 0.056 | 0.110 |
| 1985-1990 | 0.139 | 0.080 | 0.051 | 0.065 | 0.064 | 0.042 | 0.096 |
| 1990-1995 | 0.095 | 0.037 | 0.007 | 0.021 | 0.020 | -0.002 | 0.053 |
| 1995-2000 | 0.088 | 0.030 | 0.000 | 0.014 | 0.013 | -0.009 | 0.046 |
| 2000-2005 | 0.097 | 0.039 | 0.009 | 0.024 | 0.022 | 0.001 | 0.055 |
| 2005-2010 | 0.052 | 0.014 | -0.015 | -0.016 | -0.009 | -0.026 | -0.066 |
| 2010-2015 | 0.045 | -0.004 | -0.004 | 0.004 | 0.008 | -0.004 | -0.010 |

Table 4.18: Force of infection for HBV by injecting duration and time, multiplicative and additive interaction models, fitted to data from 2000 onwards.

| Multiplicative | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.452 | 0.124 | 0.058 | 0.066 | 0.000 | 0.000 | 0.511 |
| 1980-1985 | 0.327 | 0.090 | 0.042 | 0.048 | 0.000 | 0.000 | 0.370 |
| 1985-1990 | 0.211 | 0.058 | 0.027 | 0.031 | 0.000 | 0.000 | 0.239 |
| 1990-1995 | 0.117 | 0.032 | 0.015 | 0.017 | 0.000 | 0.000 | 0.132 |
| 1995-2000 | 0.099 | 0.027 | 0.013 | 0.014 | 0.000 | 0.000 | 0.113 |
| 2000-2005 | 0.081 | 0.022 | 0.010 | 0.012 | 0.000 | 0.000 | 0.092 |
| 2005-2010 | 0.049 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2010-2015 | 0.042 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Additive | | | | | | | |
| Year / injecting | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.196 | 0.131 | 0.114 | 0.124 | 0.115 | 0.079 | 0.133 |
| 1980-1985 | 0.189 | 0.124 | 0.107 | 0.117 | 0.108 | 0.072 | 0.126 |
| 1985-1990 | 0.150 | 0.085 | 0.067 | 0.078 | 0.069 | 0.033 | 0.086 |
| 1990-1995 | 0.096 | 0.031 | 0.014 | 0.024 | 0.015 | -0.021 | 0.033 |
| 1995-2000 | 0.095 | 0.030 | 0.012 | 0.023 | 0.013 | -0.022 | 0.031 |
| 2000-2005 | 0.094 | 0.029 | 0.011 | 0.022 | 0.013 | -0.023 | 0.031 |
| 2005-2010 | 0.051 | 0.012 | -0.009 | -0.012 | -0.014 | -0.023 | -0.074 |
| 2010-2015 | 0.045 | -0.004 | -0.002 | 0.002 | 0.008 | -0.005 | -0.003 |

### 4.2.3 Force of infection models for HIV by injecting duration and time

Finally, the FOI models were fitted to the HIV data. Resulting deviances are shown in Table 4.19. When fitting to the whole dataset, the additive model with interactions provided the best fit, although the multiplicative model with interactions provided a better fit to the data from the year 2000 onwards. However, all of the deviances are smaller than the degrees of freedom, indicating over-fitting. In fact, the additive main effects model could not be fitted to the data, either in the R function `optim` or using standard GLM software, despite numerous attempts to improve starting values and relax tolerances.

Table 4.19: Deviances and degrees of freedom for force of infection models for HIV according to injecting duration and time, additive vs. multiplicative and main effects vs. interaction models. Results are presented for all available data, and data from 2000 onwards. Model fitting failed for the additive main effects model.

| | | All data | | 2000 onwards | |
|---|---|---|---|---|---|
| Model | Parameterisation | Deviance | d.f. | Deviance | d.f. |
| Main effects | Additive | - | 819 | - | 510 |
| Main effects | Multiplicative | 796.6 | 819 | 482.0 | 510 |
| Interaction | Additive | 758.6 | 795 | 465.4 | 486 |
| Interaction | Multiplicative | 764.5 | 795 | 460.6 | 486 |

Figure 4.18 shows predicted HIV prevalence from the models above fitted to data from all years. At the level of individual years and injecting durations, the data are too sparse for the model to capture the observed prevalence, except for very broad trends.

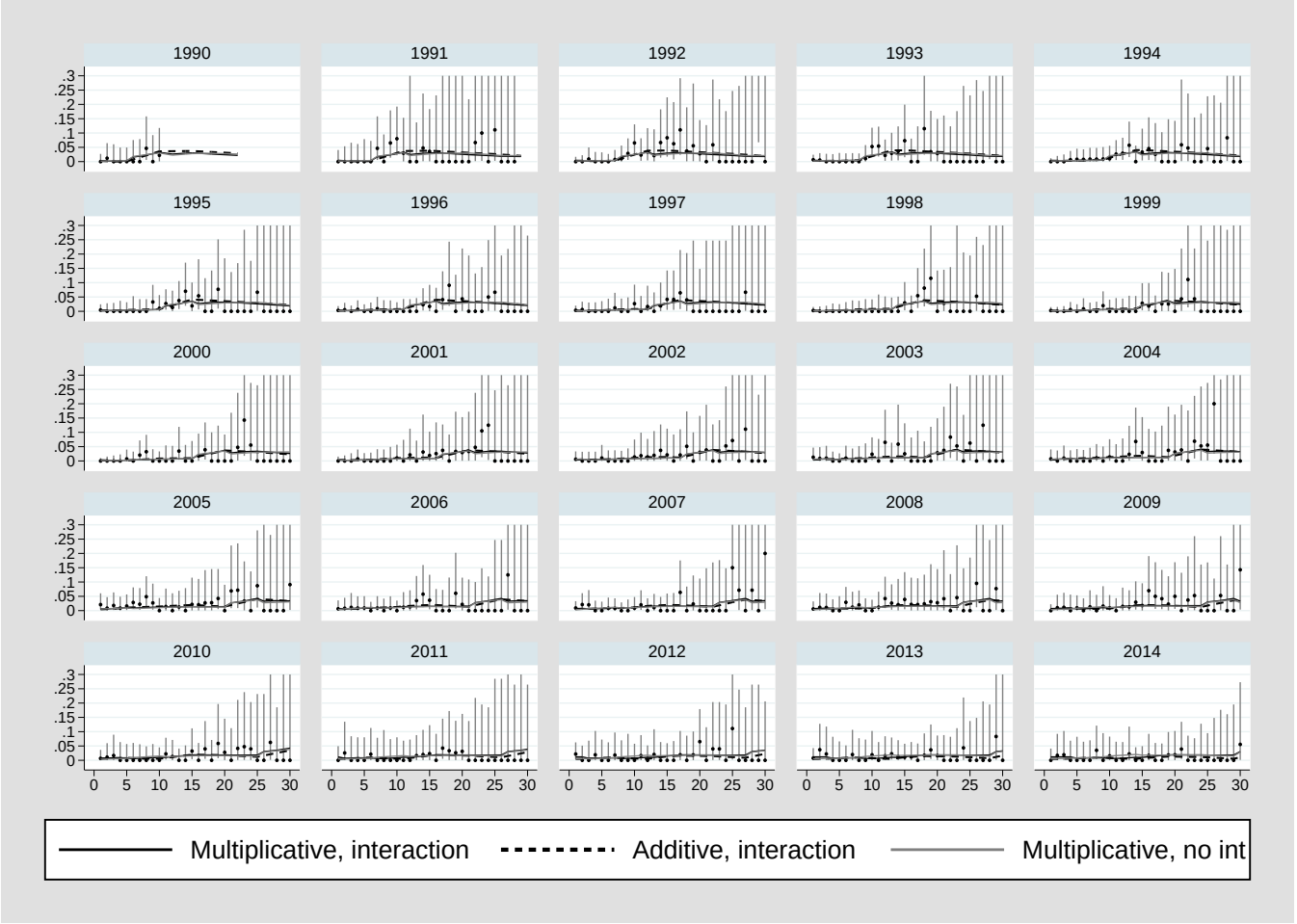


Figure 4.18: Predicted HIV prevalence by injecting duration and survey year from the force of infection models. Additive vs. multiplicative and main effects vs. interaction models are shown, although the main effects additive model is omitted (see text).

Tables 4.20 and 4.21 show the force of infection from multiplicative and additive interaction models for all data and data from 2000 onwards. Estimates are far more sensitive to choice of parameterisation and the inclusion/exclusion of pre-2000 data. Of note is that the multiplicative model predicts a near-zero FOI in the pre-1980 period whereas the additive model does not. The former is far more likely, with the first cases of HIV appearing in the early 1980s, although the data are so sparse this may not be meaningful. For all models, the highest risk was estimated to be in the first year and extremely low subsequently, with the FOI in 1st year injectors increasing somewhat in the last decade.

Table 4.20: Force of infection for HIV by injecting duration and time, multiplicative and additive interaction models.

| Multiplicative | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year / inj. | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1980-1985 | 0.0142 | 0.0034 | 0.0049 | 0.0058 | 0.0046 | 0.0035 | 0.0001 |
| 1985-1990 | 0.0012 | 0.0003 | 0.0004 | 0.0005 | 0.0004 | 0.0003 | 0.0000 |
| 1990-1995 | 0.0018 | 0.0004 | 0.0006 | 0.0007 | 0.0006 | 0.0004 | 0.0000 |
| 1995-2000 | 0.0012 | 0.0003 | 0.0004 | 0.0005 | 0.0004 | 0.0003 | 0.0000 |
| 2000-2005 | 0.0049 | 0.0012 | 0.0017 | 0.0020 | 0.0016 | 0.0012 | 0.0000 |
| 2005-2010 | 0.0071 | 0.0007 | 0.0000 | 0.0000 | 0.0013 | 0.0009 | 0.0000 |
| 2010-2015 | 0.0120 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Additive | | | | | | | |
| Year / inj. | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.0092 | 0.0056 | 0.0068 | 0.0072 | 0.0070 | 0.0055 | 0.0051 |
| 1980-1985 | 0.0025 | -0.0011 | 0.0001 | 0.0005 | 0.0003 | -0.0012 | -0.0016 |
| 1985-1990 | 0.0033 | -0.0003 | 0.0009 | 0.0014 | 0.0011 | -0.0004 | -0.0007 |
| 1990-1995 | 0.0028 | -0.0008 | 0.0003 | 0.0008 | 0.0005 | -0.0009 | -0.0013 |
| 1995-2000 | 0.0045 | 0.0009 | 0.0021 | 0.0026 | 0.0023 | 0.0008 | 0.0005 |
| 2000-2005 | 0.0053 | 0.0010 | 0.0010 | -0.0010 | 0.0010 | -0.0004 | -0.0013 |
| 2005-2010 | 0.0047 | 0.0004 | 0.0006 | 0.0005 | -0.0002 | -0.0016 | -0.0016 |
| 2010-2015 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

### 4.2.4 Force of infection models for HCV with covariates

Model results for HCV in section 4.2.1 indicated that the multiplicative model was not substantially better than the additive model when fitted to

Table 4.21: Force of infection for HIV by injecting duration and time, multiplicative and additive interaction models, fitted to data from 2000 onwards.

| Multiplicative | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year / inj. | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1980-1985 | 0.0108 | 0.0000 | 0.0038 | 0.0027 | 0.0058 | 0.0000 | 0.0014 |
| 1985-1990 | 0.0042 | 0.0000 | 0.0015 | 0.0011 | 0.0023 | 0.0000 | 0.0005 |
| 1990-1995 | 0.0046 | 0.0000 | 0.0016 | 0.0012 | 0.0025 | 0.0000 | 0.0006 |
| 1995-2000 | 0.0004 | 0.0000 | 0.0001 | 0.0001 | 0.0002 | 0.0000 | 0.0001 |
| 2000-2005 | 0.0062 | 0.0000 | 0.0022 | 0.0016 | 0.0033 | 0.0000 | 0.0008 |
| 2005-2010 | 0.0075 | 0.0003 | 0.0000 | 0.0000 | 0.0016 | 0.0000 | 0.0000 |
| 2010-2015 | 0.0120 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Additive | | | | | | | |
| Year / inj. | <1 | 1-3 | 3-5 | 5-10 | 10-15 | 15-25 | >25 |
| pre 1980 | 0.0086 | 0.0004 | 0.0037 | 0.0033 | 0.0043 | 0.0034 | 0.0011 |
| 1980-1985 | 0.0050 | -0.0033 | 0.0001 | -0.0004 | 0.0006 | -0.0003 | -0.0026 |
| 1985-1990 | 0.0072 | -0.0011 | 0.0023 | 0.0018 | 0.0028 | 0.0019 | -0.0004 |
| 1990-1995 | 0.0056 | -0.0027 | 0.0007 | 0.0002 | 0.0012 | 0.0003 | -0.0020 |
| 1995-2000 | 0.0074 | -0.0009 | 0.0025 | 0.0020 | 0.0030 | 0.0021 | -0.0002 |
| 2000-2005 | 0.0094 | -0.0004 | 0.0017 | -0.0010 | 0.0020 | -0.0009 | -0.0019 |
| 2005-2010 | 0.0086 | -0.0010 | 0.0006 | -0.0004 | 0.0001 | -0.0022 | -0.0031 |
| 2010-2015 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

the data from 2000 onwards, although goodness of fit is overstated somewhat for the additive models, which allow negative FOI estimates. As full covariate data are only available from 2000 onwards, the additive model appears adequate for exploratory purposes, keeping in mind the limitations of this model. However, it must be borne in mind that the effects of covariates are also of course assumed to be additive, which does not necessarily follow from the above. Further, the potential for a negative FOI for some combinations of injecting duration, time and covariates becomes more likely as the complexity of the model increases, further overstating the goodness of fit of the additive model.

In section 4.1.5, it was seen that there were marked differences in prevalence according to region, which also varied over time. Regional effects were therefore investigated in a FOI model with additive effects, considering the main effects for injecting duration, time period and region, plus potential interactions between the three factors. As the most important difference in

risk according to injecting duration appeared to be the excess risk in 1st year injectors, simplified forms of interactions for 1st year vs. longer injecting durations were considered. All possible combinations were fitted, and the relative merits of different models compared via AIC scores.

The model with the lowest AIC score included interactions between region and time, a simplified interaction between region and 1st year vs. longer injecting duration, but did not require an interaction between injecting duration and time. It is worth noting that this model still includes a substantial number of interaction terms, and being based on AIC scores, may be over-complex (compared to model selection under the heavier penalty for complexity applied by the BIC).

Estimates of the FOI according to region and time period, for 1st year vs. 3-5 years injecting duration are shown in Figure 4.19. There is no interaction between injecting duration and time, hence the pattern of results for 1st year vs. 3-5 years injecting is identical. However, the relative effect of 1st year vs. longer injecting durations can vary by region, and there are marked differences: a particularly striking feature is that the well-known high-risk regions of NW and London have very high FOIs in the first year, but comparable FOIs to other regions subsequently. In fact, the majority of regional variation appears to be in 1st year risk.

Other covariates investigated in section 4.1.5 were examined within the GLM framework for FOI models, with all models including the regional effects and interactions specified above. Ever receiving works was associated with a 0.017 (95% CI 0.014, 0.019) increase in FOI, with little difference between injecting duration and survey period if interactions are considered.

Use of needle exchange, which is entered as a time-varying covariate that divides pre- and post-needle exchange exposure, was predicted to change the FOI by -0.013 (95% CI -0.017, -0.008). The reduction showed no significant difference according to injecting duration, but needle exchange use appeared to be less efficacious (actually conferring an increase in risk) in earlier time periods and only becoming effective from the year 2000 onwards.

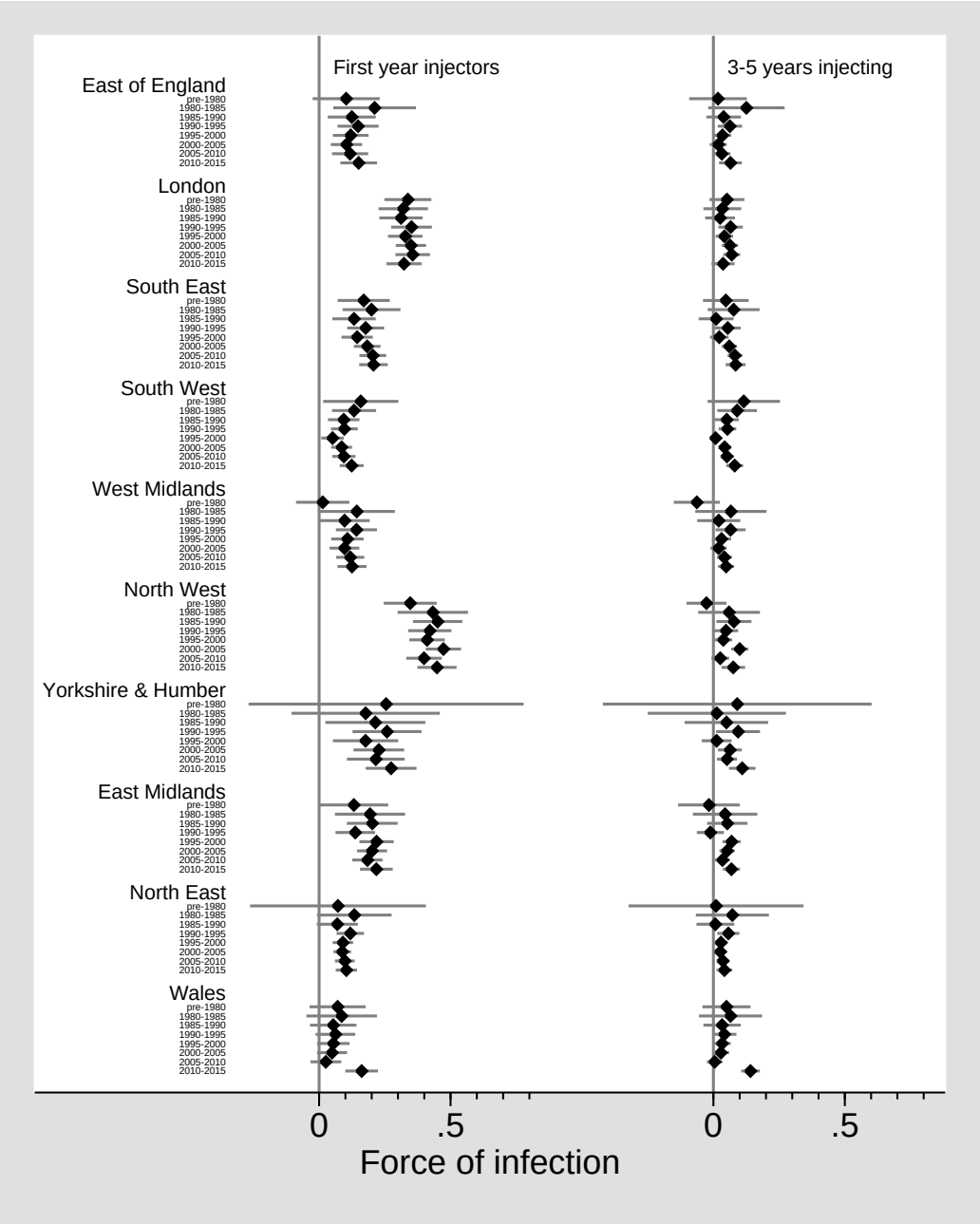


Figure 4.19: Regional difference in the force of infection for HCV, by time and 1st year vs. 3-5 years injecting.

### 4.2.5 Concluding remarks

This chapter has included a thorough investigation of patterns in BBV prevalence according to injecting duration, survey period and the key covariates available for analysis. Most of the questionnaire data on risk are only available from 2000 onwards, restricting the data available for analysis. In some cases this may simplify modelling, with interactions between injecting duration and time appearing to be less vital. It is not clear whether the earlier data may contain important information on changes in risk over time. However, if the main interest lies in current risk differences according to injecting duration and changes over recent times, this may not be particularly important.

Current status data may be modelled via standard generalised linear models for binomial data, such as logistic regression, or models that estimate the force of infection. The underlying data and information contained therein are of course the same, but with parameters related to the data in a different way. Nevertheless, the FOI may be of more direct interest in determining risk patterns: in the regional example in section 4.2.4, a striking regional pattern in 1st year injectors vs. longer injecting durations was revealed that may be less obvious from fitting logistic regression models. Similarly, by considering the effect of needle exchange as a time-varying covariate, its usage was shown to provide a general protective effect, but only in more recent years. Again, such patterns might be revealed by careful modelling in logistic regression models, but are harder to uncover. FOI models also model the relationship between time at risk and infection status more naturally, requiring fewer parameters in general; although as there is less information available from risk differences, they also tend to be estimated less precisely.

The downside of FOI models is that only the additive form of the model can be fitted within the GLM framework, whereas epidemiological applications tend to consider risk factors as multiplicative. This makes sense in general: if the overall risk of infection has changed over some time period, person $A$, who makes twice as many contacts with infectious individuals as person $B$, would naturally be expected to still have double the risk of infection, rather than a fixed additive difference. Of course, all of the outcomes considered here are infectious diseases, so it is difficult to predict in what way

time at risk and calendar time will combine due to transmission dynamics. Nevertheless, the GLM form appears to be a suitable tool for exploratory analysis of additional covariates for FOI models, similar to the suggestion of Ades and Nokes (1993) in their paper on age and time effects.

GLMs for the FOI use an unusual link for binary data, and are therefore not permitted in all statistical packages: models can be fitted in Stata using the `glm` command, although the R function `glm` will not allow such models to be fitted. An alternative here would be to use the complementary log-log link, which allows certain parametric forms for the FOI, such as the Weibull model. This may be appropriate given epidemiological knowledge of the general pattern of the FOI, but does of course impose certain shapes. In particular, the Weibull model implies an exponential change in the FOI, which if declining, will tend exponentially towards zero. Given the non-zero FOI at all injecting durations, this would not fully capture the relationships seen here.

In the next two chapters FOI models including individual frailties are considered, which are outside the scope of standard GLMs. In these analyses, multiplicative effects for calendar time and other covariates are assumed, and the effect of covariates on the FOI further examined on the basis of the findings presented in this chapter.

# Chapter 5

# Individual heterogeneity and models for multivariate data

Collecting samples from individuals in the population of interest can be expensive and time-consuming, but the serological testing itself is often relatively cheap and straightforward (moreover, given that one serological test is undertaken, *additional* testing is relatively easy). In such circumstances the same serological sample can be used to test for multiple infections (usually of a similar nature, such as childhood diseases) resulting in current status data for more than one infection. This can prove very useful from an epidemiological point of view, broadening the potential for analytical exploration.

The models considered in chapters 3 and 4 are based on the assumption that all individuals have the same risk of infection, given their age (or time exposed), calendar time and other covariate information. However, it is likely that individuals will differ according to some unmeasurable factor, particularly in the absence of covariate data, which in the context of infectious diseases is often given to be their propensity to make effective contacts, or otherwise have varying levels of exposure, susceptibility, or opportunity for infection. Such individual heterogeneity, or *frailty*, is of particular importance in the estimation and interpretation of the age-specific force of infection. In statistical models, frailty is expressed via a specified distribution and identified via the correlation that occurs in multivariate data when infections share a common route of transmission. Therefore multivariate data are central to the estimation of individual heterogeneity.

## 5.1 Frailty distributions

In a frailty model, the force of infection acting on an individual is the product of a basic rate $\lambda(t)$ and an individual-specific quantity $Z$ (see, for example, Aalen et al. (2008, p. 234)):

$$\lambda(t|Z) = Z\lambda(t).$$

Given $Z$, the probability of remaining susceptible up to time $t$ is given by

$$S(t|Z) = \exp(-ZA(t)),$$

where

$$A(t) = \int_0^t \lambda(u)\ du.$$

The function for the population proportion susceptible is found by integrating over the distribution of Z

$$S(t) = E[\exp(-ZA(t))]. \tag{5.1}$$

The Laplace transform of $Z$ is defined by

$$\mathcal{L}(c) = E(\exp(-cZ)), \tag{5.2}$$

and therefore

$$S(t) = \mathcal{L}(A(t)). \tag{5.3}$$

The population force of infection, $\lambda_p(t)$, may then be found by differentiating $-\log(S(t))$:

$$\lambda_p(t) = \lambda(t)\frac{-\mathcal{L}'(A(t))}{\mathcal{L}(A(t))}.$$

### 5.1.1 The gamma distribution for frailty

One of the simplest choices for the frailty distribution is the gamma distribution, for which the Laplace transform is easily derived. The probability

density of the gamma distribution is given as

$$f(x; k, \theta) = \frac{\theta^k}{\Gamma(k)} x^{k-1} \exp(-\theta x), \tag{5.4}$$

where $\theta$ is a rate parameter and $k$ a shape parameter. In order for the average FOI to be equal to the baseline FOI $\lambda(t)$, it is sensible to define the distribution with mean equal to 1, i.e., $\theta = k$. The frailty variance is thus given by $\delta = \frac{1}{\theta}$. Figure 5.1 shows various shapes that may be obtained under different values of $\delta$. With low values ($\delta < 0.1$) the frailty, $Z$, is narrowly distributed around 1; as $\delta$ increases this distribution spreads out, and with $\delta > 5$ the bulk of the distribution of $Z$ tends towards zero, but with a long tail stretching to higher values.

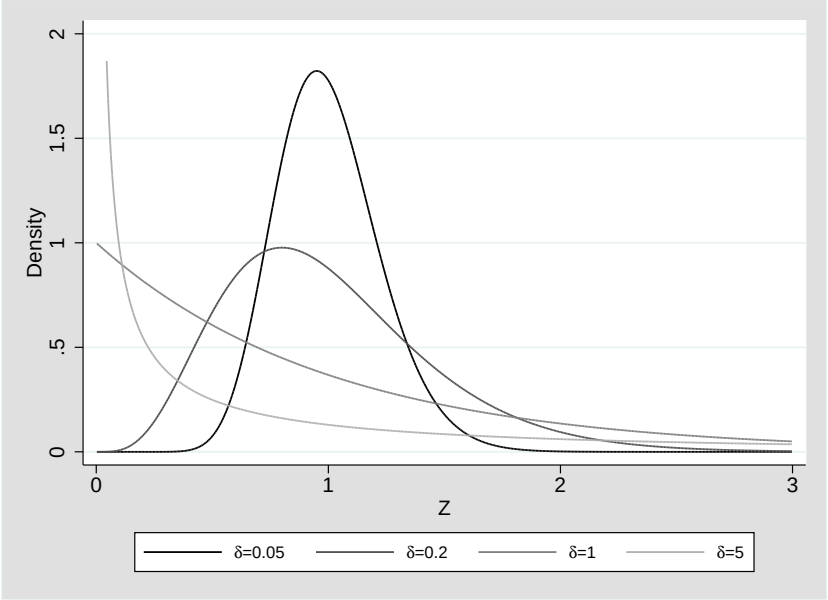


Figure 5.1: Density of gamma distributions with mean 1 and variance $\delta$=0.05, 0.2, 1 and 5.

The Laplace transform of the gamma distribution is obtained in a similar

way to its moment generating function:

$$\begin{aligned}\mathcal{L}(c) &= \int_0^\infty \exp(-cx)\frac{\theta^k}{\Gamma(k)}x^{k-1}\exp(-\theta x)\ dx \\ &= \frac{\theta^k}{\Gamma(k)}\int_0^\infty x^{k-1}\exp(-x(c+\theta))\ dx \\ &= \frac{\theta^k}{\Gamma(k)}\frac{\Gamma(k)}{(c+\theta)^k}\int_0^\infty \frac{(c+\theta)^k}{\Gamma(k)}x^{k-1}\exp(-x(c+\theta))\ dx \\ &= \frac{\theta^k}{(c+\theta)^k}.\end{aligned}$$

By the addition of terms to the numerator and denominator, an integral of another gamma distribution with rate parameter $c+\theta$ over the range $[0,\infty]$ is obtained, which is equal to 1. For gamma distributions with mean equal to 1 this can be written as:

$$\mathcal{L}(c) = \left(1+\frac{c}{\theta}\right)^{-\theta}. \tag{5.5}$$

Combining equations 5.3 and 5.5 yields the population survivor function

$$S(t) = \left(1+\frac{A(t)}{\theta}\right)^{-\theta} \tag{5.6}$$

and the population hazard rate

$$\lambda_p(t) = \frac{\lambda(t)}{1+\frac{A(t)}{\theta}}. \tag{5.7}$$

See, for instance Aalen et al. (2008, p. 236)). In some statistical analyses, ignoring sources of unexplained variation will not necessarily bias estimates (although standard errors may be underestimated), but when trying to estimate time-specific forces of infection the presence of individual heterogeneity, or frailty, can distort results. Those at higher risk will tend to experience the event in question earlier than others, while those that remain, who have a lower average frailty, will experience the event at a lower rate. This has the effect of pushing down the population force of infection over time, whereas the risk for each individual may actually be stable or increasing (see Aalen et al. (2008, Chapter 6)).

The difference between individual and population hazard is more than a philosophical point. Planning of public health interventions may depend on whether there is a genuine high risk of infection at the beginning of the time at risk compared to an *apparent* high initial risk due to heterogeneity. In the context of people who inject drugs, this might mean a shift in focus from quickly finding those that have recently initiated injecting to engage them in preventative measures, compared to increased targeting of high-risk individuals in a heterogeneous population.

Equation 5.7 shows the effect of frailty on the population hazard. When the frailty variance ($\delta = \frac{1}{\theta}$) is zero there is no difference between $\lambda_p(t)$ and $\lambda(t)$. However, as $\delta$ increases the denominator becomes greater than 1 and increases over time, decreasing the population hazard compared to that of the individual over time. Figure 5.2 shows the age-specific proportion susceptible under a constant FOI at different percentiles of the gamma distribution, under different frailty variances. When the frailty variance $\delta$ is high, a large proportion of the population have almost no risk of infection, whereas a fairly small group have a high risk and the proportion susceptible within this group declines rapidly with age.

The gamma distribution is often used for its mathematical convenience, but also has some nice properties. It can be explicitly differentiated any number of times (Aalen et al., 2008, p. 236) and heterogeneity remains constant in survivors; i.e., conditional on not having yet experienced the event, individual heterogeneity remains the same over time. This is an important feature and makes the gamma distribution the reference distribution against which other frailty distributions are compared, as other choices will result in increasing or decreasing heterogeneity in the remaining survivors over time.

### 5.1.2 The inverse Gaussian distribution

The inverse Gaussian distribution can also make for a useful choice of frailty distribution. Its probability density function is given as

$$f(x;\mu,\theta) = \sqrt{\frac{\theta}{2\pi x^3}} \exp\left[-x\frac{\theta}{2\mu^2} - \frac{\theta}{2x} + \frac{\theta}{\mu}\right]; \tag{5.8}$$

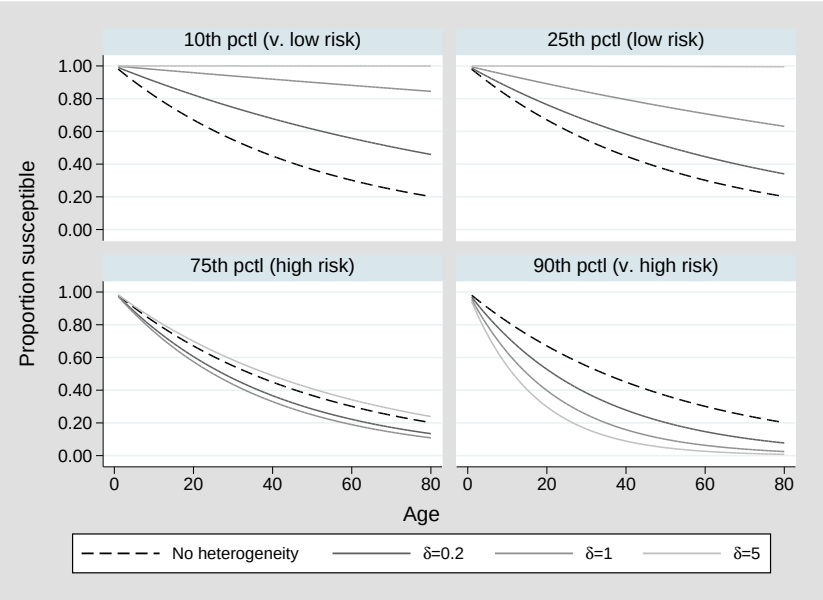


Figure 5.2: The age-specific proportion susceptible in a population with hazard=0.02 at the 10th, 25th, 75th and 90th percentile (pctl) points of different gamma frailties. Results are shown with no heterogeneity; and frailty variance $\delta$=0.2, 1 and 5.

and has mean $\mu$ and variance $\frac{\mu^3}{\theta}$. The Laplace transform is derived as follows:

$$\begin{aligned}\mathcal{L}(c) &= \int_0^\infty \exp(-cx)\sqrt{\frac{\theta}{2\pi x^3}}\exp\left[-x\frac{\theta}{2\mu^2}-\frac{\theta}{2x}+\frac{\theta}{\mu}\right]\,dx \\ &= \int_0^\infty \sqrt{\frac{\theta}{2\pi x^3}}\exp\left[-cx-x\frac{\theta}{2\mu^2}-\frac{\theta}{2x}+\frac{\theta}{\mu}\right]\,dx \\ &= \exp(-\Phi)\int_0^\infty \sqrt{\frac{\theta}{2\pi x^3}}\exp\left[-x\frac{\theta}{2\mu_c^2}-\frac{\theta}{2x}+\frac{\theta}{\mu_c}\right]\,dx,\end{aligned}$$

where $\mu_c^2 = \frac{\theta\mu^2}{\theta+2c\mu^2}$ and $\Phi = \frac{\theta\sqrt{\theta+2c\mu^2}}{\sqrt{\theta}\mu} - \frac{\theta}{\mu}$. The integral now takes the form of another inverse Gaussian pdf and is equal to 1. As interest is in frailty distributions with mean 1, the formula can be simplified to

$$\mathcal{L}(c) = \exp\left[\theta\left(1-\sqrt{1+\frac{2c}{\theta}}\right)\right]; \tag{5.9}$$

and therefore the population survivor function is

$$S(t) = \exp\left[\theta\left(1 - \sqrt{1 + \frac{2A(t)}{\theta}}\right)\right]. \tag{5.10}$$

The frailty variance is defined as $\delta = \frac{1}{\theta}$. Figure 5.3 shows the pdf of the inverse Gaussian distribution under various values of $\delta$. For small values of $\delta$ there is little difference in shape between the inverse Gaussian and the gamma distribution, but as shown in figure 5.1, as $\delta$ approaches 1 the mode of the gamma distribution is pushed towards zero, and for $\delta \geq 1$ there is no point of inflection. The inverse Gaussian however is "bell-shaped" (if potentially very skewed) for all values of $\delta$. In a practical sense, this would mean the difference between a population whose low-risk individuals had risk approaching zero (gamma distribution) compared to a population that may include very low-risk individuals but those with risk approaching zero were more rare (inverse Gaussian).

An important distinction between the two distributions is that under the inverse Gaussian distribution the heterogeneity of survivors declines over time; i.e., they become more homogeneous. If this appears to be the case, and it is believed that this declining heterogeneity is due to a selection effect (rather than changes in the heterogeneity in risk behaviour over time), then the inverse Gaussian distribution may be appropriate.

### 5.1.3 Frailty distribution families

The *power variance* function (PVF) family of distributions suggested by Hougaard (2000) provides a number of interesting possibilities for frailty distributions. PVF distributions are defined as those having the Laplace transform

$$\mathcal{L}(c; \rho, \nu, m) = \exp\left[-\rho\left(1 - \left(\frac{\nu}{\nu + c}\right)^m\right)\right]. \tag{5.11}$$

A number of distributions are special cases of the PVF family. If $\rho \to \infty$ and $m \to 0$ in such a way than $\rho m \to \eta$ then the Laplace transform approaches that of the gamma distribution with scale parameter $\nu$ and shape parameter $\eta$ (Aalen et al., 2008, p. 238). The inverse Gaussian distribution also belongs to the PVF family ($m = -\frac{1}{2}$, $\rho < 0$). Another example is the

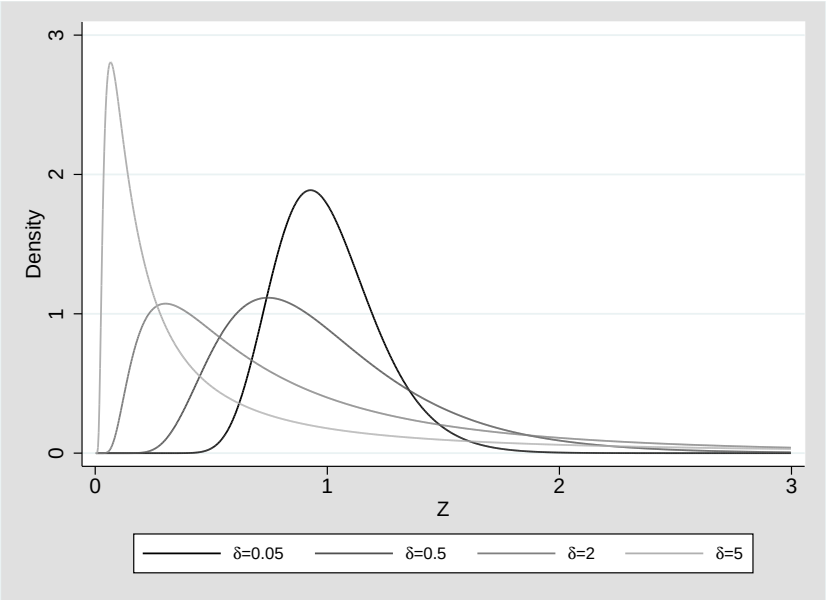


Figure 5.3: Density of inverse Gaussian distributions with mean 1 and variance $\delta$=0.05, 0.2, 1 and 5.

compound Poisson distribution. This includes a probability atom at zero, corresponding to a group with zero frailty, or a nonsusceptible group. This might arise in the context of people who inject drugs if some people never share needles or equipment, or otherwise are somehow never at risk of infection. The distribution is the sum of independent gamma variables, and hence corresponds to a *cumulative damage* model, whereby repeated "insults" increase an individual's susceptibility. This is somewhat implausible in this context, as the risk of infection is unlikely to be cumulative, although Farrington et al. (2012) describe an alternative family (the *Addams* family!) from which other discrete frailty distributions can be derived, which might be appropriate for the number of sexual partners in the modelling of sexually transmitted disease. In the context of injecting drug use, this might be the size of social group in which needle sharing might arise. The possibility of a zero-risk group is potentially interesting, but limited in that the relative frailty variance must be non-decreasing (Farrington et al., 2012).

Frailty distributions may be characterised by their relative frailty variance (RFV), the heterogeneity in survivors at time $t$, defined by Farrington et al. (2012) as:

$$RFV(t) = \text{var}(U_t|T > t) = \frac{\text{var}(U|T > t)}{E(U|T > t)^2}, \tag{5.12}$$

where $U$ is a non-negative random variable corresponding to individual frailty, and $U_t = \frac{U}{E(U|T>t)}$ is the relative frailty, conditional on surviving to time $t$. The gamma distribution is the only choice resulting in a constant RFV, whereas the inverse Gaussian (and many others) result in a decreasing RFV. The RFV is discussed further in section 5.2.3.

### 5.1.4 Time-varying frailty

In addition to the selection effect induced by particular frailty distributions, heterogeneity in risk may genuinely be decreasing (or increasing) over time. Farrington et al. (2013) describe a selection of time-varying frailty models with one or two components. The single component model is:

$$Z(t) = 1 + (Z - 1)h(t), \tag{5.13}$$

where $Z$ is a time invariant frailty of unit mean (for instance, a gamma distribution) and $0 \leq h(t) \leq 1$ is a deterministic function, such as $h(t) = \exp(-\rho t)$, which results in an exponential decay. This model was adapted slightly such that instead of heterogeneity declining to zero, a two-parameter function is used such that frailty declines toward an asymptote:

$$h(t) = \frac{\exp(-\rho t) + \exp(q)}{1 + \exp(q)}. \tag{5.14}$$

Thus $\rho$ controls the rate of decline of the relative frailty variance, and the asymptote towards which it declines is given by $\frac{\exp(q)}{1+exp(q)}$, the exponents being required to ensure the function is bounded within $[0, 1]$ and $q$ taking any real value ($q \to -\infty$ yields a zero asymptote and $q \to \infty$ no decline in relative frailty variance).

The survivor function with such a time-varying frailty distribution is

$$S(t|Z) = \exp\left[-\int_0^t \lambda(u)\left(1 + (Z-1)\left(\frac{\exp(-\rho u) + \exp(q)}{1+\exp(q)}\right)\right)\, du\right]. \tag{5.15}$$

Integrating over the frailty distribution to obtain the unconditional survivor function is not as straightforward as for the gamma and inverse Gaussian distributions. To proceed, the function is split into a "baseline" part of the expression and a "frailty" part,

$$S(t|Z) = \exp\left[-\int_0^t \lambda(u)\left(1 - \frac{\exp(-\rho u)}{1+\exp(q)} - \frac{\exp(q)}{1+\exp(q)}\right)\, du\right] \\ \exp\left[-\int_0^t Z\lambda(u)\left(\frac{\exp(-\rho u)}{1+\exp(q)} + \frac{\exp(q)}{1+\exp(q)}\right)\, du\right], \tag{5.16}$$

with the marginal form of the latter being obtained by its Laplace transform in the usual way. For a piecewise constant FOI the derivation is as follows, using a slight change in notation to chapter 3. With $k$ time bands and cutpoints $c_0 = 0, c_1, c_2 ... c_k = \infty$, the FOI is given by

$$\lambda(t) = \begin{cases} \lambda_1, & \text{if } t \leq c_1; \\ \lambda_2, & \text{if } t > c_1 \text{ and } t \leq c_2; \\ & ... \\ \lambda_k, & \text{if } t > c_{k-1}, \end{cases}$$

and the cumulative hazard $A(t)$ as

$$A(t) = \int_0^t \lambda(u)\, du = \sum_{i=1}^{k} \lambda_i \max(0, \min(t - c_{i-1}, c_i - c_{i-1})).$$

Defining the time spent in band $i$ as $t_i = \max(0, \min(t - c_{i-1}, c_i - c_{i-1}))$ and $v_i = c_{i-1} + t_i$, the terms in equation 5.16 requiring evaluation are

$$\int_0^t \lambda(u)\, du = \sum_{i=1}^{k} \lambda_i t_i,$$

$$\int_0^t \lambda(u)\exp(-\rho u)\ du = \sum_{i=1}^{k} -\frac{\lambda_i \exp(-\rho u)}{\rho}\Big|_{c_{i-1}}^{v_i}$$
$$= \sum_{i=1}^{k} \frac{\lambda_i(\exp(-\rho c_{i-1}) - \exp(-\rho v_i))}{\rho}$$

and the Laplace transform for the frailty part of equation 5.16, with $Z$ having a gamma distribution with mean 1 and variance $\frac{1}{\theta}$

$$\mathcal{L}(c) = \left(1 + \frac{c}{\theta}\right)^{-\theta}.$$

Inserting these expressions into equation 5.16 yields the unconditional survivor function

$$S(t) =$$
$$\exp\left[\sum_{i=1}^{k} -\lambda_i t_i \left(1 - \frac{\exp(q)}{1+\exp(q)}\right) - \lambda_i \left(\frac{\exp(-\rho v_i) - \exp(-\rho c_{i-1})}{\rho(1+\exp(q))}\right)\right]$$
$$\left[1 + \sum_{i=1}^{k} \frac{\lambda_i(\exp(-\rho c_{i-1}) - \exp(-\rho v_i))}{\rho\theta(1+\exp(q))} + \frac{\lambda_i t_i \exp(q)}{\theta(1+\exp(q))}\right]^{-\theta}. \quad (5.17)$$

More complex structures are of course possible. For instance, frailty may be considered to change from one level to another throughout the at-risk period, with a frailty distribution representing variability in rates of infection during childhood and a separate frailty distribution for adulthood. Farrington et al. (2013) proposed a model with two gamma components, $Z_1$ and $Z_2$, with

$$Z(t) = (1 + (Z_1 - 1)h(t))Z_2. \quad (5.18)$$

Obtaining the unconditional survivor function for the two component model requires integration over the distributions of $Z_1$ and $Z_2$, and so is not as straightforward as the Laplace transforms considered so far.

The 2-component model could conceivably be applied to people who inject drugs, with a component for new initiates ("childhood") and for experienced users. The difficulty here is the comparatively short initiation period. A more general issue with the model is that the two components are independent, whereas high or low risk is likely to persist to some extent

in experienced users. The model could potentially be extended to allow for correlations between the components, although this is not pursued further and the single component model in equation 5.13 used to capture the likely decline in heterogeneity over time.

A general disadvantage of this time-varying frailty model is that the frailty has limited support, with a minimum value of $1 - h(t)$ for equation 5.13. This implies a lower bound on the force of infection in the population, which may not be epidemiologically plausible if some individuals are at very low risk in comparison to the rest of the population. For people who inject drugs the existence of a low-risk group may be unlikely, given the high prevalence of blood-borne viruses and monotonic increase with time at risk. Another criticism of the distribution in equation 5.13 is that the frailty variance declines at a constant rate towards an asymptotic value, which may be slightly restrictive. However, in practice there is unlikely to be sufficient information to estimate such changes over time.

An alternative to the model form in 5.18 is to use functions for $h(t)$ based on powers, which do not suffer from the same issues of limited support, but are difficult to work with algebraically (Enki et al., 2014).

### 5.1.5 Piecewise constant frailties

An alternative to time-varying frailties based on deterministic functions of the form considered in section 5.1.4 is the piecewise gamma frailty proposed by Paik et al. (1994). Their definition is for a nested structure in which

$$Z_i = Z + \varepsilon_i \tag{5.19}$$

in age band $i$, where $Z$ has a gamma distribution for the overall frailty of an individual and $\varepsilon_i$ have independent gamma distributions corresponding to age-specific fluctuations in band $i$; and the mean of the two distributions sums to 1 ($\mu_1 + \mu_2 = 1$). This form is particularly easy to work with in conjunction with piecewise constant hazards: if the age bands of the baseline FOI and frailties coincide, the population survivor function is given by

$$S(t) = \exp\left[-\frac{\mu_1}{\gamma}\log(1+\gamma A(t)) - \sum_{i=1}^{k}\frac{\mu_2}{\gamma_i}\log(1+\gamma_i\lambda_i t_i)\right], \tag{5.20}$$

where $\mu_1$ and $\mu_2$ are the means of the overall and the component frailties respectively, $\gamma$ and $\gamma_i$ their scale parameters, $A(t)$ the usual cumulative hazard, $\lambda_i$ the baseline hazard in age band $i$ and $t_i$ the exposure time in age band $i$. Rather than multiplying components raised to the power of $\frac{\mu_1}{\gamma}$ and $\frac{\mu_2}{\gamma_i}$, as defined in Paik et al. (1994), the expression above sums log terms then exponentiates, which may be more numerically stable as the number of piecewise constant terms grows.

The interpretation of the Paik model is a little unusual, potentially limiting its practical application. Under the hierarchical structure individuals have an overall frailty and an age-specific frailty. These age-specific components are independent, so individuals may be high risk in one age band and low risk in another; the component part of the individual frailty is "reset" in each age band, therefore there is no inter-age band selection effect in survivors. These properties may or may not be desirable, but it seems intuitively more likely that some correlation between frailties would persist within individuals over time. In particular, this structure may not be viewed as a piecewise alternative to equation 5.13 or 5.18, in which individuals have a persistent but declining frailty or transition between two values. Farrington et al. (2012) propose a multiplicative alternative that allows compounding of frailties; this would allow dependence to be incorporated by allowing frailty components to persist across age bands, with the multiplicative relationship preserving the unit mean of the overall frailty distribution.

## 5.2 Multivariate models

With univariate data, the only information available to estimate the parameters of a given frailty distribution is via distortion in the population survivor function (as observed in Figure 5.2). This implies that the shape of the unconditional survivor function has to be assumed, or rather, the shape of the underlying hazard function. This might be reasonable in some settings, where biological or epidemiological considerations strongly suggest a particular parametric form (for instance, the Weibull distribution), but in general it would appear unsafe to make strong assumptions about the age-specific risk of infection. In particular, under the piecewise constant models that have been the focus so far there would be no information with which to es-

timate the frailty distribution, provided that the cut-points are sufficient to capture the general shape of any plausible hazard function.

Multivariate models address this problem by providing information on the fraily via the correlation between infections. Infections that share a transmission route will naturally be correlated, although correlation may occur for other reasons, such as variation in individuals' general biological susceptibility to infection. Assuming for now that transmission routes are identical, the degree of correlation determines the extent of individual heterogeneity. An individual that has a higher risk of infection with $A$ due to certain risk behaviours or susceptibility will also have a higher risk of infection with $B$ if the route of transmission is the same, inducing an association between $A$ and $B$. There may still be some interplay between the assumed frailty distribution and function for the FOI; for instance, models incorporating age and time effects, or covariates, might allow distortions from the population survivor function to "feed into" the estimated frailty; but in general if the baseline FOI function is sufficiently flexible then the estimated frailty should give a true indication of the extent of individual heterogeneity.

### 5.2.1 Implementation of simple frailty models

Infection status probabilities are denoted as follows: let $\pi_{00}(t)$ be the probability that neither infection has occurred by time $t$, $\pi_{10}(t)$ the probability that infection 1 has occurred by time $t$ but infection 2 has not, $\pi_{01}(t)$ the probability that infection 2 has occurred by time $t$ but infection 1 has not, and $\pi_{11}(t)$ the probability that both infections have occurred by time $t$. Then the survival function for remaining free from both infections may be derived as in equation 3.13. Assuming for now that the risks of infection 1 and 2 are independent gives

$$\begin{aligned}
\pi_{00}(t) &= P(T_1 > t)P(T_2 > t) \\
&= \exp\left(-\int_0^t h_1(u)\ du\right) \exp\left(-\int_0^t h_2(u)\ du\right) \\
&= \exp\left(-\int_0^t h_1(u)\ du - \int_0^t h_2(u)\ du\right),
\end{aligned}$$

where $T_i$ is the time infection $i$ occurs and $h_i(t)$ is the hazard for infection $i$. It is simplest to then derive the probabilities for a single infection $\pi_{10}(t)$ and $\pi_{01}(t)$ in terms of the marginal probability of remaining free from the *other* infection, minus the probability of *neither* infection. The probability of both infections occurring is then the remaining probability. With $A_i(t) = \int_0^t h_i(u)du$, the cumulative hazard for infection $i$,

$$\begin{aligned}
\pi_{00}(t) &= \exp\left(-A_1(t) - A_2(t)\right) \\
\pi_{10}(t) &= \exp\left(-A_2(t)\right) - \pi_{00}(t) \\
\pi_{01}(t) &= \exp\left(-A_1(t)\right) - \pi_{00}(t) \\
\pi_{11}(t) &= 1 - \pi_{00}(t) - \pi_{10}(t) - \pi_{01}(t).
\end{aligned} \tag{5.21}$$

The log-likelihood for paired infection data then takes the product multinomial form

$$\sum_t \sum_{i,j=0}^{1} n_{ijt} \log(\pi_{ij}(t)),$$

where $n_{ijt}$ is the number of individuals with disease status $(i, j)$ at time $t$. In the case that some individuals are missing status data for one infection, these individuals contribute to the likelihood via marginal terms, e.g.,

$$\sum_t \sum_{i=0}^{1} n_{i.t} \log(\pi_{i0}(t) + \pi_{i1}(t))$$

for data on infection 1 only.

The formulae outlined above are not particularly useful in themselves, as no further information is gained from analysing the joint data compared to fitting separate models for each infection, but the equations form the basis of the multivariate frailty models. For a gamma distributed frailty, using the expression $S(t|Z) = \exp\left(-ZA(t)\right)$, then integrating out $Z$ as in equation

5.6, leads to

$$\begin{aligned}
\pi_{00}(t) &= \left(1 + \frac{A_1(t) + A_2(t)}{\theta}\right)^{-\theta} \\
\pi_{10}(t) &= \left(1 + \frac{A_2(t)}{\theta}\right)^{-\theta} - \pi_{00}(t) \\
\pi_{01}(t) &= \left(1 + \frac{A_1(t)}{\theta}\right)^{-\theta} - \pi_{00}(t) \\
\pi_{11}(t) &= 1 - \pi_{00}(t) - \pi_{10}(t) - \pi_{01}(t)
\end{aligned} \tag{5.22}$$

The equations for the inverse Gaussian distribution may be derived in a similar way. These models may readily be extended to an age- and time-specific force of infection by substituting cumulative force of infection functions $A(a,t) = \int_0^t h(u, t-a+u)du$ for infections 1 and 2; and indeed any of the general forms of covariate model described in section 3.2.5. The likelihood then follows the same multinomial form, but is of course indexed by age, time, and any other covariates. As with the univariate models, the observed probabilities of infection may be related to true infection status if information is available on the sensitivity and specificity of the tests. The tests for the infections considered in this thesis have imperfect sensitivity but near-100% specificity, so only sensitivity need be considered. This results in the following formulae:

$$\begin{aligned}
p_{00} &= \pi_{00} + (1-S_1)\pi_{10} + (1-S_2)\pi_{01} + (1-S_1)(1-S_2)\pi_{11} \\
p_{10} &= \pi_{10}S_1 + S_1(1-S_2)\pi_{11} \\
p_{01} &= \pi_{01}S_2 + S_2(1-S_1)\pi_{11} \\
p_{11} &= \pi_{11}S_1S_2,
\end{aligned} \tag{5.23}$$

where $p_{ij}$ are expected proportions observed for infection 1 and 2, $\pi_{ij}$ the true probabilities, and $S_1$ and $S_2$ the sensitivity of tests for infection 1 and 2 respectively. This approach is based on the assumption that the sensitivity of the two tests is independent and the sensitivity of one test is not altered by positive infection status for the other test.

Model fitting will generally require bespoke code to maximise the likelihood under frailty models. Extending such models to include covariates in piecewise constant models with additive effects is relatively simple. Covari-

ates are simply added to the linear predictor in the GLM form of the model, with a parameter representing the additive difference in the force of infection, multiplied by the exposure time at that covariate level, being added to the cumulative FOI (see equations 3.22 and 3.23). For multiplicative models the relationship between covariates, exposure time and the cumulative hazard is more difficult to specify in a general form. For instance, for a model with piecewise age-specific FOI $\exp(\mu_i)$ for age band $i$ and proportional hazards for a covariate $x$ with hazard ratio $\exp(\beta)$ the survivor function is given by

$$\begin{aligned} S(a,x) = \exp[-(&\exp(\mu_1 + \beta x)A_1(a) + \exp(\mu_2 + \beta x)A_2(a) + \\ &... + \exp(\mu_k + \beta x)A_k(a)]. \end{aligned} \tag{5.24}$$

It is straightforward to extend the model in equation 5.24 to incorporate additional covariates, interactions with age, time and so on, but will require adapting the code used to run the model (in contrast to additive models, which will only require specification of a different design matrix). Therefore testing alternative model forms can be time-consuming.

An alternative approach is to create multiple data rows for each observation, where each row consists of a particular age, time and covariate pattern in the individual's exposure history, and the exposure time at this covariate combination. With indicator variables for a particular covariate combination, the FOI is the exponentiated sum of the relevant parameters, which is multiplied by the exposure time. For instance, in equation 5.24 there is a separate row in the data corresponding to each of the terms $\exp(\mu_i + \beta x)A_i(a)$. These contributions to the cumulative hazard are then summed across each individual's history, and it is far easier to derive general-purpose model code to calculate each such contribution to the cumulative hazard (model code is provided in appendix section 9.2.2). Table 5.1 shows an example of such data, which are similar in form to that used for time-varying covariates in survival analysis.

This formulation requires larger datasets and increased computation time (the summation across rows and columns must be performed at each iteration of the numerical search procedure), but the advantage is generalisable model code.

The time-varying frailty models require some care in implementation,

Table 5.1: Example data for split exposure periods, with three age bands $(a_1, a_2, a_3)$ and three time bands $(t_2, t_3$; baseline period is 1), a covariate $x$ that can modify the hazard in the first age period and the exposure time $t$ for individual $i$ at exposure period $j$.

| $i$ | $j$ | $a_1$ | $a_2$ | $a_3$ | $t_2$ | $t_3$ | $x$ | $t$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 3 |
| 1 | 2 | 1 | 0 | 0 | 1 | 0 | 1 | 2 |
| 1 | 3 | 0 | 1 | 0 | 1 | 0 | 0 | 7 |
| 1 | 4 | 0 | 1 | 0 | 0 | 1 | 0 | 3 |
| 1 | 5 | 0 | 0 | 1 | 0 | 1 | 0 | 4 |
| 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 5 |
| 2 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

as unlike the simple frailty models, which consist of the cumulative hazard within an expression for the Laplace transform, time-varying frailty models include a number of terms involving cumulative hazards and different parameters. In particular, the piecewise constant model defined in equation 5.17 includes cumulative sums of the exposure times and FOI up to each age band, which for models that include both age and time effects becomes complicated. Again, splitting observations into exposure periods simplifies things somewhat; with exposure periods sorted in their temporal sequence the necessary sums are easier to calculate. Model code is provided in appendix section 9.2.3.

### 5.2.2 Separable mixing and shared parameters

It is important to remember when modelling data on infectious diseases that changes in the FOI according to age (or exposure time) and calendar time are the result of changes in the frequency and patterns of contact between infected and susceptible individuals. This may be accounted for in infectious disease models via the formulation of a *contact matrix*, representing the relative frequencies of effective contacts ("mixing") between individuals of age group $i$ and age group $j$ (see for instance, Farrington et al. (2001)). A specific form of contact frequencies is *separable mixing*, in which the contacts of age $i$ are distributed according to the activity level in each age group $j$. If the contact functions for each infection are proportional, then the age-

specific FOI for the two infections must also be proportional (Farrington et al., 2001). Specifying the age-specific FOI for infection 1 as $\lambda_{a1} = \lambda_a$, if the age-specific FOI for infection 2 is proportional to that for infection 1, it is given by $\lambda_{a2} = c\lambda_a$, where $c$ is a constant. Likelihood ratio tests or other model fit statistics may then be used to test the null hypothesis that the FOIs are proportional. If this null hypothesis is rejected, then the model with non-proportional FOIs is preferable.

This test may also be extended to the case where there is individual heterogeneity; in fact, when heterogeneity is present but not allowed for in the analysis, the test may reject the separable mixing model even if this model is sufficient. If separable mixing is tested for while incorporating frailty appropriately however, the test is then valid. For instance, the multinomial proportions for infection status under the gamma frailty model in equation 5.22 would be:

$$\begin{aligned}
\pi_{00}(t) &= \left(1 + \frac{A_1(t) + cA_1(t)}{\theta}\right)^{-\theta} \\
\pi_{10}(t) &= \left(1 + \frac{cA_1(t)}{\theta}\right)^{-\theta} - \pi_{00}(t) \\
\pi_{01}(t) &= \left(1 + \frac{A_1(t)}{\theta}\right)^{-\theta} - \pi_{00}(t) \\
\pi_{11}(t) &= 1 - \pi_{00}(t) - \pi_{10}(t) - \pi_{01}(t).
\end{aligned} \tag{5.25}$$

Calendar time may also be considered. HCV, HBV and HIV are likely to differ in temporal effects due to improvements in HBV vaccination and HIV treatment (which decreases infectivity), while until relatively recently there has been no such direct intervention on infection risk for HCV. In this case, the temporal effects for two infections might differ, while having proportional injecting duration (age) effects. In this case, if the model is specified in terms of injecting duration-specific hazard ratios (HR) as in equation 3.19 then the HRs would be the same for the two infections.

In general, proportionality tests may be applied for age, time or any other other covariate, to assess whether the change in risk according to certain factors is proportional for the two infections. The interpretation of the test as a test for separable mixing may become somewhat muddied for more complex models in which some factors are shared and some are not. In par-

ticular, temporal changes due to HBV vaccination and HIV treatment may not have a proportional effect at different injecting durations, and therefore the contact matrix may effectively change over time. Nevertheless, tests of proportionality for different factors may yet provide insights into shared elements of risk behaviour.

### 5.2.3 Associations between infections and further investigation of heterogeneity

The correlation between infections provides information on shared frailty (Farrington et al., 2001), and therefore investigation of the dependence structure can provide an indication of suitable frailty distributions. Archimedean copulas provide a framework for doing this, in which a bivariate distribution on the unit square is specified in terms of its marginal distributions, in this context the marginal survivor functions $S_1(t)$ and $S_2(t)$ for infections 1 and 2 respectively, and an associated dependence function $C$ (Genest and Rivest, 1993). This allows the dependence structure to be investigated independently of the marginal effects.

The Clayton copula is particularly useful for bivariate survival data (Clayton, 1978). The resulting cross-ratio function (CRF) can be written as

$$\theta^*(t_1, t_2) = \frac{S(t_1, t_2)\mathrm{D}_1\mathrm{D}_2 S(t_1, t_2)}{[\mathrm{D}_1 S(t_1, t_2)][\mathrm{D}_2 S(t_1, t_2)]}, \tag{5.26}$$

where $\mathrm{D}_j$ denotes the derivative operator $\delta/\delta t_j$. The CRF may be interpreted as the ratio of the hazard rates for event 1 given event 2 has, or has not yet, occurred (and vice versa) (Oakes, 1989). A key feature of this measure is its frailty interpretation, as $\theta^*(t_1, t_2)$ depends only on $(t_1, t_2)$ through $S(t_1, t_2)$ (Oakes, 1989).

Clayton (1978) provides an alternative derivation of this model based on the Cox model (Cox, 1972) and frailty with a gamma distribution, which has constant RFV. Alternative models would therefore not result in constant $\theta^*(t_1, t_2)$ when the underlying RFV is constant and would therefore not be suitable for the purpose of choosing an appropriate frailty distribution. Plots of the CRF can thus be used to suggest an appropriate frailty distribution, with the gamma distribution suggested by a constant CRF, and decreasing

or increasing CRFs suggesting alternative distributions.

For current status data, the joint survivor function $S(t_1, t_2)$ is not observable and the CRF cannot be evaluated. Unkel and Farrington (2012) proposed a measure of association that tracks the RFV over time in bivariate current status data, which can help to choose an appropriate frailty distribution. This measure is defined as the value $\Phi$ solving the implicit equation

$$\left(p_1(t)^{1-\exp(\Phi)} + p_2(t)^{1-\exp(\Phi)} - 1\right)^{1/[1-\exp(\Phi)]} = p_{00}(t), \tag{5.27}$$

where $p_{00}(t)$ is the proportion of individuals susceptible to both infections at time $t$ and $p_1(t)$ and $p_2(t)$ the proportion susceptible to infection 1 and 2. This approach is preferable to other measures of association, such as the odds ratio, which can increase over time even if the RFV is not increasing (Unkel and Farrington, 2012).

Examination of plots of $\Phi$ against time at risk (injecting duration in this context) can help to understand whether the RFV is constant or changes over time, with a common pattern in epidemiology being a decreasing RFV. Such plots may also be constructed according to subgroups based on covariate information, such as gender and calendar time, to examine whether the evolution of the RFV differs. If important differences are revealed, stratified analysis may then be undertaken to examine differences in heterogeneity between the subgroups. Stratification also allows the baseline FOI and any other covariates to differ between subgroups, which may or may not be desirable. An alternative approach would be to allow some shared parameter values across strata in the structural part of the model, but a stratified frailty distribution; for instance, different frailty variances for males and females, a fixed HR for females compared to males and common effects for calendar time and injecting duration.

The FOI for a stratified frailty variance model with age-specific FOI $\lambda_i(t)$ in subgroup $i$ and frailty variance $Z_i$ in subgroup $i$ is therefore

$$\lambda_i(t|Z_i) = Z_i\lambda_i(t). \tag{5.28}$$

The marginal survivor functions are obtained as in section 5.1 for any of

the frailty distributions considered so far, with the $\delta_i$ parameters for frailty variance (and the rate and asymptote parameters $\rho$ and $q$ for the TVF model) being estimated separately for each subgroup $i$ of the data. For instance, under a gamma frailty the marginal survivor function with FOI and frailty variances as in equation 5.28 becomes

$$S_i(t) = \left(1 + \frac{A_i(t)}{\theta_i}\right)^{-\theta_i}. \tag{5.29}$$

This model is readily extended to include calendar time and any other covariates, and may include a mixture of parameters that are specific to each subgroup, common across subgroups, or based on proportional FOIs across subgroups. If differences in frailty variance are being examined across subgroups, then the baseline FOI should also vary across these subgroups, at least allowing for differences under a proportional hazards assumption (for example a FOI of $\lambda(t)$ in males and $c\lambda(t)$ in females, where $c$ is a constant). As ever, the frailty variance is being estimated conditional on the baseline FOI structure, so some care is required to incorporate relevant sources of measured heterogeneity, which might include interactions between the subgroup variable with age or time.

Further extensions might allow some aspects of the more complex models, such as the TVF model, to vary but not others. For instance, stratified $\delta$ but a common shape for the decline in frailty variance or asymptote parameters. The frailty variance might also be allowed to vary according to more than one covariate factor. However, further complexity is not considered subsequently, and only the basic stratified frailty variance models are considered.

# Chapter 6

# Bivariate models for blood-borne viruses in people who inject drugs

In this chapter the methods described in chapter 5 are applied to the UAM data on people who inject drugs (PWID). Chapter 4 focussed on two main areas: firstly, how the risk of infection for HCV, HBV and HIV changed with injecting duration, over calendar time, and whether injecting duration-specific risk changed over time (interactions). The second aim was to examine a broad range of covariates using demographic and risk factor data collected by the UAM survey that might be of potential interest for modelling. In this chapter the aim is to conduct further modelling of covariates within a force of infection (FOI) model while incorporating a frailty distribution that represents individual heterogeneity. The use of different distributions that reflect different types of individual heterogeneity is investigated, in order to understand how risk evolves over the course of injecting career.

The combination of covariate models and frailty distributions allows the interplay between these two factors to be explored. One obvious question is whether the frailty variance is decreased by the addition of covariate information; as the frailty distribution accounts for unmeasured heterogeneity, accounting for some of the differences in risk should of course decrease the level of residual variation. Secondly, the frailty distribution itself might be altered by the inclusion of covariates, especially where the distribution of

covariate levels changes with exposure or calendar time, or there are interactions or time-varying covariates. Including covariates may therefore lead to different patterns of residual variation with increasing time at risk. The inclusion of individual heterogeneity will also answer the more common question considered in survival analysis and analysis of current status data: whether a FOI/hazard that apparently decreases with time at risk is due to selection effects induced by heterogeneity.

Finally, individual heterogeneity may vary according to different risk groups, such as by region or gender. Such differences can be explored via stratified analysis. Alternatively, the problem may be framed within a covariate model with group-specific frailty variances but some covariate effects common to the different groups. This approach may be preferable to fitting separate models to each stratum, which might be inefficient and lead to estimation problems if the data are too finely subdivided.

The chapter is organised as follows. Firstly, patterns of relative frailty variance are examined via the measure of association $\Phi$ described in section 5.2.3 to determine appropriate functional forms for the frailty distribution. Multiplicative models are then used to examine the injecting duration- and time-specific force of infection, with and without interactions, under a range of frailty distributions. The following section then focusses on the inclusion of covariates: demographic information on region, age and gender, and risk factor information on ever received works (needles, syringes etc.) from another person, ever use and age at first use of needle exchange, ever/age first imprisoned, and men who have sex with men. Finally, stratified frailty variances are implemented in order to explore whether certain subgroups exhibit greater levels of individual heterogeneity.

The focus of the analysis is on the bivariate HCV-HBV data, with HCV being the the key outcome of interest and the correlation with HBV providing information on individual variability. Data are far more limited for HIV, although the key models are examined for HCV-HIV and HBV-HIV pairings, with a particular view as to how the HIV data can inform more complex structures for the correlation between infections. The HBV-HIV pair is therefore also of particular interest, as these infections have a wholly shared transmission route, namely both injecting and sexual contacts. All analyses are based on data from 2000 onwards that have complete informa-

tion on the key covariates.

## 6.1 Measures of bivariate association: heterogeneity in risk of BBV infection in people who inject drugs

The measure of association $\Phi$ described in section 5.2.3 tracks the relative frailty variance (RFV) over exposure time, showing how frailty evolves in survivors and suggesting an appropriate structure for the frailty distribution (Unkel and Farrington, 2012). Figure 6.1 shows $\Phi$ by injecting duration for the three infection pairs, HCV-HBV, HCV-HIV and HBV-HIV. For all three pairings there is a general decline in RFV with injecting duration. For HCV-HBV, $\Phi$ drops sharply over the first three years of injecting, before declining more slowly subsequently, although still significantly above zero (which would indicate no heterogeneity). A similar pattern is observed for HCV-HIV, although HIV infections are sparse and estimates more variable across injecting duration, with wide confidence intervals. For HBV-HIV there is also some decline, but more slowly, and heterogeneity is greater than the other pairings, although again, data are sparse.

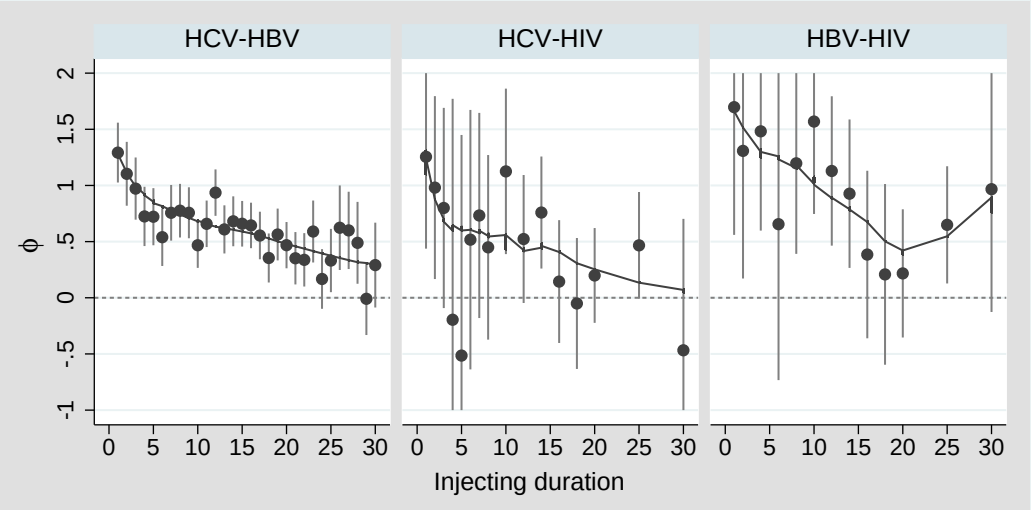


Figure 6.1: Associations between the three infection pairs by injecting duration, estimates of $\Phi$ and 95% confidence intervals, and Lowess curve. Injecting durations are grouped where data are sparse to prevent zero cells.

It is worth considering the routes of transmission for the different infections, with all being transmitted via sharing of injecting equipment but the addition of sexual transmission for HIV and HBV (which can be considered relatively uncommon for HCV). The Lowess curves for HCV-HBV and HCV-HIV pairs are more similar than the HBV-HIV pair (albeit with wide confidence intervals for the HIV pairs). This is as would be expected if the correlation is due to heterogeneity in injecting behaviour.

The HCV-HIV pair may have a slightly lower correlation, and in particular decline fastest towards zero. Data are too sparse to be conclusive, but this might point to a smaller component of shared risk of infection following the high-risk initiation period. Conversely, the HIV-HBV pair, which have identical transmission routes, show the highest level of correlation, which may reflect the additional shared heterogeneity resulting from the sexual transmission route. Although conclusions are uncertain, the observed patterns are consistent with epidemiological considerations. Also of note is that sexual exposure may occur prior to injecting for HBV and HIV, which could affect the evolution of the RFV in unpredictable ways, depending on the baseline forces of infection and correlation between the routes of transmission.

It is also of interest to see whether the RFV may have different patterns over time and according to covariate levels. Data are too sparse for further subdivision of the HIV pairs, but this can be investigated for HCV-HBV. The following plots therefore show the measure $\Phi$ for HCV-HBV by injecting duration according to survey year, region, age at first use, gender, ever received works and ever imprisonment (it was not possible to examine whether men who have sex with men and those that have never used a needle exchange have different patterns, as both these groups are too small).

Figure 6.2 shows $\Phi$ according to survey year, which is divided into three 5-year periods from 2000 onwards. The relative frailty variance appears to change pattern across periods, with a continuous decline for 2000-2004, a sharp decline over the first 5 years of injecting and more constant thereafter for 2005-2009, and fairly constant for 2010-2014.

Results for region are shown in Figure 6.3, with North East combined with Yorkshire and Humber and East and West Midlands combined due to sparsity of data. $\Phi$ appears to be high initially but declining for London, compared to a lower, but more stable value in the North West. Results

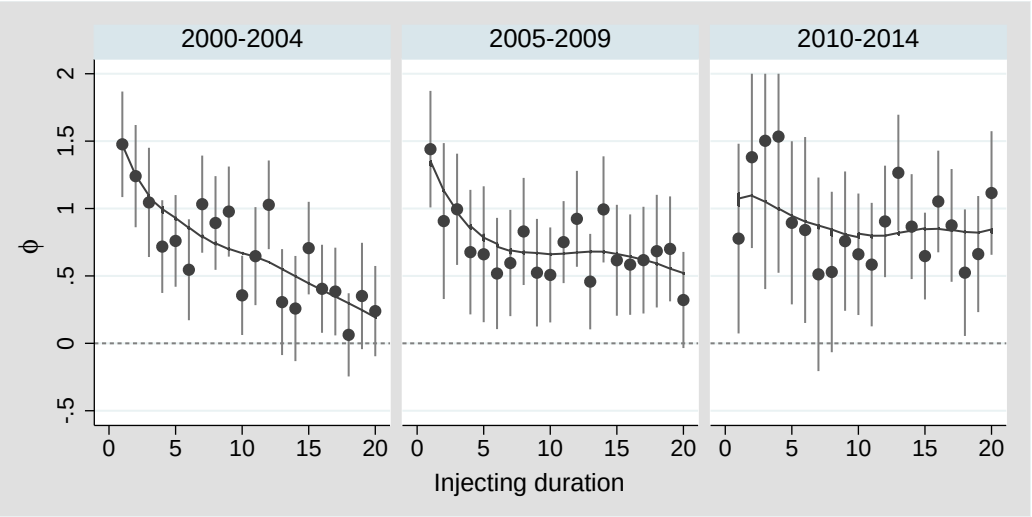


Figure 6.2: Association between HCV-HBV by survey period and injecting duration, estimates of Φ and 95% confidence intervals, and Lowess curve. Injecting durations are grouped where data are sparse to prevent zero cells.

are less precise for other regions, but appear to point to different regional patterns, with a sharp decline and lower correlation subsequently in the South West, higher correlation in Wales, and an unusual "bathtub" (decreasing then increasing) shape for the combined North East and Yorkshire and Humber region.

Figure 6.4 shows Φ for males and females. The decline in RFV is slightly sharper for males compared to females, and interestingly HCV-HBV infection in males has a generally stronger association, indicating that females may be more homogeneous in their risk of infection.

Figure 6.5 shows Φ by age at first use. The pattern for those that began injecting at less than 18 years of age shows a slow decline over injecting career, while those aged 18-24 have a steeper decline over the first 5 years and more stable thereafter. The pattern is similar for those that began injecting at 25 or older, but with a lower value of Φ at initiation.

There was little difference in pattern according to whether ever received works or not (Figure 6.6), although Φ may remain slightly elevated at longer injecting durations in those that report never receiving works. The difference is plausible: those that report never sharing equipment are a mixture of those that genuinely have never shared and those that have incorrectly answered the question (if all responses were genuine, prevalence in this group would be

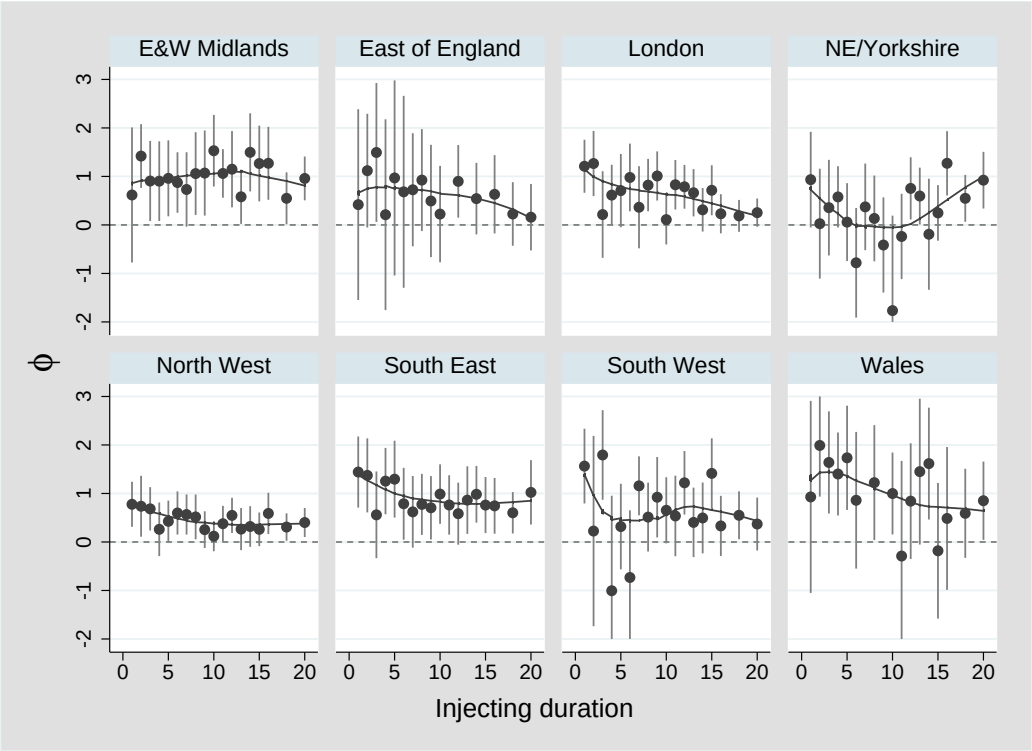


Figure 6.3: Association between HCV-HBV by region and injecting duration, estimates of Φ and 95% confidence intervals, and Lowess curve. Injecting durations and some regions are grouped where data are sparse to prevent zero cells.

extremely low, as those infected with a blood-borne virus are highly likely to have acquired infection via sharing injecting equipment). Conversely, those that report sharing equipment would probably be less likely to have misreported their answer, and therefore be a more homogeneous group in terms of risk.

Finally, Figure 6.7 shows Φ by ever-imprisoned status. Φ is at around the same level at initiation, but declines more quickly to a slightly lower value for those ever imprisoned. For most PWID, imprisonment is likely at some point in injecting career (69% of those in the data from 2000 onwards have been imprisoned) and therefore those that have never been imprisoned are the rarer, and apparently more heterogeneous group. There may also be an aspect of under-reporting of imprisonment, similar to that for ever receiving works, which would result in a mixture of risk-types in the never-imprisoned group.

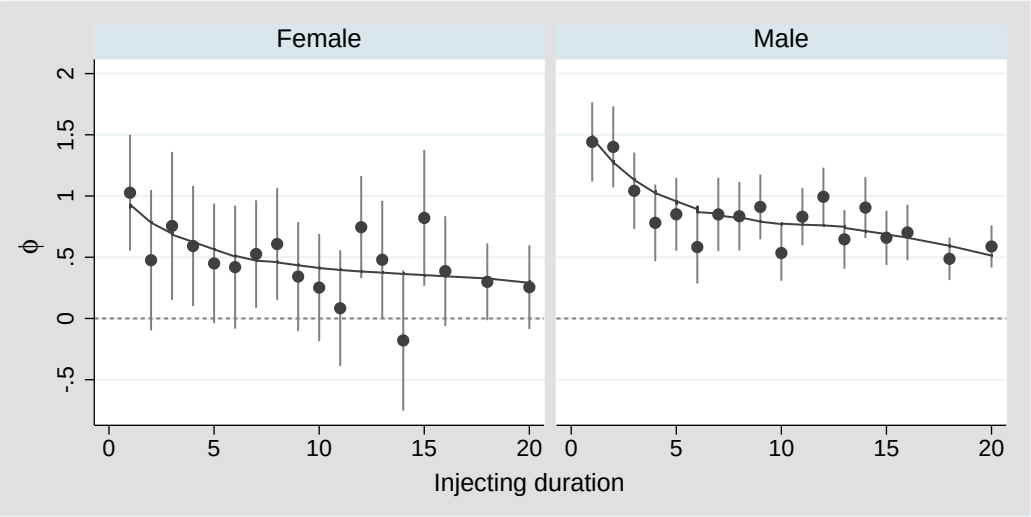


Figure 6.4: Association between HCV-HBV by gender and injecting duration, estimates of Φ and 95% confidence intervals, and Lowess curve. Injecting durations are grouped where data are sparse to prevent zero cells.

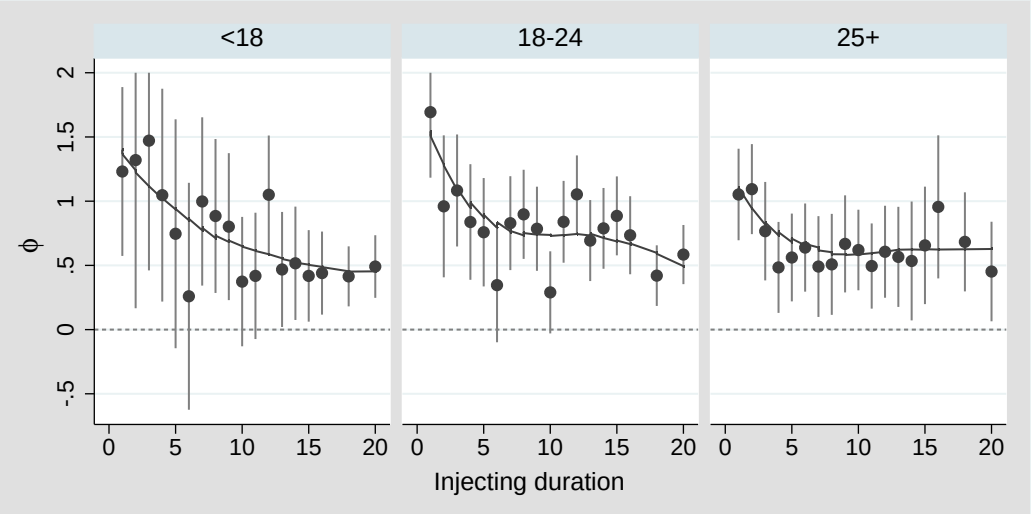


Figure 6.5: Association between HCV-HBV by age at first use and injecting duration, estimates of Φ and 95% confidence intervals, and Lowess curve. Injecting durations are grouped where data are sparse to prevent zero cells.

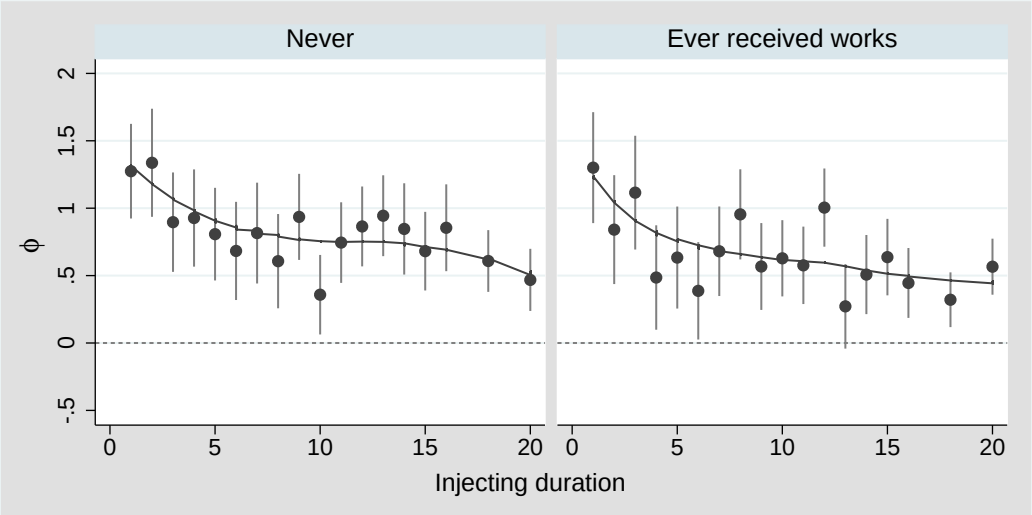


Figure 6.6: Association between HCV-HBV by ever receiving works and injecting duration, estimates of Φ and 95% confidence intervals, and Lowess curve. Injecting durations are grouped where data are sparse to prevent zero cells.

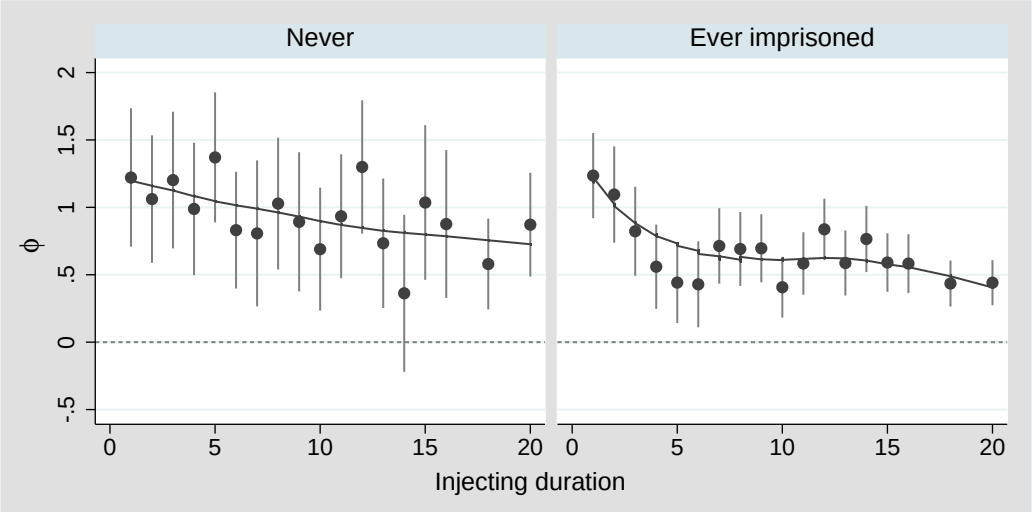


Figure 6.7: Association between HCV-HBV by ever-imprisonment and injecting duration, estimates of Φ and 95% confidence intervals, and Lowess curve. Injecting durations are grouped where data are sparse to prevent zero cells.

### 6.1.1 Conclusions

The data from all three pairs of infections point strongly toward a declining frailty variance in survivors, which is consistent for the majority of subgroups across a range of covariates. However, the pattern of decline changes markedly across some subgroups, and according to survey year. Stratified plots of $\Phi$ can reveal such differences, but may not capture the structure of residual heterogeneity after accounting for time and other covariates; in other words, although the patterns may appear to vary across subgroups, a common frailty distribution might still capture the residual heterogeneity adequately in more complex models. This can be assessed via the predicted association between infections, with $\Phi$ calculated from the bivariate prevalence predicted by a frailty model and plotted against the observed values to examine how well the chosen distribution for residual frailty fits. These plots require some care in interpretation, representing the estimated association between infections conditional on both the frailty distribution and any modelled covariates.

## 6.2 Bivariate frailty models: trends and risk patterns

The inclusion of frailty distributions is first examined within models with injecting duration and time effects only, with and without interactions between them. Data from 2000 onwards are considered, for which key covariates are non-missing (equipment sharing, needle exchange, imprisonment and men who have sex with men. Region, age and sex are always complete). Within this subset of data, all individuals have data on at least one infection test within each bivariate pair (HCV-HBV, HCV-HIV and HBV-HIV). Therefore the same dataset is used in all subsequent analyses.

Calendar time and injecting duration are defined as in section 4.2: pre-1980, 1980-1985, 1985-1990, 1995-2000, 2000-2005, 2005-2010 and 2010 onwards; and injecting duration as $\leq 1$, 1-3, 3-5, 5-10, 10-15, 15-25 and $> 25$ years. Four models are considered for the frailty distribution: independence (no frailty), gamma, inverse Gaussian and a single-component time-varying frailty (TVF) model as described in section 5.1.4. The gamma distribution

provides a constant relative frailty variance (RVF), such that the heterogeneity of survivors is unchanged, whereas the inverse Gaussian distribution results in declining RVF due to selection effects. The TVF model uses a gamma distribution that declines in variability over time towards an asymptote (equation 5.14).

### 6.2.1 Estimated frailty variance and model fit

Table 6.1 shows frailty variances and AIC scores for the eight possible models (four frailty distributions, with and without interaction between time and injecting duration) fitted to the bivariate HCV-HBV data. Under the no interaction models, the frailty variance $\delta$ is greater for the inverse Gaussian than the gamma model, and highest for the TVF model. These values are for the frailty variance at $t = 0$, when all individuals are susceptible, but this variability decreases over time for the non-gamma frailty models. The $\rho$ parameter for the TVF model is 0.99 (see equation 5.14), such that the frailty variance declines rapidly towards the asymptotic lower bound; the estimate of $q$ is 0.74 such that this lower bound is 0.68, around two-thirds of the initial frailty variance.

Table 6.1: Bivariate model statistics for HCV-HBV data, with and without interaction between injecting duration and time and under four different frailty distributions. Frailty variance $\delta$, -log-likelihood, number of parameters $k$ and AIC score.

| | No interaction | | | | Interaction | | | |
|---|---|---|---|---|---|---|---|---|
| Distribution | $\delta$ | -LL | d.f. | AIC | $\delta$ | -LL | $k$ | AIC |
| Independence | NA | 22798.1 | 28 | 45652.3 | NA | 22770.3 | 52 | 45644.6 |
| Gamma | 0.97 | 22445.8 | 29 | 44949.6 | 0.98 | 22424.5 | 53 | 44955.0 |
| Inv. Gaussian | 1.94 | 22422.0 | 29 | 44901.9 | 1.92 | 22398.2 | 53 | 44902.3 |
| Time-varying | 3.16 | 22416.0 | 31 | 44893.9 | 3.18 | 22392.0 | 55 | 44894.1 |

There is a substantially better fit to the data under the frailty models compared to independence, and a fair improvement under the inverse Gaussian and TVF models compared to the gamma model, indicating that the RFV in survivors declines over time in some way. The TVF model provides some further improvement on the inverse Gaussian, although the two extra parameters for the rate of decline and asymptotic minimum will provide more

flexibility so this cannot necessarily be viewed as evidence for a time-varying frailty rather than a selection effect; it may be that neither distribution is quite right.

There is very little change in any of the frailty parameter estimates when including an interaction between injecting duration and time. For all frailty models the AIC score is increased by the addition of interaction terms, although there is an improvement for the independence model. Figure 6.8 shows the observed and predicted bivariate proportions of HCV-HBV under the inverse Gaussian model. The model fit appears reasonable from visual inspection, although there is some lack of fit at longer injecting durations. This occurs for models with and without interactions and is due to the fixed effects of injecting duration in the pre-survey period: as the cumulative hazard must increase monotonically, there is a lower bound on prevalence at longer injecting durations. There is little that can be done to address this, as interaction terms outside the range of the survey data are not identifiable. As in section 4.2.1, models with and without interactions are very similar for the data from 2000 onwards, and the predicted prevalences are also similar for other frailty distributions.

Figure 6.9 shows estimates of the measure of association $\Phi$ based on the predicted bivariate prevalence from the different frailty models, with interactions between injecting duration and calendar time. This is close to constant for the gamma model, as would be expected, but increases and decreases slightly with injecting duration due to changes in covariates for calendar time. The predicted value of $\Phi$ for the inverse Gaussian model decreases over time, as expected, due to the selection effect. Finally, under the TVF model $\Phi$ decreases rapidly over the first 3 years of injecting duration, but then continues to decrease, and more rapidly than the inverse Gaussian. This may be counter-intuitive, as under the estimated parameters for the TVF distribution the frailty variance is very close to its asymptotic value within a few years and is no longer declining substantially. However, the association measure $\Phi$ at time $t$ reflects the cumulative impact of the association over the interval $(0, t]$, rather than its impact at time $t$; this is unavoidable with current status data, as event times are not observed. For all frailty distributions, estimates of $\Phi$ are very close for the models with and without interactions.

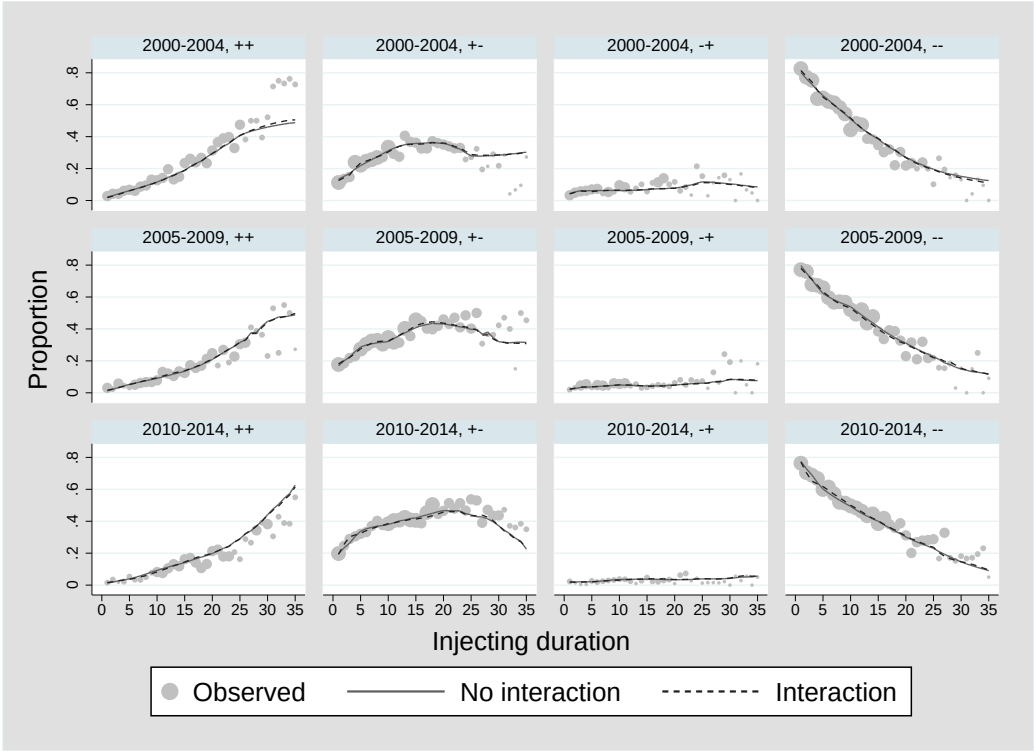


Figure 6.8: Observed and predicted bivariate HCV-HBV prevalence (positive +, negative -) by injecting duration and survey period, inverse Gaussian frailty. Models with and without interactions between injecting duration and calendar time effects. Observed proportions are plotted according to sample size.

Assessing the absolute fit of the models is complicated by the sparsity of the multinomial data, with HBV-positive HCV-negative individuals comparatively rare. Of the possible survey year, injecting duration and bivariate infection status combinations, 21.6% of cells are zero, 13.6% are ones and 61% of all cells less than or equal to 5. Categories are therefore first collapsed before calculating the deviance, according to the categorisation of survey year and injecting duration used for the logistic regression models in section 4.1.1 and removing injecting durations of $> 35$ years. This results in 60 combinations of survey period/injecting duration category and only 13.3% of cells less than or equal to 5. The deviance of the inverse Gaussian model with no interaction between calendar time and injecting duration effects was 332.4 on 151 degrees of freedom, and the interaction model 296.7 on 127 degrees of freedom, indicating neither fit the data very well. Exam-

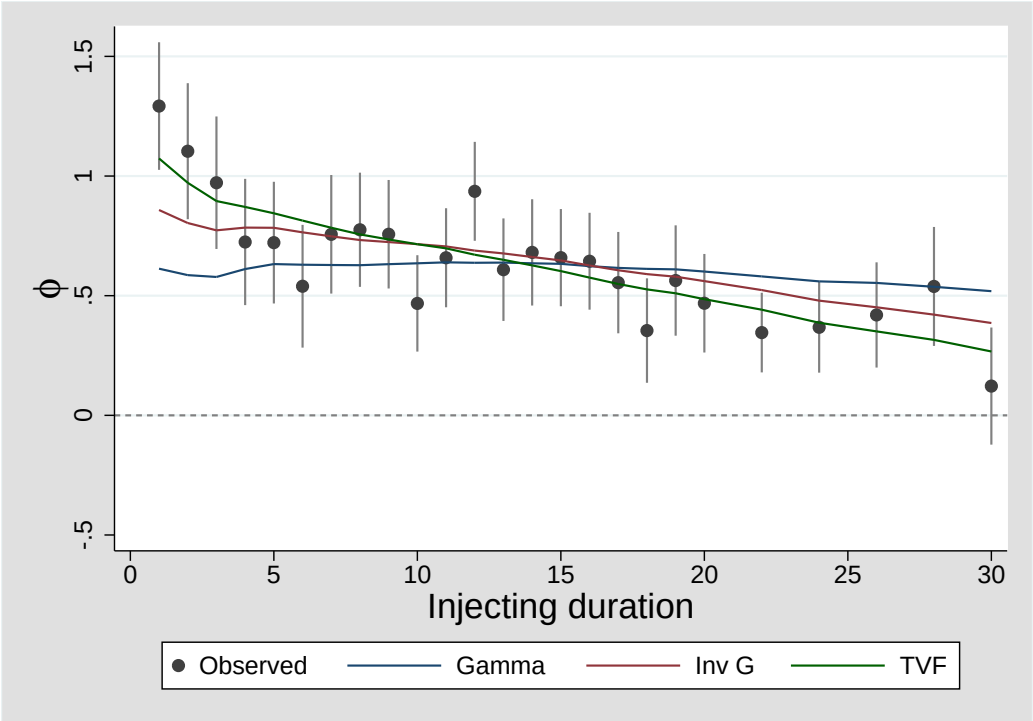


Figure 6.9: Observed and predicted estimates of association between HCV and HBV under different frailty distributions, with interactions between injecting duration and calendar time. Injecting durations are grouped where data are sparse to prevent zero cells.

ination of the data points with high deviance contributions reveals little in the way of a regular pattern, with areas of poor fit spread across injecting durations and survey years, and pointing more to overdispersion rather than inadequate model structure.

To further assess the fit of the model the univariate deviance for HCV and HBV were examined separately. As data are less sparse, survey year is not grouped and injecting duration can be grouped at a finer level (un-aggregated for durations up to 10 years, 2-year bands from 10-20 years, 5-year bands subsequently). The univariate deviance for HCV was 282.7 on 255 degrees of freedom for the model with no interaction between injecting duration and calendar time and 254.6 on 243 degrees of freedom for the interaction model, and for HBV a deviance of 417.5 on 255 degrees of freedom for the no interaction model and 397.5 on 243 degrees of freedom for the interaction model. The fit is therefore adequate for HCV but somewhat poor for HBV.

Plots of deviance residuals are shown in figures 6.10 and 6.11 for HCV and

HBV. As would be expected from the deviances, the variability of deviance residuals is higher for HBV than HCV. There do not appear to be obvious systematic patterns in the residuals for HCV, although there is perhaps more variability in residuals at shorted injecting durations. For HBV there are some systematic patterns, but this tends to occur in individual survey years; for instance most of the residuals are negative in 2009 and 2014. This was noted in section 4.1.2: such "local" patterns can only be captured within the model if the structural part of the model includes parameters for individual years. Further, the FOI model is parameterised in terms of cumulative risk over calendar time and injecting duration, so sharp changes in prevalence in individual survey years would only be captured by interaction terms at the level of individual years.

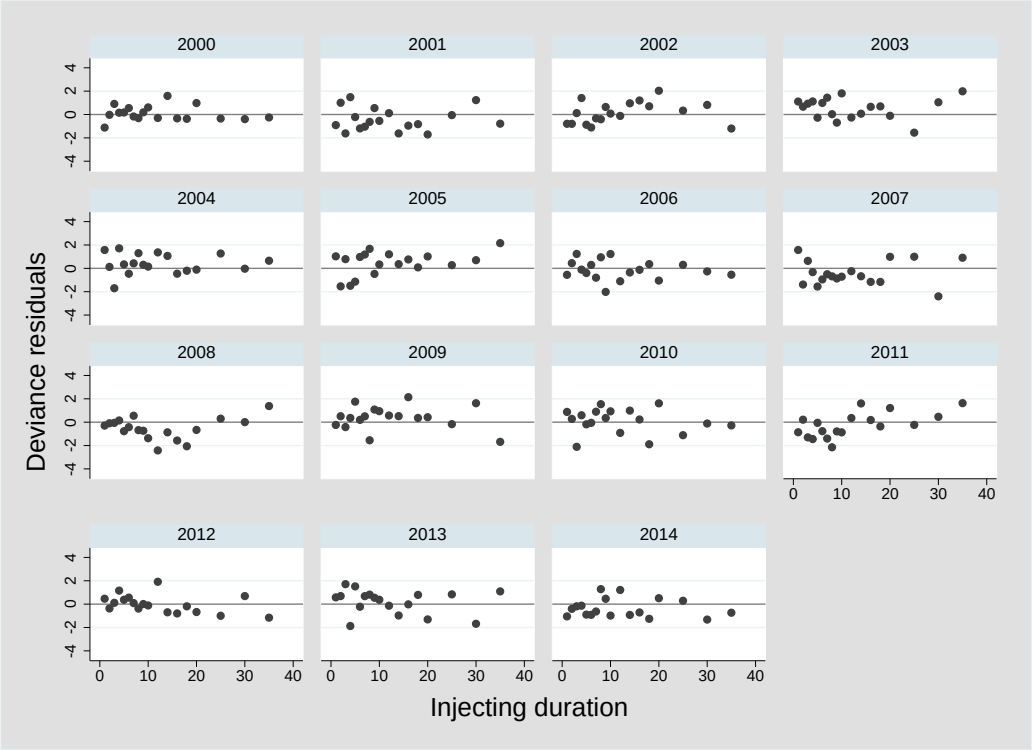


Figure 6.10: Deviance residuals for HCV under bivariate HCV-HBV model, with survey year and injecting duration interaction and inverse Guassian frailty.

For the HCV-HIV and HBV-HIV pairings there were some model fitting issues: there is far less information to estimate the frailty variance, in particular changes over time, which is made yet more difficult with the inclusion

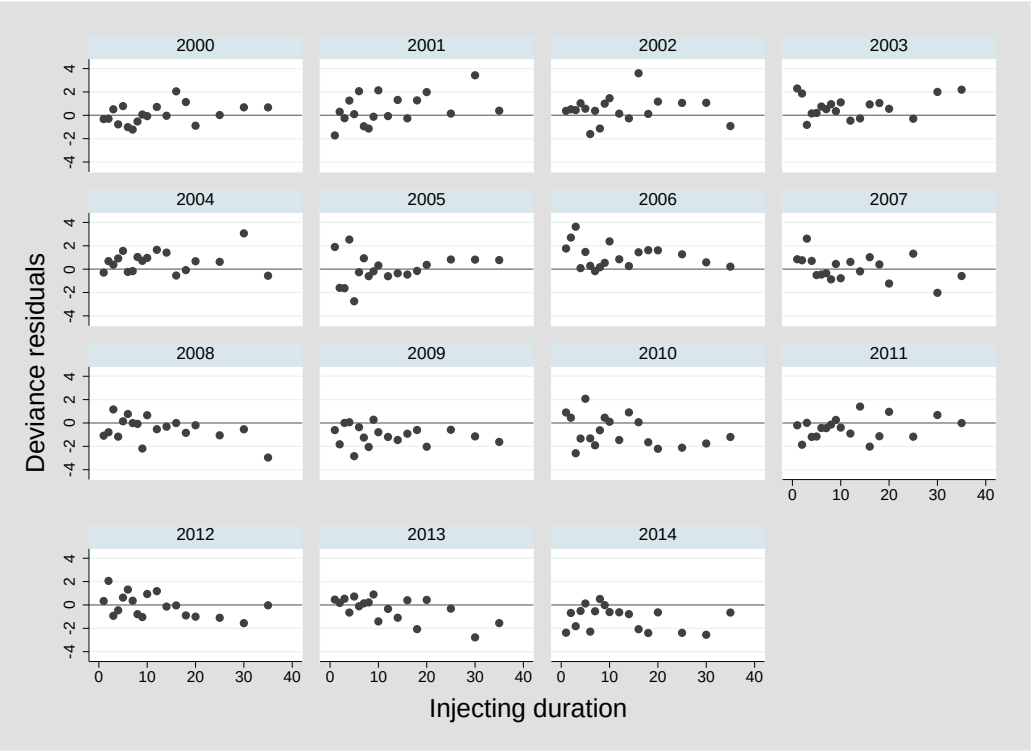


Figure 6.11: Deviance residuals for HBV under bivariate HCV-HBV model, with survey year and injecting duration interaction and inverse Guassian frailty.

of interactions between injecting duration and time. Table 6.2 shows the model fit statistics and frailty variances for the bivariate HCV-HIV model. Patterns are similar to those for HCV-HBV, but somewhat less conclusive, with a declining frailty variance being preferred to the gamma model but little difference between the inverse Gaussian and TVF models, and a very high estimate of the frailty variance under the latter (9.17). For the TVF model the estimate for $\rho$, the rate of decline in relative frailty variance, is 73.0, resulting in an immediate decline to the asymptote, estimated at 0.47. The interaction model results appear unreliable, with the gamma frailty being preferred to the inverse Gaussian, and an extremely high estimate of the frailty variance for the TVF model (361). Results are also inconclusive as to whether the interaction terms are necessary or not, with a better AIC score for the independence and gamma models with the interaction, but worse for the two other models.

Table 6.3 shows the model fit statistics for the bivariate HBV-HIV mod-

Table 6.2: Bivariate model statistics for HCV-HIV data, with and without an interaction between injecting duration and time and under four different frailty distributions. Frailty variance $\delta$, -log-likelihood, number of parameters $k$ and AIC score.

| | No interaction | | | | Interaction | | | |
|---|---|---|---|---|---|---|---|---|
| Distribution | $\delta$ | -LL | d.f. | AIC | $\delta$ | -LL | $k$ | AIC |
| Independence | - | 15048.6 | 28 | 30153.2 | - | 15020.9 | 52 | 30145.8 |
| Gamma | 0.84 | 15042.9 | 29 | 30143.9 | 0.56 | 15008.4 | 53 | 30122.8 |
| Inv. Gaussian | 0.89 | 15033.4 | 29 | 30124.7 | 0.82 | 15015.1 | 53 | 30136.3 |
| Time-varying | 9.17 | 15032.3 | 31 | 30126.6 | 360.8 | 15004.2 | 55 | 30118.4 |

els. The pattern is similar to the HCV-HBV results, and appear more orderly: the AIC scores decline for the independence, gamma, inverse Gaussian, and TVF models in turn for models with and without interactions, and there is a fairly consistent preference for the no interaction model over the interaction model. The frailty variance is estimated to be higher for the inverse Gaussian model, which is to be expected, given that the variance in survivors will decline. The TVF model has a high variance (11.5 no interaction, 8.7 interaction) and a less rapid rate of decline than the other pairings($\rho = 0.43$ and 0.18 for no interaction/interaction models). However, the asymptotic minimum is far lower at 0.26 for the no interaction model and practically zero for the interaction model, indicating far less heterogeneity in survivors over time.

Table 6.3: Bivariate model statistics for HBV-HIV data, with and without an interaction between injecting duration and time and under four different frailty distributions. Frailty variance $\delta$, -log-likelihood, number of parameters $k$ and AIC score.

| | No interaction | | | | Interaction | | | |
|---|---|---|---|---|---|---|---|---|
| Distribution | $\delta$ | -LL | d.f. | AIC | $\delta$ | -LL | $k$ | AIC |
| Independence | NA | 10378.8 | 28 | 20813.7 | NA | 10368.6 | 52 | 20841.2 |
| Gamma | 1.42 | 10352.1 | 29 | 20762.2 | 1.44 | 10328.1 | 53 | 20762.3 |
| Inv. Gaussian | 2.34 | 10349.3 | 29 | 20756.5 | 2.30 | 10325.9 | 53 | 20757.7 |
| Time-varying | 11.46 | 10341.8 | 31 | 20745.5 | 8.73 | 10320.9 | 55 | 20751.9 |

Figure 6.12 shows the observed and predicted proportions of bivariate HCV-HIV status by survey period and injecting duration. The fit is generally good for those not infected with HIV, but proportions of HIV positive are

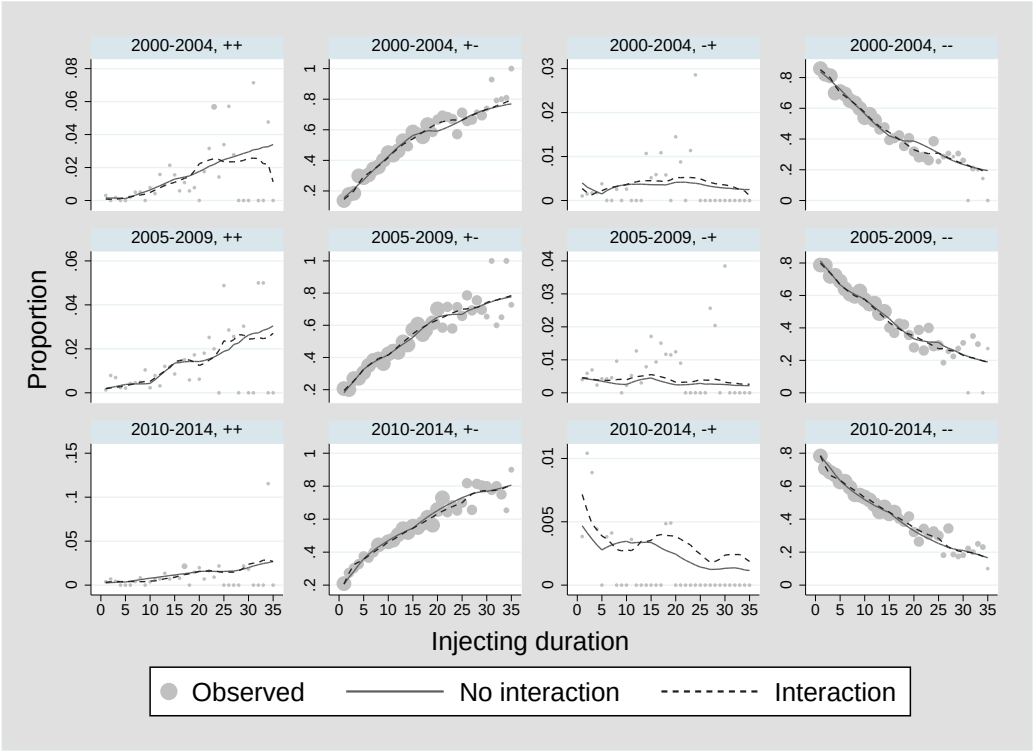


Figure 6.12: Observed and predicted bivariate HCV-HIV prevalence (positive +, negative -) by injecting duration and survey period, gamma frailty. Models with and without interactions between injecting duration and calendar time effects, and no-interaction model with shared injecting duration effect. Observed proportions are plotted according to sample size; y-axis scales vary for clarity.

very low, and models both with and without interactions do not capture the variability in observed proportions very well due to the sparsity of the data.

Figure 6.13 shows the observed and predicted proportions of bivariate HBV-HIV status by survey period and injecting duration. As with the HCV-HBV model, there is some lack of fit for HBV at longer injecting durations. Proportions of HIV positive are very low, and models both with and without interactions do not capture the variability in observed proportions very well. The figure also displays predictions according to a shared injecting duration effect (see section 6.2.2).

Figure 6.14 shows estimates of the measure of association $\Phi$ between HCV and HIV based on the predicted bivariate prevalence from the different frailty models, with interactions between injecting duration and calendar time. The

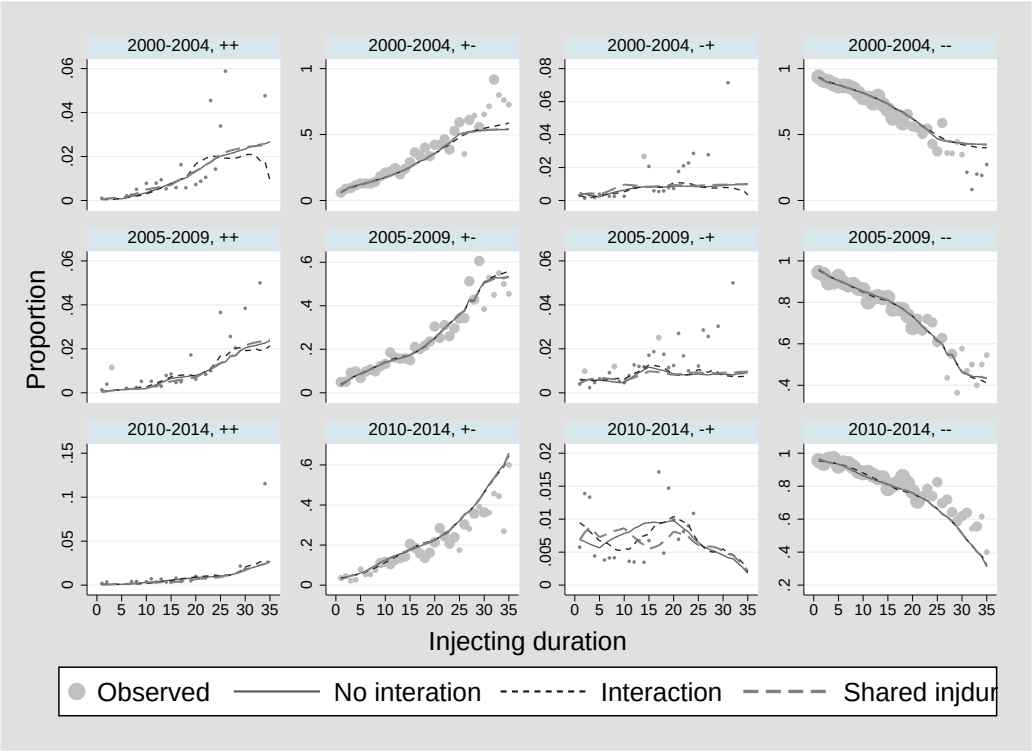


Figure 6.13: Observed and predicted bivariate HBV-HIV prevalence (positive +, negative -) by injecting duration and survey period, gamma frailty. Models with and without interactions between injecting duration and calendar time effects, and no-interaction model with shared injecting duration effect. Observed proportions are plotted according to sample size; y-axis scales vary for clarity.

inverse Gaussian model shows a slight decline in association as injecting duration increases, but is generally comparable to the gamma frailty model. The TVF model captures the decline in $\Phi$ better; this is due to the very high estimate of initial frailty variance and subsequent high rate of decline, although this was somewhat less for the model without interactions. For the other frailty distributions, there was little difference between models with or without interactions.

Figure 6.15 shows estimates of the measure of association $\Phi$ between HBV and HIV based on the predicted bivariate prevalence from the different frailty models, with interactions between injecting duration and calendar time. The inverse Gaussian model shows a slightly higher association at shorter injecting durations but $\Phi$ declines to the same level as the estimates

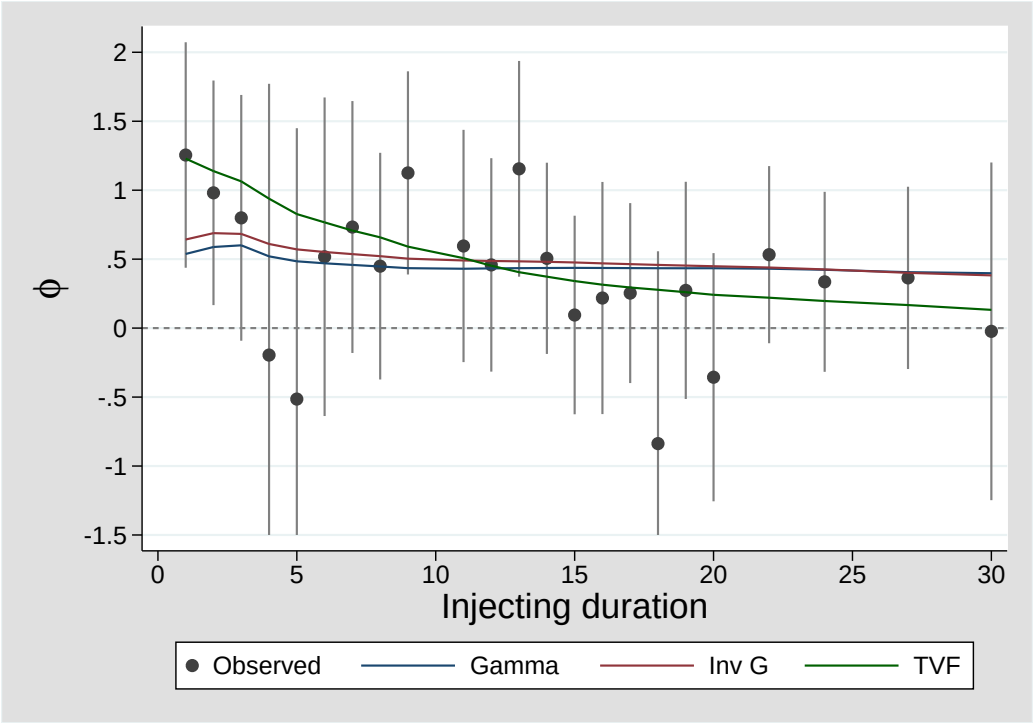


Figure 6.14: Observed and predicted estimates of association between HCV and HIV under different frailty distributions, with interactions between injecting duration and calendar time effects. Injecting durations are grouped where data are sparse to prevent zero cells.

from the gamma distribution at longer durations (over 20 years). Neither fit the observed pattern well for shorter injecting durations. The TVF model however captures the decline in $\Phi$ better over this period and although data are sparse, appears the most plausible model for the observed frailty. There are only minor differences in the estimates for models with and without interactions.

Assessment of model fit is again difficult, as data are even sparser for infection pairs including HIV, with 46% of cells equal to zero and 72% with counts of five or less. Grouping survey year and injecting duration as for HCV-HBV gives a deviance of 308.9 on 151 degrees of freedom for the gamma frailty model with no interaction between injecting duration and survey year effects, and 270.5 on 127 degrees of freedom for the interaction model. Again, this appears to be partly due to overdispersion and the constrained lower bound on prevalence at longer injecting durations, with models with and without interaction terms overestimating HBV prevalence at longer injecting

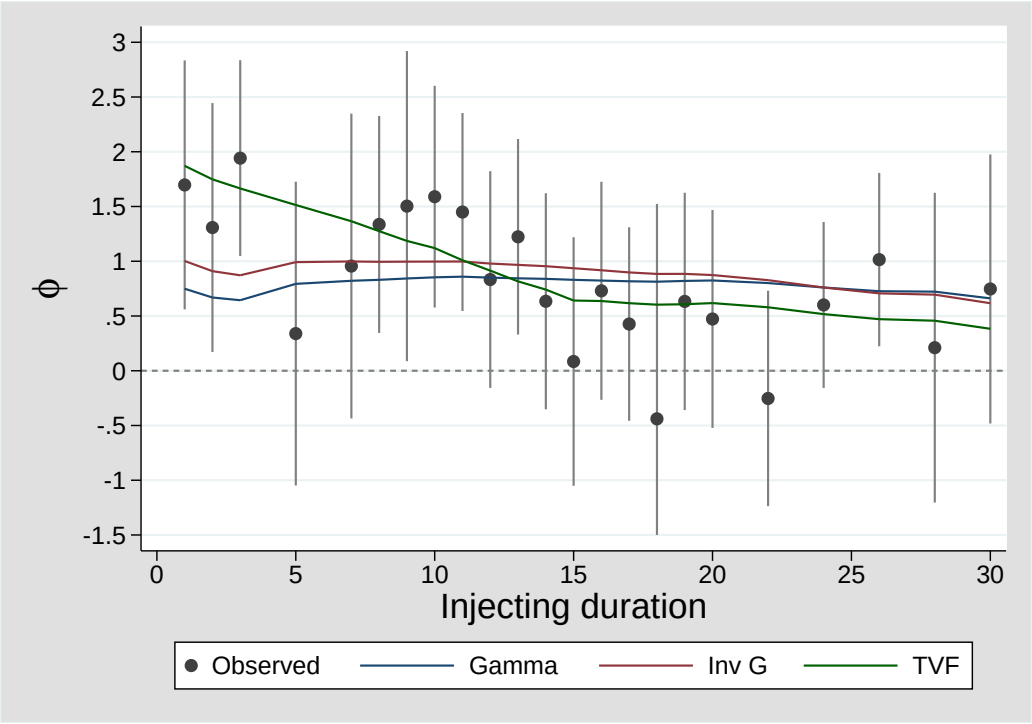


Figure 6.15: Observed and predicted estimates of association between HBV and HIV under different frailty distributions, with interactions between injecting duration and calendar time effects. Injecting durations are grouped where data are sparse to prevent zero cells.

durations in more recent years.

### 6.2.2 Trends and risk patterns under bivariate frailty models

Parameter estimates for the baseline FOI by calendar time and injecting duration-specific HRs for HCV and HBV are shown in Table 6.4, for each of the frailty distributions, in the model with no interaction between injecting duration and calendar time. As observed in section 4.2, there is a substantially increased risk of infection in the 1st year of injecting (compared to 3-5 years) for both HCV and HBV. This is reduced somewhat under the frailty models compared to the independence model, but still very high. Accounting for heterogeneity therefore only partly explains the high prevalence of BBVs in 1st year injectors. Subsequent to the first year of injecting, the FOI stays relatively constant for HCV and HRs for injecting durations of 5-10 and

10-15 years are fairly similar under the different frailty models. Beyond 15 years the estimates differ according to the choice in frailty distribution. This is due to a combination of sparse data and the constraining effect described above.

Temporal patterns for HCV suggest that risk was highest in the early 1980s and declined slightly until 1995, was markedly reduced in 1995-2000 but then slowly increasing from this point onwards. Temporal patterns are fairly similar for the different models (the baseline rates vary due to the different estimates of injecting duration, but the patterns are similar) except for the period before 1980, which are affected by the choice of frailty distribution. All of the frailty models estimate a low or practically zero risk (unlike the independence model) prior to 1980, which may be epidemiologically plausible.

For HBV, as well as a very high FOI in the 1st year, the FOI for years 1-3 is 2-3 times higher than the baseline group of 3-5 years. The estimated FOI is then similar for 3-5 years and 5-10 years, but estimates are practically zero from 10 years onwards, such that all of the estimated risk of infection occurs in the first 10 years of injecting. Therefore although observed HBV prevalence increases monotonically with injecting duration, according to the model this is solely due to higher levels of risk in the past, which has decreased steadily since 1980.

Obtaining parameter estimates for injecting duration-specific hazard ratios is problematic for HIV. This is due to sparsity of the outcome, and a general pattern of injecting duration-specific prevalence that causes difficulties in fitting: prevalence rises sharply at initiation, but then does not increase monotonically, staying at around the same level from durations 2-10 years before increasing again. This results in HRs for injecting duration-specific risk being estimated as practically zero or extremely high in some categories, with variable patterns in estimated risk depending on the infection pair (HCV-HIV or HBV-HIV) and chosen frailty distribution (although an excess risk in the first year is a common feature). Despite this instability in estimated injecting duration-specific risk, estimates for temporal trends are plausible and similar across different models and infection pairs. Nevertheless, it would be preferable to model the data according to a reasonable pattern of risk, rather than one in which the estimated at-risk periods only

Table 6.4: Parameter estimates for bivariate HCV-HBV model under different frailty distributions: independence (Indep), gamma, inverse Gaussian (Inv G) and time-varying frailty (TVF). Model with main effects for calendar time and injecting duration (no interaction).

| | **HCV** | | | | **HBV** | | | |
|---|---|---|---|---|---|---|---|---|
| Parameter | Indep | Gamma | Inv G | TVF | Indep | Gamma | Inv G | TVF |
| *Time-specific baseline rate (3-5 years injecting)* | | | | | | | | |
| pre-1980 | 0.054 | 0.000 | 0.016 | 0.000 | 0.059 | 0.145 | 0.179 | 0.225 |
| 1980-1985 | 0.075 | 0.119 | 0.153 | 0.145 | 0.032 | 0.057 | 0.074 | 0.095 |
| 1985-1990 | 0.069 | 0.098 | 0.124 | 0.115 | 0.020 | 0.029 | 0.037 | 0.048 |
| 1990-1995 | 0.066 | 0.103 | 0.129 | 0.128 | 0.012 | 0.016 | 0.020 | 0.025 |
| 1995-2000 | 0.046 | 0.062 | 0.072 | 0.069 | 0.009 | 0.012 | 0.015 | 0.018 |
| 2000-2005 | 0.060 | 0.089 | 0.107 | 0.103 | 0.007 | 0.009 | 0.010 | 0.012 |
| 2005-2010 | 0.066 | 0.099 | 0.119 | 0.114 | 0.003 | 0.004 | 0.005 | 0.006 |
| post-2010 | 0.076 | 0.117 | 0.142 | 0.136 | 0.002 | 0.003 | 0.003 | 0.004 |
| *Injecting duration-specific hazard ratios* | | | | | | | | |
| 1st year | 3.21 | 2.39 | 2.14 | 2.38 | 11.60 | 9.34 | 8.38 | 7.18 |
| 1-3 yr | 0.83 | 0.72 | 0.69 | 0.72 | 3.06 | 2.62 | 2.39 | 1.92 |
| 3-5 yr | 1 (ref) | 1 (ref) | 1 (ref) | 1 (ref) | 1 (ref) | 1 (ref) | 1 (ref) | 1 (ref) |
| 5-10 yr | 0.66 | 0.69 | 0.69 | 0.64 | 1.04 | 0.85 | 0.82 | 0.53 |
| 10-15 yr | 0.79 | 1.04 | 1.02 | 0.87 | 0.00 | 0.00 | 0.00 | 0.00 |
| 15-25 yr | 0.68 | 1.35 | 1.27 | 0.99 | 0.00 | 0.00 | 0.00 | 0.01 |
| 25+ yrs | 0.76 | 3.22 | 2.31 | 1.48 | 0.00 | 0.00 | 0.00 | 0.01 |

occur at initiation and 5-10 or 10-15 years, with a FOI of practically zero at other times.

Here, assuming common hazard ratios for injecting duration-specific risk becomes useful. HBV and HIV have identical transmission routes, and although risk has clearly changed over time, the relative hazards according to injecting duration should be broadly comparable. The likelihood ratio test (LRT) for separate injecting duration HRs vs. shared parameters, under the gamma model with no interaction between injecting duration and time, gave a $p$-value of 0.844, so no evidence of a poorer fit. Conversely, assuming shared hazard ratios for calendar time effects provided an extremely small LRT $p$-value ($p < 0.0001$) and therefore the assumption of common time trends is not tenable. The assumption of shared common parameters had some impact on the estimate frailty variance, decreasing from 1.42 to 1.29 with common injecting duration parameters, and to 0.54 under common time parameters.

Table 6.5 shows parameter estimates for calendar time and injecting du-

ration under separate and shared injecting duration effects. The time effect (which is expressed as a hazard ratio vs. 2005-2010) is little changed for HBV under shared effects, but shifted somewhat for HIV with a generally higher risk of infection prior to 1995, although following the same declining pattern, and higher risk post-2010. Estimates of injecting duration-specific risk differ for both infections, with the estimates being an average of the patterns of risk for the individual infections, but more heavily weighted toward the HBV estimate. Of note is that although the shared HRs appear markedly higher than for HBV for 1st year and 1-3 years injecting durations, this is more to do with the estimated FOI in the baseline category, 3-5 years, being "pushed" downwards, as all the HRs are increased. Figure 6.13 shows estimated bivariate prevalence from the shared parameter model in comparison with the non-shared model. There is little visual difference in estimated HBV prevalence for those that are HIV negative, and for HIV prevalence the data are too sparse to tell whether the estimates are better or worse.

For comparison, the equivalence of injecting duration and calendar time-specific HRs was tested for the other infection pairs, under the same model and frailty structure above. For HCV-HIV, there was also no difference in model fit under common injecting duration-specific hazard ratios ($p = 0.795$) and little difference with common time effects ($p = 0.221$). The frailty variance was again reduced, from 0.84 to 0.46 under common injecting duration parameters, and 0.59 under common time parameters. For HCV-HBV the fit was significantly worse when either effect was assumed to be the same ($p < 0.0001$). However, there was little change in the frailty variance (0.97 under no shared parameters, 0.90 with shared injecting duration parameters, 1.03 with shared time parameters).

### 6.2.3 Conclusions

Under the model with no interaction between injecting duration and time effects and a gamma frailty distribution, the estimated frailty variances were 0.97 for HCV-HBV, 0.84 for HCV-HIV and 1.42 for HBV-HIV, with generally similar results under the interaction models. These values represent a moderate level of individual heterogeneity, with the hazard ratio between

Table 6.5: Estimated hazard ratios from bivariate HBV-HIV model under gamma frailty distribution with separate and shared injecting duration effects. Model with main effects for calendar time and injecting duration (no interaction).

| | **Separate** | | **Shared** | |
|---|---|---|---|---|
| Parameter | HBV | HIV | HBV | HIV |
| *Time-specific baseline rate (3-5 years injecting)* | | | | |
| pre-1980 | 48.6 | 2.20 | 42.5 | 3.32 |
| 1980-1985 | 16.1 | 1.67 | 15.1 | 2.56 |
| 1985-1990 | 7.81 | 1.23 | 7.41 | 1.42 |
| 1990-1995 | 4.08 | 1.15 | 3.93 | 1.81 |
| 1995-2000 | 3.05 | 0.43 | 3.04 | 0.60 |
| 2000-2005 | 2.13 | 1.08 | 2.14 | 1.06 |
| 2005-2010 | 1 (ref) | 1 (ref) | 1 (ref) | 1 (ref) |
| post-2010 | 0.70 | 1.18 | 0.68 | 1.78 |
| *Injecting duration-specific hazard ratios* | | | | |
| 1st year | 7.89 | 571.2 | 20.9 | |
| 1-3 yr | 2.15 | 0.00 | 5.56 | |
| 3-5 yr | 1 (ref) | 1 (ref) | 1 (ref) | |
| 5-10 yr | 0.78 | 62.0 | 2.71 | |
| 10-15 yr | 0.00 | 141.8 | 0.31 | |
| 15-25 yr | 0.00 | 0.00 | 0.01 | |
| 25+ years | 0.00 | 0.00 | 0.01 | |

the 75th percentile and the 25th percentiles of the estimated gamma distribution being 4.7 for HCV-HBV, 4.1 for HCV-HIV and 7.3 for HBV-HIV. There is strong evidence of a declining relative frailty variance, and the inverse Gaussian distribution provides a better fit and with a higher variance, predictably so as the correlation between infections naturally declines due to a selection effect, and therefore must be higher to start with.

The TVF model provides the most detail on how the frailty variance changes over time, although it is also the most sensitive to the parameterisation of the baseline FOI. For HCV-HBV there is a rapid decline in variance, dropping to around two-thirds the initial value. For HCV-HIV the initial variance is very high, but the decline is near-immediate, to around half the initial value. Due to the rapid decline the period of time that the frailty variance is near its initial value at $t = 0$ is extremely short. For the HBV-HIV pair the decline in variance is slower, but tends toward a low asymptotic

value, indicating homogeneity in long-term injectors.

None of the frailty models fully captured the pattern of relative frailty variance indicated by plots of $\Phi$ for HCV-HBV, which showed a sharp initial decline followed by a slower decline in RFV. As the frailty variance in the TVF model declines toward an asymptote, it can only capture one of these features unless additional parameters are introduced. The observed plots indicate a slight underestimation of the variability in the first 1-3 years of injecting, although the slow decline is present in the model predictions. For HBV-HIV, the TVF model gives a more plausible approximation of the observed pattern of $\Phi$. There is something of a dip in RFV around 4-7 years, but data are extremely sparse and the global pattern appears to have been captured well.

As noted in section 6.1, the two routes of transmission are only shared by HBV and HIV, with this pairing showing the strongest association and a smaller decline in RFV over time. Conversely, the HCV-HIV pairing showed the least association, with a generally smaller frailty variance in most models. Following this line of reasoning, heterogeneity in risk may be underestimated for HCV, given that the transmission route is only partially shared, although it appears the overlap between HCV and HBV risk is still high.

Identification of parameters is a problem for the HIV data. Tests of common parameters for each infection indicate that the effect of injecting duration is comparable for HIV and HBV. However, results also showed no evidence for a difference between the HCV and HIV injecting duration-specific HRs, although there was a significant difference for the HCV and HBV HRs. Clearly there is a lack of data on HIV to conclusively demonstrate similar patterns of risk according to injecting duration. Nevertheless, a common effect for injecting duration-specific risk for HBV and HIV is plausible, due to a wholly shared route of transmission. Conversely, there was strong evidence for differences in risk according to calendar time for HBV and HIV, confirming that direct and indirect infection control measures (HBV vaccination, HIV treatment) are likely to have had a different impact on levels of risk over time for these infections. For the HIV pairings, the estimated frailty variance was sensitive to the choice of shared parameters. However, the frailty variance for HCV-HBV showed little change, despite the shared parameter models giving a markedly poorer fit, so this again may be more

due to the sparsity of the HIV data, rather than a systematic effect of constraints on the injecting duration parameters.

## 6.3 Bivariate frailty models with covariates

In this section covariates are introduced into the frailty models to examine their effect on the force of infection and residual frailty. This is a somewhat different approach to model building than that usually employed, which would seek to identify relevant covariates and model structure before examining whether random effects were required to account for residual variation. However, the analysis of multivariate current status data largely focuses on unmeasured heterogeneity, and it is the introduction of covariate data that is relatively novel here.

The key factors of interest are the covariates identified in sections 4.1.4 and 4.2.4: demographic variables for region, age and sex, and risk factor information on sharing equipment, needle exchange use, imprisonment and men who have sex with men (MSM). There are 10 regions, with SE England set as the baseline group, and the age variable is defined in terms of age at first use, with $< 18$, $18 - 24$ and $25+$. Gender is binary (1 for females vs. 0 for males) and estimates the difference between females and men who do not have sex with men: a binary variable for MSM (1) vs. non-MSM (0) is included to estimate differences in MSM risk in males. Ever receiving "works" (injecting equipment) is binary, and the needle exchange use and imprisonment variables are defined in terms of ever occurring plus a time-varying covariate for pre- and post-first occurrence. This formulation is intended to unpick the possible effects for being the type of individual that is *ever* imprisoned or uses a needle exchange and the change in risk subsequent to having been first imprisoned or starting to use a needle exchange. An example of these two aspects of risk would be that heavier users may be more likely to have ever used a needle exchange, but experience a decline in risk once usage is initiated.

In most models in section 6.2 there was not a strong preference for including interactions between injecting duration and time, and little difference in the estimates of parameters for the frailty distribution. Therefore only the main effects for injecting duration and time are considered in this section,

which reduces the number of models that need be considered and avoids estimation problems that occur where data are sparse and a number of covariates included. Model building is conducted hierarchically in four steps, starting from the model with no interaction between injecting duration and calendar time effects in section 6.2 (1). Models 2-4 then successively add covariates to the model: (2) region (this being the most important factor identified in section 4.1.4), (3) age and sex, and (4) the risk factor covariates. The sequence is repeated for each of the three infection pairs under the four frailty models, giving 16 possible models for each pair of infections.

### 6.3.1 Estimated frailty variance and model fit

Table 6.6 shows the estimated frailty parameters under models with different covariate structures and frailty distributions. For the Gamma and inverse Gaussian models there is a sequential decline in the frailty variance $\delta$ as covariates are added to the model; however the reductions in variance are only 20% and 26% respectively. The pattern is similar for the TVF model, with a marked reduction in $\delta$ for the model with region vs. no covariates, although initial variance is slightly higher in the full model. In all the TVF models with covariates, the initial variance is lower but the asymptotic value to which the frailty variance declines is between 0.80 and 0.84, compared to 0.68 without covariates. The decline toward the asymptote is near-immediate under the covariate models, with $\rho$ ranging from 18.7 to 41.1, but very rapid under the model without the covariates also: with $\rho = 0.99$ the frailty variance is two-thirds of the way from the initial value to the asymptote within 1 year. These results may point to a high variance on initiation of injecting, which rapidly stabilises, although it is worth noting that the information to detect changes in frailty is low and models, and their estimates, are somewhat sensitive.

Model fit is steadily improved by the addition of covariates, with AIC scores decreasing by 1414.5, 167.1 and 619.0 as region, age and sex, and risk factors are added sequentially to the model (inverse Gaussian frailty). For the full covariate models, the inverse Gaussian distribution provides the lowest AIC score, with a difference in AIC score of 28.4 compared to the gamma distribution and 6.6 compared to TVF. The TVF model, while providing the

Table 6.6: Estimated variance parameters for HCV-HBV data under covariate models with different frailty distributions. No covariates (No cov); region; region, age and gender (R,A,G) and all covariates and risk factors (Full). Frailty variances $\delta$ and parameters for time-varying model: $\rho$, rate of decline; and asymptotic proportion to which the initial variance at $t = 0$ declines.

| Frailty | (1) No cov | (2) Region | (3) R,A,G | (4) Full |
|---|---|---|---|---|
| Gamma $\delta$ | 0.97 | 0.86 | 0.83 | 0.77 |
| Inv. Gaussian $\delta$ | 1.94 | 1.58 | 1.52 | 1.44 |
| Time-varying $\delta$ | 3.16 | 1.83 | 1.74 | 1.96 |
| $\rho$ | 0.99 | 41.1 | 18.7 | 24.8 |
| Asymptote | 0.68 | 0.84 | 0.84 | 0.80 |

best fit without covariates, appears unnecessarily complex once covariate information is incorporated. Nevertheless, there is still significant residual heterogeneity, and this appears to follow the same pattern of declining RFV identified in sections 6.1 and 6.2.

The absolute fits of the covariate models are difficult to assess, as the data are now cross-classified according to a number of factors. Examining the deviance according to data collapsed by survey period and injecting duration categories, as in section 6.2.1, reveals a modest improvement in fit but with a large number of additional parameters (deviance of 265.2 on 115 degrees of freedom, vs. 332.4 on 151 degrees of freedom). However, a substantial improvement would not be expected, as the covariates are not part of the survey year/injecting duration cross-classification used in the calculation of the deviance.

Figure 6.16 shows estimates of the measure of association $\Phi$ obtained under the full covariate model. Under the gamma frailty model the estimate declines over time: due to the covariate mix changing with injecting duration, accounting for covariates has partly explained the declining RFV. The inverse Gaussian predictions are similar, but with a more marked decline, as expected. Interestingly, the TVF results are near-identical to those predicted from the inverse Gaussian model, and therefore do not capture the initial high RFV at the start of injecting.

Table 6.7 shows the estimated frailty variances for the bivariate HCV-HIV models. Unlike HCV-HBV, there was not an orderly reduction in frailty

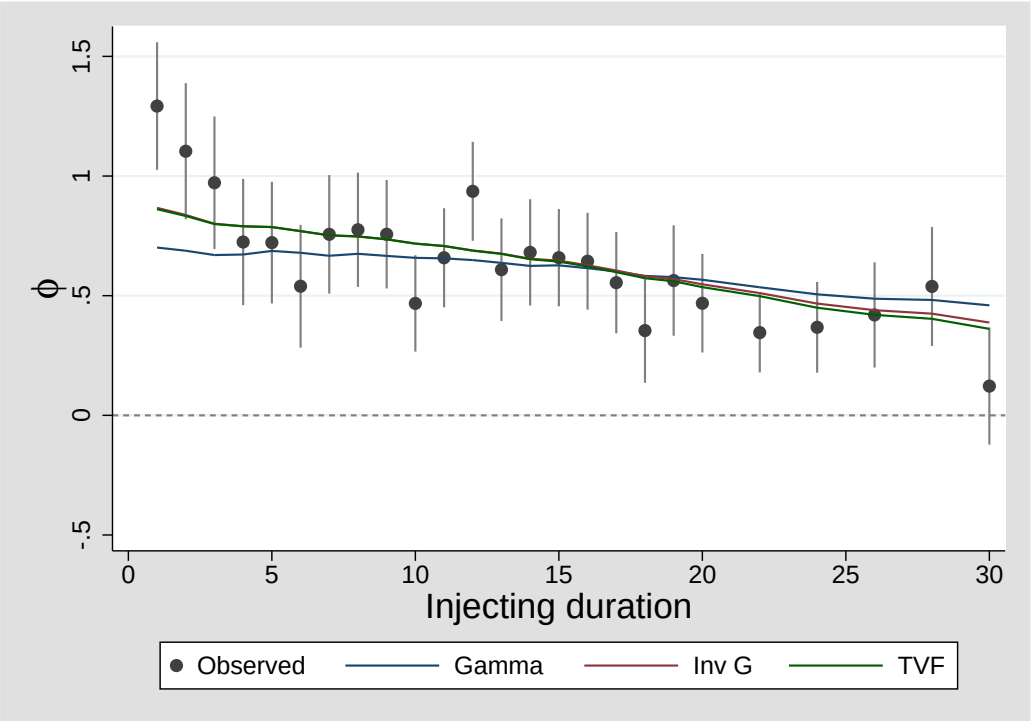


Figure 6.16: Observed and predicted estimates of association between HCV and HBV under different frailty distributions and including full covariate information. Injecting durations are grouped where data are sparse to prevent zero cells. *Inv G*=inverse Gaussian, *TVF*=time-varying frailty.

variance with the addition of extra parameters, with the variance increasing for both gamma and inverse Gaussian distributions when region was added to the model. The AIC scores indicated substantial improvement with the addition of covariates, reducing by 1249, 120 and 617 as each set of covariates was added to the model (gamma frailty). With region, age and sex, and under the full model, the gamma frailty gave the lowest AIC score. However, the TVF model could not be fitted at all, with estimation problems when region was entered in the model. Simplifications of the baseline FOI, grouping of regions or examining covariates other than region may have allowed some form of this model to be fitted, but as the HCV-HIV pair is of less interest this was not pursued further.

Table 6.8 shows the estimated frailty variances for the bivariate HBV-HIV models. There is some evidence that the frailty variance declines with the addition of covariates, but the reduction is small and does not sequentially decrease with the addition of each set of covariates, as for HCV-HBV. The

Table 6.7: Estimated variance parameters for HCV-HIV data under covariate models with different frailty distributions. No covariates (No cov); region; region, age and gender (R,A,G) and all covariates and risk factors (Full). Frailty variances $\delta$; time-varying model not fitted.

| Frailty | (1) No cov | (2) Region | (3) R,A,G | (4) Full |
|---|---|---|---|---|
| Gamma $\delta$ | 0.84 | 1.16 | 0.92 | 0.75 |
| Inv. Gaussian $\delta$ | 0.89 | 0.95 | 0.96 | 0.86 |

TVF model shows very different patterns under different covariates. For region only, the initial variance is lower but does not decrease as markedly over time, to an asymptotic value of 0.86 of its initial value. When age and gender are added, the initial variance is higher than with region alone and decreases to an asymptote of near-zero, but the decrease is very slow ($\rho$=0.07). Finally, with the other covariate information (full model) the decline in RFV is faster than with region, age and sex, but with a lower initial variance and a higher asymptote.

Table 6.8: Estimated variance parameters for HBV-HIV data under covariate models with different frailty distributions. No covariates (No cov); region; region, age and gender (R,A,G) and all covariates and risk factors (Full). Frailty variances $\delta$ and parameters for time-varying model: $\rho$, rate of decline; and asymptotic proportion to which the initial variance at $t = 0$ declines.

| Frailty | (1) No cov | (2) Region | (3) R,A,G | (4) Full |
|---|---|---|---|---|
| Gamma $\delta$ | 1.42 | 1.29 | 1.26 | 1.30 |
| Inv. Gaussian $\delta$ | 2.34 | 1.93 | 2.01 | 2.03 |
| Time-varying $\delta$ | 11.46 | 3.46 | 6.08 | 3.69 |
| $\rho$ | 0.43 | 0.51 | 0.07 | 0.15 |
| Asymptote | 0.26 | 0.86 | 0.00 | 0.34 |

Figure 6.17 shows predictions of the measure of association $\Phi$ for the TVF model under different sets of covariates. With no covariates, the estimate of $\Phi$ declines steadily over injecting duration and slightly overestimates the strength of association at longer injecting durations (over 15 years). Estimates appear worse with region only: the declining RFV is not captured by the model and only a slight decline is predicted, with overestimation of association at longer injecting durations. With age and sex also included the pattern is captured better and the predicted association is similar to the

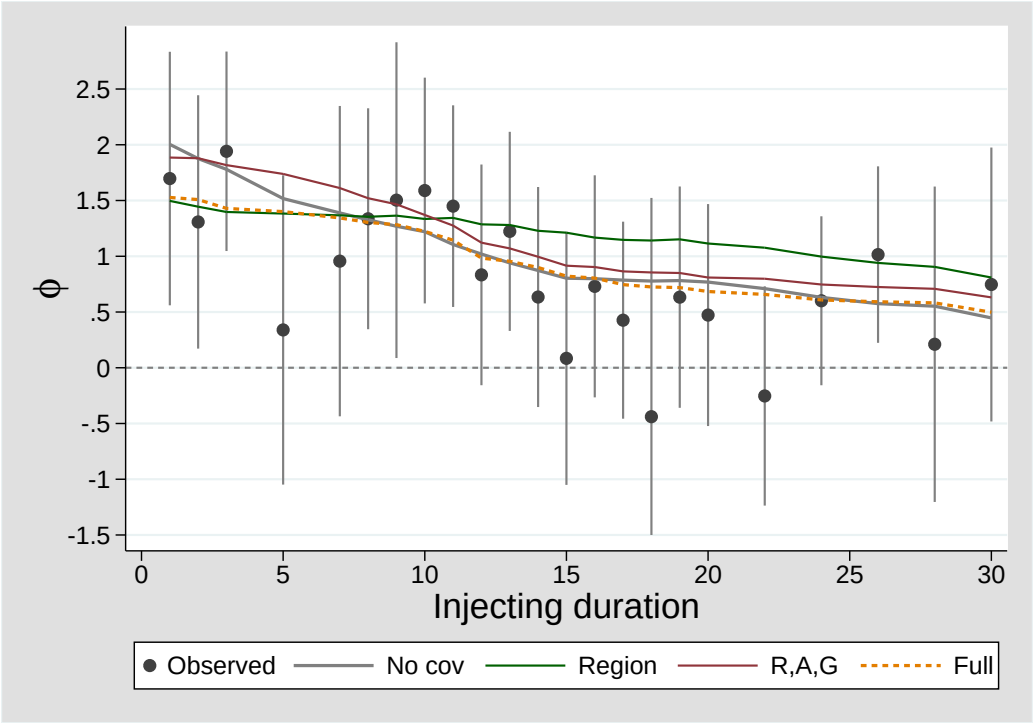


Figure 6.17: Observed and predicted estimates of association between HBV and HIV under different covariate models under the time-varying frailty model. *R,A,G*=region, age and gender. Injecting durations are grouped where data are sparse to prevent zero cells.

no covariate model. Under the full model the pattern is shifted again, with similar estimates as the region-only model for injecting durations less than 10 years, but a lower predicted association at longer durations.

### 6.3.2 Covariate effects

The inclusion of covariates altered the estimated trend over calendar time for HCV and HBV, largely due to the inclusion of region as a covariate (Figure 6.18). For HCV there was a smoother U-shape in the post-1980 period: the FOI falls from a peak in 1980-1985 to its lowest point in 1995-2000 before rising again. Without covariates, there is a sharper dip in 1995-2000 and a far lower FOI in the pre-1980 period. For HBV the fall in FOI over time is very steep in any case, but including region suggests that the fall in risk from 1990-1995 to post-2010 is slightly less than the estimates without covariates.

Region is a very important factor, as shown in chapter 4, and crucial

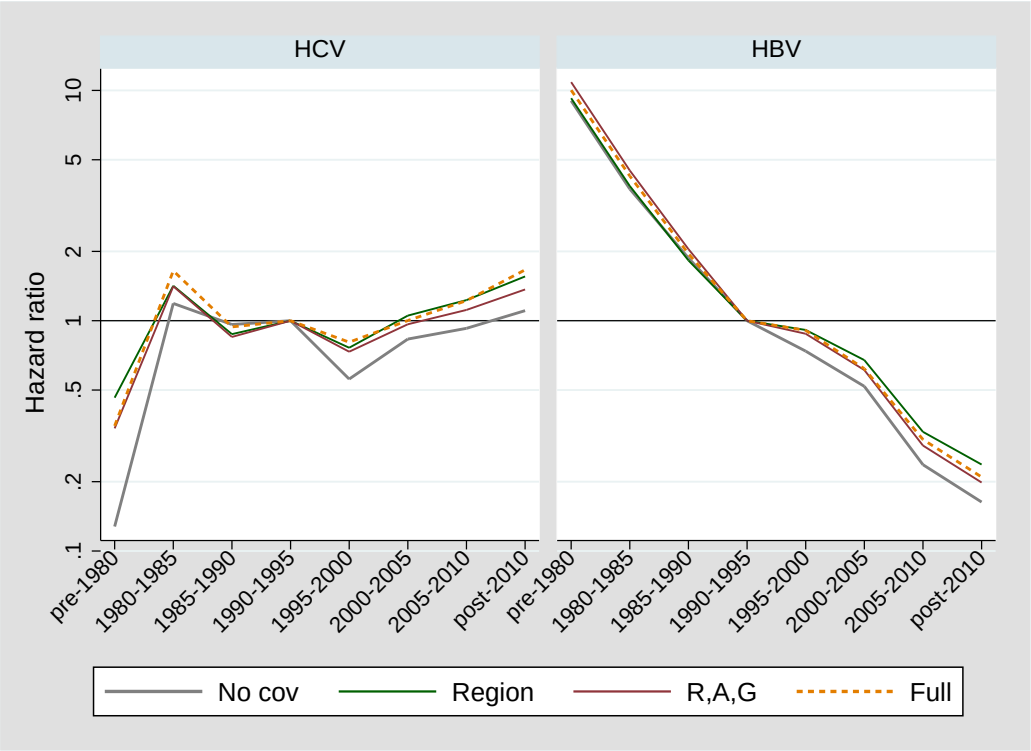


Figure 6.18: Estimated temporal trends for HCV and HBV under different covariate models with inverse Gaussian frailty, hazard ratios for 5-year survey periods vs. 1990-1995. No covariates; region; region, age and gender (R,A,G) and full model (all the above plus risk factor information).

to understanding trends over time due to changing patterns of the injecting epidemic and the sampling frame of the UAM data. Figure 6.19 shows the proportions of individuals at risk over time across regions. Of those sampled by the survey who were injecting in the 1970s, a large proportion were in the London region ($35-40\%$), but this proportion has declined steadily over time, to 10% or less since 2010. The proportion of those injecting in the North West grew steadily from 1980 to 1995, from 10% to 21%, but declined subsequently to 10% or less since 2010. These areas have markedly higher prevalence of BBVs, and if region is not adjusted for then estimated trends over time will be distorted by the changing composition of the sample. This is seen in the estimated trends, with the underlying risk increasing more noticeably for HCV and declining less markedly for HBV once region is adjusted for. The declining frailty variance observed in Figure 6.16 may also be due to changes in sampled regions; as the high-risk areas of London and

the North West are proportionally less-represented in the sample over time, the population may have become more homogeneous. Having accounted for regional effects in the analysis, the variability of residual frailty declines less over time. Therefore the predicted RFV in Figure 6.16 declines under the gamma model, despite the gamma distribution having constant RFV.

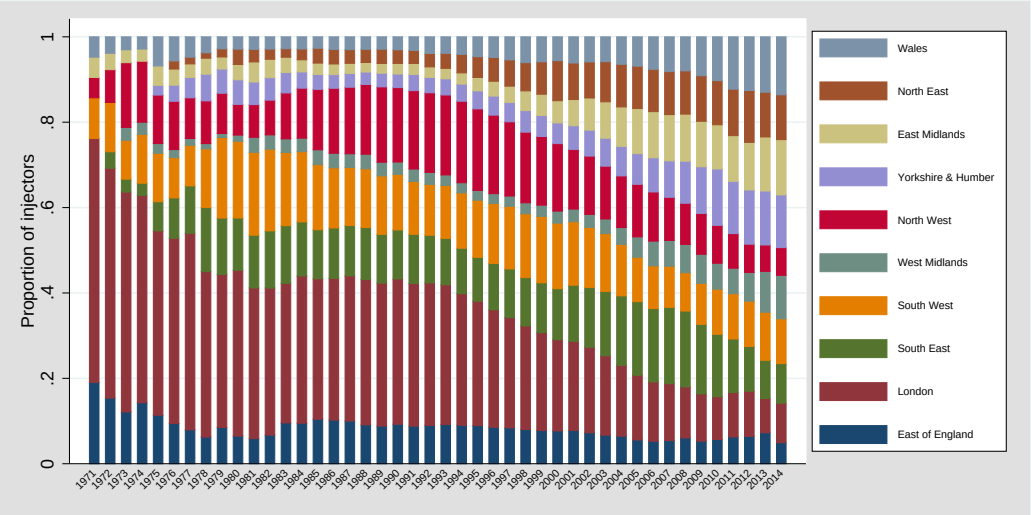


Figure 6.19: Proportion of injectors across regions over time, from those participating in the UAM survey from 2000 onwards with complete covariate information. The injecting career length of each participant is used to calculate the numbers injecting in 1971, 1972, etc. in each region. Bars are stacked in the same order that they appear in the legend on the right.

Figure 6.20 shows the estimated temporal trend from the full covariate model under different frailty distributions. The choice of frailty distribution has little impact on the estimated trend, but all trends with frailty differ from that under independence, particularly in early years. For HCV, the pre-1980 risk is much lower under the frailty distribution but higher for 1980-1985, which is epidemiologically plausible. For HBV, the decrease in risk over time is somewhat steeper under the frailty models.

Patterns of injecting duration-specific risk for HCV and HBV were not changed markedly by the inclusion of different covariates, and the estimated HRs for the covariate effects were similar across different frailty distributions. The best-fitting model is therefore focussed on, which included all covariate effects and an inverse Gaussian frailty distribution. Similarly, the estimated HRs for HIV were fairly consistent for different frailty distribu-

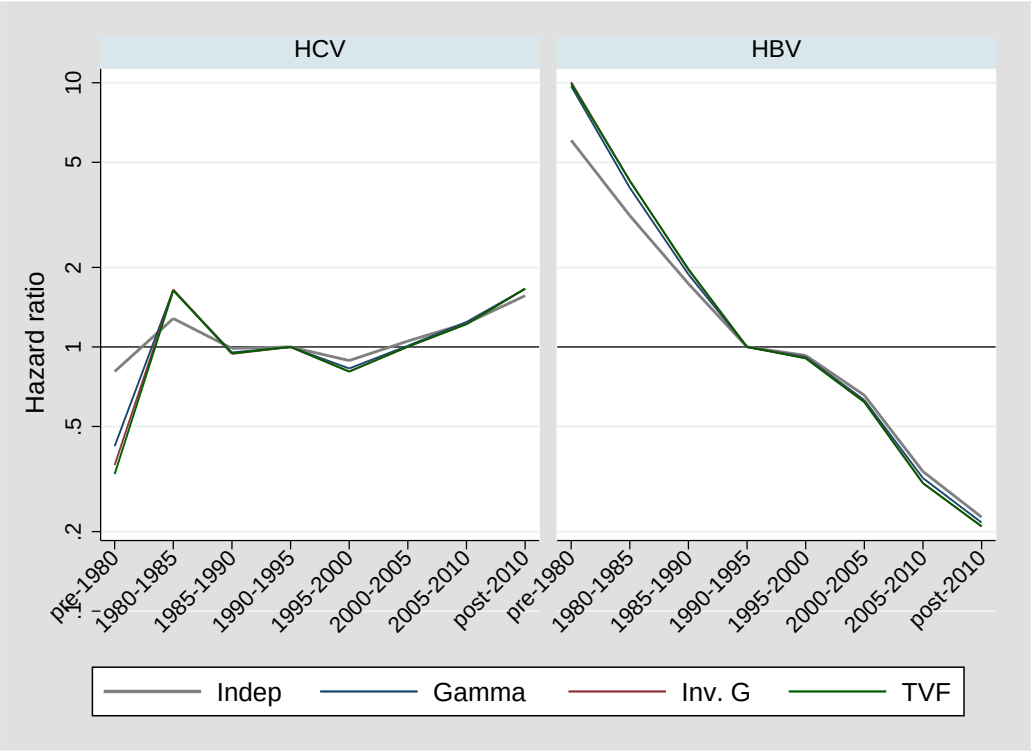


Figure 6.20: Estimated temporal trends under different frailty models with all covariates, hazard ratios for 5-year survey periods vs. 1990-1995. Independence, gamma, inverse Gaussian (Inv. G) and time-varying frailty (TVF) distributions.

tions, adjustment for other covariates, and for both HCV-HIV and HBV-HIV pairings. Therefore HIV results are taken from the HBV-HIV model, also with inverse Gaussian frailty for consistency. Estimated hazard ratios and 95% confidence intervals from these models are shown in Table 6.9.

The patterns of estimated HRs, in terms of higher/lower risk for different regions, ages of first use, gender and risk factors, are very similar to the estimated odds ratios from generalised linear models of prevalence in section 4.1.4. Briefly, London has a higher FOI for HCV and HBV, and much higher for HIV; and the North West has a higher FOI for HCV, and particularly high for HBV. Older age at first use is associated with a higher FOI for HCV and HBV, but younger age at first use for HIV. The interesting additions are the time-varying aspects of risk factors: ever-use of needle exchange is possibly associated with an increased FOI for HCV (HR=1.18, 95% CI: 0.94-1.49) but the period following first use also has a higher FOI (HR=1.19, 95%

Table 6.9: Hazard ratios for covariates for HCV, HBV and HIV under bivariate inverse Gaussian frailty models; HCV-HBV estimates and HIV estimates from HBV-HIV bivariate model. *NX*: needle exchange.

| | HCV | HBV | HIV |
|---|---|---|---|
| *Region* | | | |
| East England | 0.58 (0.50, 0.67) | 0.82 (0.63, 1.07) | 0.84 (0.39, 1.81) |
| London | 1.35 (1.21, 1.50) | 1.28 (1.07, 1.53) | 4.43 (2.87, 6.84) |
| South East | 1 (ref) | 1 (ref) | 1 (ref) |
| South West | 0.56 (0.50, 0.63) | 0.70 (0.55, 0.89) | 0.58 (0.31, 1.10) |
| West Midlands | 0.51 (0.44, 0.58) | 0.42 (0.33, 0.53) | 0.40 (0.15, 1.06) |
| North West | 1.70 (1.53, 1.90) | 2.04 (1.67, 2.48) | 1.37 (0.81, 2.30) |
| Yorskshire & Humber | 1.04 (0.90, 1.21) | 0.53 (0.38, 0.73) | 0.09 (0.01, 0.64) |
| East Midlands | 0.88 (0.78, 0.99) | 0.80 (0.65, 0.97) | 0.93 (0.51, 1.72) |
| North East | 0.46 (0.41, 0.52) | 0.77 (0.61, 0.99) | 0.29 (0.12, 0.71) |
| Wales | 0.47 (0.40, 0.54) | 0.62 (0.49, 0.77) | 0.41 (0.17, 1.03) |
| *Demographics* | | | |
| Age first used $< 18$ | 0.94 (0.87, 1.01) | 0.80 (0.71, 0.91) | 1.75 (1.29, 2.39) |
| Age first used 18-24 | 1 (ref) | 1 (ref) | 1 (ref) |
| Age first used 25+ | 1.32 (1.23, 1.42) | 1.54 (1.38, 1.72) | 1.21 (0.84, 1.73) |
| Female vs. male | 1.38 (1.28, 1.48) | 1.22 (1.08, 1.37) | 0.78 (0.55, 1.09) |
| *Risk factors* | | | |
| Ever rec'd works | 1.56 (1.47, 1.65) | 1.23 (1.10, 1.38) | 1.56 (1.18, 2.06) |
| Ever used NX | 1.18 (0.94, 1.49) | 0.90 (0.67, 1.19) | 1.59 (0.73, 3.49) |
| Post-first NX | 1.19 (1.02, 1.37) | 0.90 (0.78, 1.02) | 0.57 (0.35, 0.92) |
| Ever imprisoned | 1.48 (1.27, 1.73) | 1.20 (0.93, 1.55) | 0.95 (0.54, 1.65) |
| Post-first prison | 1.32 (1.13, 1.54) | 1.17 (0.93, 1.49) | 0.75 (0.41, 1.37) |
| MSM | 1.07 (0.90, 1.27) | 1.21 (0.95, 1.54) | 4.01 (2.56, 6.28) |

CI: 1.02-1.37). For HBV, both ever using a needle exchange and post-first use period have a reduced FOI, although the 95% CI for the HR crosses 1. For HIV, there is a possible, but uncertain, increase in FOI for those ever using a needle exchange, but reduced risk following first use (HR=0.57, 95% CI: 0.35-0.92). For imprisonment, the FOI for both ever being imprisoned and the period following imprisonment is increased for HCV, somewhat less for HBV, and shows little effect for HIV. As for the analysis in section 4.1.4, ever receiving works is associated with an increased risk of all BBVs, and MSM a substantially increased risk of HIV infection.

### 6.3.3 Age, gender and individual frailty

In this section patterns of HCV and HBV infection by age and gender are further explored. The base model considered here includes injecting duration and time (without interaction), an overall region effect (vs. the South East) and examines age at first use ($< 18$, $18 - 24$ and $25+$), females vs. males and an interaction between age at first use and gender. Multiplicative covariate effects are assumed and an inverse Gaussian distribution used to model individual variability.

Despite fairly strong differences in risk according to age at first use and gender, models fit was not improved by including interactions between them; in other words, the effect of age at first use was similar for males and females. There was also no evidence of an interaction between gender and injecting duration, with little improvement in model fit, despite the additional parameters. There was however evidence of a difference in injecting duration-specific risk according to age at first use, with a reduction of 5.7 in AIC under the latter. Examination of the interaction terms showed that the strongest effect was in those starting injecting at age 25 or older, with an excess risk in the first year but a broadly comparable FOI subsequently. A simplified model including only interaction terms for the first year risk of injecting provided a similar AIC score, with an increase of just 0.4 vs. the more complex model.

In section 6.1 there were apparent differences in the measure of association, $\Phi$, according to gender and possibly age at first use. The base model was extended to incorporate separate frailty variances for males and females, and

for the three categories of age at first use. For the former, a somewhat higher frailty variance was estimated for males ($\delta = 1.64$) vs. females ($\delta = 1.17$), compared to an overall frailty variance of ($\delta = 1.52$). AIC scores indicated a modest improvement under the stratified frailty variance model, with a decrease of 3.0 under stratified frailty variances. A similar model was fitted with separate frailty variances for those aged $< 18$, $18 - 24$ and $25+$ at first use, giving frailty variances of 1.35, 1.68 and 1.47 respectively. AIC scores indicated that there was no benefit to the additional parameters however, with an *increase* of 2.1 for the more complex model.

### 6.3.4 Conclusions

The inclusion of covariates reduced the frailty variance, which represents residual heterogeneity. Although model fit is enormously improved, the reduction in residual heterogeneity is relatively modest, indicating substantial variability in risk that is not captured by the covariates. A wealth of questionnaire data is available in the UAM data, but it seems that there are underlying differences in risk that are not easily captured.

The inclusion of covariates may have altered the shape of the residual frailty distribution, with the inverse Gaussian model being preferred over the TVF model for HCV-HBV, whereas the more complex model was favoured without covariates. Model predictions are virtually identical for the inverse Gaussian and TVF models, and estimates of $\Phi$ do not match the observed pattern as well as the models without covariates in section 6.2.1. This might be partly an identifiability problem: the covariate information is more easily identifiable than the correlation between infections, with the latter potentially being distorted by the inclusion of covariates. The measure of association is based on the 2x2 tables of bivariate status, with the proportion that are positive for both (++) typically small, especially at short injecting durations where the RFV is changing most rapidly. Therefore underestimates of the ++ cell (resulting in a weaker estimated association) may have less impact on maximising the likelihood in comparison to fitting to the mean prevalence according to different covariates. In other words, the univariate fit of each infection (or just one if the second is sparse) may take precedence.

Shifts in the estimated frailty distribution also occur for HBV-HIV. When

region alone is added to the model, the resulting estimates of $\Phi$ do not match the observed pattern well, although once age and sex are also added to the model, the pattern is a better match. In general, estimates of the shape of the frailty distribution appear quite sensitive to the inclusion of covariates.

The estimated hazard ratios for covariates are generally comparable to the odds ratios obtained under logistic regression models in section 4.1.4. The function that links the covariates to the mean of the response variable differs, but ultimately the two models are both estimating ratios of infection prevalence at different levels of the covariate. In order to uncover more detail with regard to the effect of risk factors on the FOI, covariate interactions or time-varying covariates are required. However, there is limited scope for increasing model complexity due to sparsity of data.

The time-varying covariates here provided some interesting but unusual results, being parameterised in terms of ever-occurrence and change in risk post-first occurrence. For needle exchange use, there is a different pattern for each infection: increased risk for ever and post-first use for HCV, a slight decrease for both for HBV, and increased risk for ever use but decreased risk post-first use for HIV. It is somewhat difficult to make sense of these results, with an increase in the risk of HCV infection following use of harm reduction services, and inconsistencies in the direction of effects across infections. Possibly they are the result of the combination of the propensity for injecting and sexual risk, the correlations between these factors and their relative importance for each infection; or, there may be no sensible epidemiological interpretation. Both ever and post-first imprisonment are associated with an increased risk of HCV infection, slightly increased risk of HBV infection, but little association for HIV. If the relative importance of injecting to sexual risk is high for HCV, sexual risk somewhat important for HBV and more important than injecting risk for HIV, and prison is a high-risk environment for injecting but not for sexual transmission, this may be plausible. Further, individuals that have ever been imprisoned may have lower sexual risks and vice versa; a purely speculative idea, but one that would explain the attenuation of the risk associated with imprisonment for HBV and HIV.

Further investigation of age at first use and females vs. males confirmed that risk of HCV and HBV infection in males is more variable than that of females. This was indicated by plots of the association between infections in

section 6.1 but the inclusion of a stratified frailty variance, while accounting for relevant covariate information, provides a basis for statistical hypothesis testing. Further exploration of patterns of RFV might be considered in this way (for instance, via group-specific parameters in time-varying frailty models) but are likely to be hampered by a lack of information.

An interaction between age at first use and injecting duration was also noted, in particular an apparent excess risk in the first year of injecting for those starting at older ages. This may reflect the different circumstances in which individuals start injecting, with more comparable levels of risk post-initiation. However, another explanation is misreporting of age at first use, as these results would also be consistent with a portion of individuals erroneously giving their current age as the age they first injected, and in fact having been at risk for longer.

## 6.4 Concluding remarks

This chapter has shown that the inclusion of demographic and risk factor information as covariates reduces residual individual heterogeneity in the risk of infection in people who inject drugs, but the effect is modest. Despite identifying several risk factors that are strongly predictive, there is still substantial variability in the risk of infection. The data clearly point to a declining RFV, such that the pool of uninfected individuals become more homogeneous over time. Under the inverse Gaussian model, this is attributed to a selection effect, while the TVF model assumes the decline is due to changes in the underlying risk of individuals. The TVF model provides more flexibility and in the absence of covariates improves model fit, although still does not quite capture the patterns of relative frailty variance for HCV-HBV. With the inclusion of covariates, simpler frailty distributions are preferred. However, rather than this being due to a simpler distributional structure (whose variance does not decline over time) it may be that the estimated frailty distribution is being distorted by the presence of covariates, as the estimate association between infections sometimes provided a poorer approximation of the pattern of RFV when additional covariate information were included.

A full understanding of individual heterogeneity in people who inject

drugs requires appropriate modelling of the different routes of transmission, incorporating both injecting and sexual risk. People who inject drugs may have high sexual risks (BMA Board of Science, 2013, Chapter 4) due to decreased inhibitions (both a cause and consequence of drug use) and a general variability in levels of inhibition could result in a correlation between injecting and sexual risk. In order to identify the sexual component, or any structural form for the frailty variance that involves different components of risk (i.e., beyond the shared frailty model) requires the use of trivariate data. The HIV data are however very sparse and therefore the identification of complex correlation structures is challenging.

One way to make progress would be to either simplify the baseline FOI or place constraints on model parameters, an example of the latter being the shared injecting duration-specific hazard ratios explored in section 6.2.2. Such constraints will generally take the form of assumptions: there is no evidence against a common injecting duration-specific risk for HBV and HIV, but it is only considered as a modelling possibility because data are insufficient to distinguish between different parameter estimates for the two infections in the first place. Estimation is then conditional on the assumed model structure - as with any statistical model of course, but the assumptions here cannot be verified by the data. In section 6.2.2, the estimated frailty variance was altered by the assumptions of shared parameters, and indeed, the estimated frailty, particularly the TVF model, sensitive to covariates and other model structure where data are sparse. The situation is similar in ways to the estimation of frailty in a univariate model, which is only detectable via divergence from an assumed model structure. In the bivariate case, under a collection of covariates which may or may not be shared, estimates of the frailty distribution may be influenced by divergence from the assumed model structure in unpredictable ways.

# Chapter 7

# Trivariate models and components of risk

## 7.1 Introduction

The extension from univariate to bivariate analysis considered in chapters 5 and 6 allows for individual heterogeneity to be estimated and a more detailed investigation of the infection process. Trivariate data allow yet more complex structures to be considered, in terms of different aspects of heterogeneity in individual risk for different infections. A key consideration for HCV, HBV and HIV in people who inject drugs (PWID) is that injecting risk is common to all infections, but there is potentially an additional component of sexual risk for HBV and HIV, which is negligible for HCV. There is also the more general issue that individual variability may have components that are not wholly shared by all three infections, for instance, variation in biological susceptibility.

Components of individual heterogeneity in survival analysis have largely been considered within the framework of *correlated gamma frailties*: the correlation in frailties for different outcomes occurring due to there being a common component for all outcomes, but with additional components specific to each outcome. For current status data this presents a problem: as with shared frailty models, heterogeneity is only identified via the correlation between infections, and components that are specific to a single infection cannot be identified at any single time point. Models in which frailty com-

ponents are common to at least two infections are therefore required, as parameters for the variance components can be identified directly from the correlation structure of the data.

This chapter is organised as follows. The correlated gamma frailty model is reviewed, and two forms of additive model based on shared components are proposed, based on the correlated frailty model. Multiplicative component structures are then considered. For both types of model, simplifications of the full (saturated) model are likely to be required in practice, and possible sub-models are explored. Some epidemiological considerations are then discussed in terms of suitable models for trivariate data on blood-borne infections in PWID. The proposed models are then fitted to the UAM data.

## 7.2 Correlated frailties and component frailty models

### 7.2.1 The correlated gamma frailty model

A limitation of the shared frailty model is that individual heterogeneity is assumed to have an identical effect on the risk of each infection, and (equivalently) that there are no aspects of individual frailty that are specific to one infection; or at least, any such factors are not considered in the analysis. This is a necessary assumption unless the baseline hazard is parameterised to have a certain shape, from which any deviation would be assumed to arise from some component of individual frailty (Farrington et al., 2001).

A more flexible form of frailty model is the *correlated frailty model*, in which the frailties for each infection are positively correlated, but not identical (Yashin et al., 1995). This is achieved by specifying additive components, usually with gamma distributions, with the overall frailty for each infection being the sum of a shared component, plus individual frailty components that are specific to each infection. This approach has generally been applied to bivariate data, but could readily be extended to higher-order multivariate data.

For general survival analysis models, the bivariate survivor function for

the shared gamma frailty model with frailty variance $\sigma^2$ can be written as

$$S(t_1, t_2) = \left( S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1 \right)^{-\frac{1}{\sigma^2}}$$

(see for example, Wienke et al. (2005)). The bivariate correlated frailty model described by Yashin et al. (1995) is defined in terms of additive components. The components $Y_i$ $(i = 0, 1, 2)$ are independent random variables that have gamma distributions with scale parameter 1, and $Z_i = \sigma_i^2(Y_0 + Y_i)$ for $i = 1, 2$. $Y_0$ therefore corresponds to the shared component of individual frailty and $Y_1$, $Y_2$ the non-shared components, while the $\sigma_i$ terms control the variance. With $k_0$, $k_1$ and $k_2$ positive parameters, $Y_i \sim \Gamma(k_i, 1)$ for $i = 0, 1, 2$, $\sigma_i^2 = (k_0 + k_i)^{-1}$ for $i = 1, 2$ and $\rho = k_0[(k_0 + k_1)(k_0 + k_2)]^{-\frac{1}{2}}$, the bivariate survivor function proposed by Yashin et al. (1995) is

$$S(t_1, t_2) = S_1(t_1)^{1-\rho\sigma_1/\sigma_2} S_2(t_2)^{1-\rho\sigma_2/\sigma_1} [S_1(t_1)^{-\sigma_1^2} + S_2(t_2)^{-\sigma_2^2} - 1]^{-\rho/(\sigma_1\sigma_2)}, \tag{7.1}$$

The shared frailty model is therefore a special case of the correlated model where $\rho = 1$, or equivalently $k_1 = k_2 = 0$, and individual frailty is identical for both infections.

The bivariate correlated frailty model may readily be extended to a trivariate or higher order model, and a general formulation is given in Yashin et al. (1995) for the multivariate survivor function, which can be written as

$$S(t_1, ..., t_n) = \left( \sum_{i=1}^{n} S_i(t_i)^{-\frac{1}{k_0+k_i}} - n + 1 \right)^{-k_0} \prod_{i=1}^{n} S_i(t_i)^{1-k_0/(k_0+k_i)}, \tag{7.2}$$

where $k_0$ is the inverse of the variance of the gamma frailty component common to all outcomes and $k_i$ for the components specific to each outcome $i$.

### 7.2.2 The correlated gamma frailty model for current status data

For current status data observed at a single time point $t$, equation 7.2 can be written in terms of the cumulative hazard functions $A_i(t)$ for infection $i$

as:

$$S_{1,\ldots,n}(t) = \left(1 + \sum_{i=1}^{n} \frac{A_i(t)}{k_0 + k_i}\right)^{-k_0} \prod_{i=1}^{n} \left(1 + \frac{A_i(t)}{k_0 + k_i}\right)^{-k_i}. \tag{7.3}$$

Having defined the multivariate survivor functions for up to $n$ infections, this can be used as the basis of deriving the $2^n$ table of probabilities for each possible infection status in a similar way as the bivariate case, as in equation 5.22. This chapter is primarily concerned with models for trivariate data, in which case we define univariate, bivariate and trivariate survivor functions with cumulative hazards $A_i(t)$ at time $t$ for infection $i$ as:

$$\begin{aligned}
S_1(t) &= \left(1 + \frac{A_1(t)}{k_{123} + k_1}\right)^{-(k_{123}+k_1)} \\
S_{12}(t) &= \left(1 + \sum_{i=1}^{2} \frac{A_i(t)}{k_{123} + k_i}\right)^{-k_{123}} \prod_{i=1}^{2} \left(1 + \frac{A_i(t)}{k_{123} + k_i}\right)^{-k_i} \\
S_{123}(t) &= \left(1 + \sum_{i=1}^{3} \frac{A_i(t)}{k_{123} + k_i}\right)^{-k_{123}} \prod_{i=1}^{3} \left(1 + \frac{A_i(t)}{k_{123} + k_i}\right)^{-k_i},
\end{aligned} \tag{7.4}$$

where $k_{123}$ is the inverse of the variance of the component shared by all three infections and $k_i$ those specific to infection $i = 1, 2, 3$. $S_2(t)$, $S_3(t)$, $S_{13}(t)$ and $S_{23}(t)$ are defined similarly, and the probabilities of trivariate infection status $p_{abc}(t)$ with $a = 0, 1$ for infection 1, $b = 0, 1$ for infection 2 and $c = 0, 1$ for infection 3 at time $t$ are:

$$\begin{aligned}
p_{000}(t) &= S_{123}(t) \\
p_{100}(t) &= S_{23}(t) - p_{000}(t) \\
p_{010}(t) &= S_{13}(t) - p_{000}(t) \\
p_{001}(t) &= S_{12}(t) - p_{000}(t) \\
p_{110}(t) &= S_3(t) - p_{000}(t) - p_{100}(t) - p_{010}(t) \\
p_{101}(t) &= S_2(t) - p_{000}(t) - p_{100}(t) - p_{001}(t) \\
p_{011}(t) &= S_1(t) - p_{000}(t) - p_{010}(t) - p_{001}(t) \\
p_{111}(t) &= 1 - p_{000}(t) - p_{100}(t) - p_{010}(t) - p_{001}(t) - p_{110}(t) - p_{101}(t) - p_{011}(t)
\end{aligned} \tag{7.5}$$

Equation 7.5 also holds for models that involve both age and time and any other covariates in the cumulative hazard functions featuring in equation 7.4. The general approach of combining survivor functions for 1,2,...$n$ infections to obtain the $2^n$ cell probabilities can readily be extended to higher-order multivariate data.

Correlated frailty models have been shown to be identifiable for bivariate event times (Iachine, 2004), although the information contained in current status data is weaker: at a single time point, frailty can only be estimated via the correlation between infections. Therefore any component of variability attributable to a single infection cannot be reliably identified without fairly strong assumptions. From the perspective of statistical modelling of bivariate current status data, given an unconstrained hazard function and some specified frailty distribution for the common frailty component, an improvement in model fit might still be achieved via the addition of infection-specific components. It may be tempting to interpret such an improvement as evidence of infection-specific components of frailty, but the improvement could also arise from particular selection effects or time-varying aspects of the underlying frailty that are not captured by the assumed functional form. Even if the functional form is correct, the amount of information available to estimate frailty components is likely to be very low.

### 7.2.3 Shared component models

For current status data, information on individual frailty should ideally be identifiable solely from the correlation structure of the data at a single time point, thereby avoiding the need to gain information on individual frailty via distortions in the baseline hazard function(s). For a $2 \times 2$ table of bivariate infection status at a single time point, there are three degrees of freedom: given the total sample size, once three cells are known the fourth is determined. Therefore the two cumulative hazards may be estimated, with one degree of freedom remaining to estimate the correlation between the two infections, and thereby the frailty variance.

Given the 8 cells of the $2 \times 2 \times 2$ table for trivariate data, there are 7 degrees of freedom with which to estimate parameters. Three cumulative hazards need to be estimated, leaving 4 degrees of freedom for the estimation

of the correlation (frailty) structure. As components cannot be specific to a single infection, there is a unique saturated model consisting of frailty components for each pair of infections, plus a frailty parameter that acts on all three infections. This is conceptually similar to log-linear modelling of three binary covariates, in which the baseline is the sum-to-one constraint, the cumulative hazards are main effects, and frailty parameters are interaction terms (three 2-way, one 3-way).

Assuming that additive gamma frailty components are to be considered, where the overall frailty for each infection consists of a weighted sum of variables with a gamma distribution, the general form for the trivariate saturated component frailty model is

$$\begin{aligned} Z_1 &= w_{123}Y_{123} + w_{12}Y_{12} + w_{13}Y_{13} \\ Z_2 &= w_{123}Y_{123} + w_{12}Y_{12} + w_{23}Y_{23} \\ Z_3 &= w_{123}Y_{123} + w_{13}Y_{13} + w_{23}Y_{23}, \end{aligned} \tag{7.6}$$

where the 123 subscript refers to components common to all three infections and 12, 13 and 23 the pairwise components; $w_{123}$, $w_{12}$, $w_{13}$ and $w_{23}$ are real numbers corresponding to the weight of each frailty component, and $Y_{123}$, $Y_{12}$, $Y_{13}$ and $Y_{23}$ are independent random variables with a gamma distribution. This general form requires some constraints to be usable in practice, namely that the $w$ terms and random variables combine such that $Z_1$, $Z_2$ and $Z_3$ each have unit mean. Two possible specifications that achieve this are (1) to use gamma variables with unit mean and set the sum of the weights to equal one, and (2) to use gamma variables with scale parameter 1 and use properties of sums of such gamma variables to derive another gamma distribution that has mean one. Both models are *additive shared component* models, and the first is referred to as the *fixed weight* model and the second the *variable weight* model. The reason that one or the other must be chosen is that there is not sufficient information to estimate both the variances of the random variables and the weight terms, again assuming that the information from time-variation in the correlation structure is likely to be weak and unreliable for this purpose and we wish to be able to estimate the frailty structure from a single time point. Therefore if the variances are unconstrained the weight terms must be fixed. The variable weight model is a

modification of the approach used for the correlated frailty model described in section 7.2.2, where the weights and variances are linked by the same parameters. Properties of each of these models are discussed subsequently.

### 7.2.4 Fixed weight component model

In the fixed weight component model $Y_{123} \sim \Gamma(k_{123}, k_{123})$, $Y_{12} \sim \Gamma(k_{12}, k_{12})$, $Y_{13} \sim \Gamma(k_{13}, k_{13})$ and $Y_{23} \sim \Gamma(k_{23}, k_{23})$. The four $k$ parameters for the shape and scale of the component gamma distributions with mean one are the inverses of the variances of each frailty component, for example $\delta_{123} = \frac{1}{k_{123}}$. These four parameters use all of the available degrees of freedom for the correlation structure, and therefore the weight parameters $w$ must be set at fixed values.

Although the weight terms are not strictly identifiable, there may still be aspects of the data that would result in a better or worse fit for certain combinations of $w$ terms. Firstly, as already mentioned, the way in which the correlation structure evolves over time might provide weak information on the relative importance (weights) of each component. These would be subtle differences, as the difference between a component with low variance and high weight and one with high variance and low weight is likely to be slight. Again, this would also require that the frailty distributions themselves are of the correct form and not inherently time-varying. Secondly, component models in general require all of the correlations to be positive, as frailty variances must be non-negative. Although not usually considered as a potential issue for the shared frailty model as infections do not usually show a negative correlation, the saturated model in equation 7.6 requires that each pairwise correlation and the three-way (or "residual") correlation are all non-negative. This means that the saturated model does not guarantee a perfect fit to the data, even at a single time point, and this could result in one or more of the weight terms being estimated at zero in order to accommodate the negatively-correlated component. However, such issues would point to the model being generally incorrect, and not a good basis for obtaining estimates of the weight terms.

In the absence of any external information, a natural choice for the weight terms is just to set all of them equal to $\frac{1}{3}$. A priori information might

indicate that the component shared by all infections, $Y_{123}$ has greater or lesser importance than the others; for instance $w_{123} = 0.5$, $w_{12} = w_{13} = w_{23} = 0.25$, and different values could potentially be explored in sensitivity analyses. It is of course possible to assign different weight terms to each component, and potentially different contributions for a component to the risk of each infection, although the expert knowledge required for appropriate choices is likely to be lacking.

Assuming for now that $w_{123} = w_{12} = w_{13} = w_{23} = \frac{1}{3}$, the overall frailty terms for infections 1, 2 and 3 are

$$\begin{aligned} Z_1 &= (Y_{123} + Y_{12} + Y_{13})/3 \\ Z_2 &= (Y_{123} + Y_{12} + Y_{23})/3 \\ Z_3 &= (Y_{123} + Y_{13} + Y_{23})/3. \end{aligned} \tag{7.7}$$

The variances of the overall frailties (specified in terms of their component variances) are

$$\begin{aligned} \text{var}(Z_1) &= (\delta_{123} + \delta_{12} + \delta_{13})/9 \\ \text{var}(Z_2) &= (\delta_{123} + \delta_{12} + \delta_{23})/9 \\ \text{var}(Z_3) &= (\delta_{123} + \delta_{13} + \delta_{23})/9. \end{aligned} \tag{7.8}$$

With $A_i(t)$ the baseline cumulative hazard function for infection $i$, the trivariate survivor function, bivariate survivor function for infections 1 and 2 and univariate survivor function for infection 1 are:

$$\begin{aligned} S_{123}(t) = &\left(1 + \frac{A_1(t) + A_2(t) + A_3(t)}{3k_{123}}\right)^{-k_{123}} \left(1 + \frac{A_1(t) + A_2(t)}{3k_{12}}\right)^{-k_{12}} \\ &\left(1 + \frac{A_1(t) + A_3(t)}{3k_{13}}\right)^{-k_{13}} \left(1 + \frac{A_2(t) + A_3(t)}{3k_{23}}\right)^{-k_{23}} \end{aligned}$$

$$\begin{aligned} S_{12}(t) = &\left(1 + \frac{A_1(t) + A_2(t)}{3k_{123}}\right)^{-k_{123}} \left(1 + \frac{A_1(t) + A_2(t)}{3k_{12}}\right)^{-k_{12}} \\ &\left(1 + \frac{A_1(t)}{3k_{13}}\right)^{-k_{13}} \left(1 + \frac{A_2(t)}{3k_{23}}\right)^{-k_{23}} \end{aligned}$$

$$S_1(t) = \left(1 + \frac{A_1(t)}{3k_{123}}\right)^{-k_{123}} \left(1 + \frac{A_1(t)}{3k_{12}}\right)^{-k_{12}} \left(1 + \frac{A_1(t)}{3k_{13}}\right)^{-k_{13}}. \quad (7.9)$$

The bivariate survivor functions for pairs 1-3 and 2-3 and univariate functions for infections 2 and 3 are defined similarly. The probabilities for trivariate infection status are then defined as in equation 7.5.

It is important to note that the shared frailty model is not a special case of this model when the variances of the pairwise components are zero. If the variances of the pairwise components in equation 7.6 are zero then the overall frailties become

$$\begin{aligned} Z_1 &= w_{123}Y_{123} + w_{12} + w_{13} \\ Z_2 &= w_{123}Y_{123} + w_{12} + w_{23} \\ Z_3 &= w_{123}Y_{123} + w_{13} + w_{23}, \end{aligned} \quad (7.10)$$

such that $Z_1$, $Z_2$ and $Z_3$ have limited support and can only take values in the range $(1 - w_{123}, \infty)$. This model may therefore have a tendency to provide non-zero variance estimates for the pairwise components, even if individual variability genuinely is predominantly related to a single component shared by all infections, as the potentially limited support may not be compatible with the correlation structure in the data. The parameters for the fixed weight model should therefore be interpreted as the relative importance of each component, rather than the amount of variation directly attributable to each component.

Also of note is that if $k_{123} = k_{12} = k_{13} = k_{23} = k$ and the weight terms are equal, then the sum of gamma distributions for each $Z_i$ have the form of another gamma distribution with variance $\frac{1}{3k}$. The univariate survivor function in equation 7.9 will therefore be equivalent to that of a shared frailty model; however, the bivariate and trivariate functions are not.

The fixed weight model is therefore an alternative to the shared frailty model which allows for non-perfect correlation in the frailties $Z_i$ for each infection, but with limited support for frailties that have low correlation in the case that the variance of one or more components approaches zero. The correlations between the $Z_i$ terms, which again may be expressed more

naturally in terms of their component variances, are

$$\begin{aligned}
\rho_{12} &= \left( \frac{(\delta_{123} + \delta_{12})^2}{(\delta_{123} + \delta_{12} + \delta_{13})(\delta_{123} + \delta_{12} + \delta_{23})} \right)^{\frac{1}{2}} \\
\rho_{13} &= \left( \frac{(\delta_{123} + \delta_{13})^2}{(\delta_{123} + \delta_{13} + \delta_{12})(\delta_{123} + \delta_{13} + \delta_{23})} \right)^{\frac{1}{2}} \\
\rho_{23} &= \left( \frac{(\delta_{123} + \delta_{23})^2}{(\delta_{123} + \delta_{23} + \delta_{12})(\delta_{123} + \delta_{23} + \delta_{13})} \right)^{\frac{1}{2}} .
\end{aligned} \tag{7.11}$$

It can be seen from equation 7.11 that the values of the variance components restrict the range of the correlations, although to a lesser degree than the correlated frailty model described in section 7.2.2.

Another consideration for this model is the variability of survivors and how the relative frailty variance (RFV) changes over time (Unkel and Farrington, 2012). Except for the special case where all of the components have equal variance and the weights are equal, the overall frailties $Z_1$, $Z_2$ and $Z_3$ will not have a gamma distribution and therefore do not have the property of constant RFV.

### 7.2.5 Variable weight component model

As mentioned previously the way in which the gamma components are defined for the correlated frailty model proposed by Yashin et al. (1995) specifies the weights of the general form for additive components in equation 7.6 via the $k$ terms. This approach also has the property that the resulting overall frailties $Z_i$ have gamma distributions. This approach therefore does not have the potentially limited support for the overall frailty of the fixed-weight approach described in section 7.2.4, and will preserve the property of constant RFV that the shared gamma frailty model has.

The components are expressed as independent gamma variates:

$$\begin{aligned}
Y_{123} &\sim \Gamma(k_{123}, 1) \\
Y_{12} &\sim \Gamma(k_{12}, 1) \\
Y_{13} &\sim \Gamma(k_{13}, 1) \\
Y_{23} &\sim \Gamma(k_{23}, 1)
\end{aligned} \tag{7.12}$$

and the overall frailties defined as

$$\begin{aligned} Z_1 &= \frac{1}{k_{123}+k_{12}+k_{13}}(Y_{123}+Y_{12}+Y_{13}) \\ Z_2 &= \frac{1}{k_{123}+k_{12}+k_{23}}(Y_{123}+Y_{12}+Y_{23}) \\ Z_3 &= \frac{1}{k_{123}+k_{13}+k_{23}}(Y_{123}+Y_{13}+Y_{23}). \end{aligned} \tag{7.13}$$

With $A_i(t)$ the baseline cumulative hazard function for infection $i$, the trivariate survivor function, bivariate survivor function for infections 1 and 2 and univariate survivor function for infection 1 are:

$$\begin{aligned} S_{123}(t) = &\left(1+\frac{A_1(t)}{k_{123}+k_{12}+k_{13}}+\frac{A_2(t)}{k_{123}+k_{12}+k_{23}}+\frac{A_3(t)}{k_{123}+k_{13}+k_{23}}\right)^{-k_{123}} \\ &\left(1+\frac{A_1(t)}{k_{123}+k_{12}+k_{13}}+\frac{A_2(t)}{k_{123}+k_{12}+k_{23}}\right)^{-k_{12}} \\ &\left(1+\frac{A_1(t)}{k_{123}+k_{12}+k_{13}}+\frac{A_3(t)}{k_{123}+k_{13}+k_{23}}\right)^{-k_{13}} \\ &\left(1+\frac{A_2(t)}{k_{123}+k_{12}+k_{23}}+\frac{A_3(t)}{k_{123}+k_{13}+k_{23}}\right)^{-k_{23}} \end{aligned}$$

$$\begin{aligned} S_{12}(t) = &\left(1+\frac{A_1(t)}{k_{123}+k_{12}+k_{13}}+\frac{A_2(t)}{k_{123}+k_{12}+k_{23}}\right)^{-(k_{123}+k_{12})} \\ &\left(1+\frac{A_1(t)}{k_{123}+k_{12}+k_{13}}\right)^{-k_{13}}\left(1+\frac{A_2(t)}{k_{123}+k_{12}+k_{23}}\right)^{-k_{23}} \end{aligned}$$

$$S_1(t) = \left(1+\frac{A_1(t)}{k_{123}+k_{12}+k_{13}}\right)^{-(k_{123}+k_{12}+k_{13})}. \tag{7.14}$$

The functions for $S_{13}$, $S_{23}$, $S_2$ and $S_3$ are expressed similarly, and the trivariate cell probabilities derived as in equation 7.5. These equations can be expressed more compactly by substituting $\sigma_i^2 = \frac{1}{k_{123}+k_{ij}+k_{ik}}$ as in equation 7.1, but writing in terms of $k$ terms illustrates the similarities and differences to the fixed weight model expressed in equation 7.9. In the case where $k_{123} = k_{12} = k_{13} = k_{23} = k$, this model is equivalent to the fixed weight model with $k_{123} = k_{12} = k_{13} = k_{23} = k$, as the gamma contributions have equal weights and variances.

The variable weight model has properties that may make it implausible for many applications. The variances of the overall frailties (specified in terms of their component variances) are

$$\begin{aligned}
\text{var}(Z_1) &= \frac{1}{1/\delta_{123} + 1/\delta_{12} + 1/\delta_{13}} \\
\text{var}(Z_2) &= \frac{1}{1/\delta_{123} + 1/\delta_{12} + 1/\delta_{23}} \\
\text{var}(Z_3) &= \frac{1}{1/\delta_{123} + 1/\delta_{13} + 1/\delta_{13}}.
\end{aligned} \tag{7.15}$$

Under the fixed weight model in section 7.2.4, the larger the variance components in equation 7.8, the larger the overall frailty variance. Under the variable weight model, if all the components are large then the frailty variance is also large, however smaller components will increase the denominator in equation 7.15 and thus limit the size of the overall frailty variance.

There are further constraints on the correlations between frailties for the different infections. The correlations between the $Z_i$ terms, which again may be expressed more naturally in terms of their component variances, are

$$\begin{aligned}
\rho_{12} &= \left( \frac{1/\delta_{123}^2 + 1/\delta_{12}^2}{(1/\delta_{123}^2 + 1/\delta_{12} + 1/\delta_{13})\,(1/\delta_{123}^2 + 1/\delta_{12} + 1/\delta_{23})} \right)^{\frac{1}{2}} \\
\rho_{13} &= \left( \frac{1/\delta_{123}^2 + 1/\delta_{13}^2}{(1/\delta_{123}^2 + 1/\delta_{12} + 1/\delta_{13})\,(1/\delta_{123}^2 + 1/\delta_{13} + 1/\delta_{23})} \right)^{\frac{1}{2}} \\
\rho_{23} &= \left( \frac{1/\delta_{123}^2 + 1/\delta_{23}^2}{(1/\delta_{123}^2 + 1/\delta_{12} + 1/\delta_{23})\,(1/\delta_{123}^2 + 1/\delta_{13} + 1/\delta_{23})} \right)^{\frac{1}{2}}.
\end{aligned} \tag{7.16}$$

Therefore pairs of infections with higher overall frailty variances will tend to have a lower correlation under the variable weight model, whereas the converse is true under the fixed weight model (equation 7.11). These are similar issues to those of the correlated frailty model described in section 7.2.2.

The fixed weight model therefore makes more sense from an epidemiological perspective and is easier to interpret: frailty components directly represent the variability in transmission risks that are shared between infections. The components of the variable weight model do not have such a natural interpretation, and infections with higher overall frailty variances

are less correlated than those with a smaller variance. Given that the frailty models considered here estimate components of frailty that are shared between infections, this is a rather unusual property and it is questionable whether such a model is useful.

### 7.2.6 Multiplicative component models

The two models consisting of additive gamma components proposed in sections 7.2.4 and 7.2.5 are restricted by the potentially limited support for the overall frailty components for the fixed weight model and restrictions on the correlation between frailties for the variable weight model. Fixed weight models are also not nested, so it is not possible to build up a sequence of models for comparison; for instance, two or more components compared to a shared frailty model. Models based on multiplicative frailties do not suffer these issues. The general form of a 2-component multiplicative frailty is:

$$Z = Y_1 Y_2, \tag{7.17}$$

where $Y_1$ and $Y_2$ are independent random variables with unit mean. Provided both variables have support over the range $(0, \infty)$ then $Z$ will also not be limited, and if $Y_1$ or $Y_2$ have zero variance then the model will naturally reduce to some simpler model that does not have limited support. The multiplicative version of the shared component model for trivariate data in equation 7.6 is

$$\begin{aligned} Z_1 &= Y_{123} Y_{12} Y_{13} \\ Z_2 &= Y_{123} Y_{12} Y_{23} \\ Z_3 &= Y_{123} Y_{13} Y_{23}, \end{aligned} \tag{7.18}$$

which would reduce to a shared frailty model if the pairwise components have zero variance.

Multiplicative models based on gamma components do not have an algebraic expression for the survivor functions: only one of the gamma components can be integrated out via the Laplace transform, leaving some part of the function that must be evaluated numerically (or approximated) for the remaining random variable(s). In the univariate case for instance, with

cumulative hazard function $A(t)$ and gamma-distributed frailty components $x_1$, $x_2$ with shape parameters $\theta_1$, $\theta_2$ and rate parameters $k_1$, $k_2$, the survivor function is derived as

$$\begin{aligned}
S(t) = &\int_0^\infty \int_0^\infty \exp(-A(t)x_1x_2)\frac{\theta_1^{k_1}}{\Gamma(k_1)}x_1^{k_1-1} \\
&\exp(-\theta_1x_1)\frac{\theta_2^{k_2}}{\Gamma(k_2)}x_2^{k_2-1}\exp(-\theta_2x_2)dx_1dx_2 \\
= &\int_0^\infty \frac{\theta_2^{k_2}}{\Gamma(k_2)}x_2^{k_2-1}\exp(-\theta_2x_2)\frac{\theta_1^{k_1}}{(A(t)x_2+\theta_1)^{k_1}} \\
&\int_0^\infty \exp(-x_1(A(t)x_2+\theta_1))\frac{(A(t)x_2+\theta_1)^{k_1}}{\Gamma(k_1)}x_1^{k_1-1}dx_1dx_2 \\
= &\int_0^\infty \frac{\theta_2^{k_2}}{\Gamma(k_2)}x_2^{k_2-1}\exp(-\theta_2x_2)\frac{\theta_1^{k_1}}{(A(t)x_2+\theta_1)^{k_1}}dx_2 \\
= &\frac{\theta_1^{k_1}\theta_2^{k_2}}{\Gamma(k_2)}\int_0^\infty \frac{x_2^{k_2-1}\exp(-\theta_2x_2)}{(A(t)x_2+\theta_1)^{k_1}}dx_2.
\end{aligned}$$

Under the full shared component model, with an overall frailty term and three pairwise components, the trivariate and bivariate functions involve four frailty components. As only one can be integrated out algebraically, the resulting functions require 3-dimensional integration, and therefore substantial computation time. Simpler models are therefore likely to be needed in practice, or an alternative approach to estimation used.

## 7.3 Identifiability and model simplification

Although the models described in sections 7.2.4 and 7.2.5 are technically identifiable, the information available to estimate differences in variances in the "saturated" forms is likely to be very low. In particular, the sparsity of HIV data in the UAM study is likely to result in a lack of identifiability of components shared by all infections and the pairwise components. A small simulation study was conducted under the fixed weight model (equation 7.9) and indicated that even with large amounts of data in all of the trivariate $2 \times 2 \times 2$ table, the log-likelihood surface did not have a well-defined optimum. Briefly, the $k_{123}$ parameter can be higher (lower) and the pairwise

components lower (higher) with negligible difference to the likelihood.

The model also has some issues with convergence. As

$$\begin{aligned}\lambda_1|Z_1 &= \lambda_{0,1}(Y_{123} + Y_{12} + Y_{13})/3\\ \lambda_2|Z_2 &= \lambda_{0,2}(Y_{123} + Y_{12} + Y_{23})/3\\ \lambda_3|Z_3 &= \lambda_{0,3}(Y_{123} + Y_{13} + Y_{23})/3,\end{aligned}$$

if the variance of any component is high, then the mean of the frailty variance will still be 1, but the majority of the distribution has extremely low values. For instance, for a gamma distribution with variance of 100, over 90% of the distribution is less than 0.01. As variance components tend to infinity, the baseline hazard is therefore effectively reduced by about $\frac{1}{3}$, and if two components have variance approaching infinity (or very high) then the reduction is $\frac{2}{3}$. The baseline hazard is thus inflated by a factor of 1.5 or 3 if 1 or 2 components are effectively reduced to zero in this way. Care must therefore be taken, as the model can easily converge to a local optimum at very high estimated variance components, with individual frailty being attributed wholly to $k_{123}$ or to the pairwise components.

The multiplicative form of component model does not have quite the same issue with local optima, but there is likely to still be little information to identify individual components of frailty. Further, as only one component can be integrated out algebraically, the saturated model would require 3-dimensional numerical integration, which is time-consuming to carry out with a sufficient degree of precision. Simpler models would therefore be preferable in both the additive and multiplicative framework.

### 7.3.1 Pairwise-only models

Pairwise-only components may be a pragmatic approach for an investigation of frailties for different infections with no specific hypothesis. The general form of the overall frailty for trivariate data under this model is

$$\begin{aligned}Z_1 &= w_{12}Y_{12} + w_{13}Y_{13}\\ Z_2 &= w_{12}Y_{12} + w_{23}Y_{23}\\ Z_3 &= w_{13}Y_{13} + w_{23}Y_{23}.\end{aligned} \tag{7.19}$$

The survivor functions for the fixed weight model (equation 7.9) reduce to

$$\begin{aligned}
S_{123}(t) &= \left(1+\frac{A_1(t)+A_2(t)}{2k_{12}}\right)^{-k_{12}} \\
&\quad \left(1+\frac{A_1(t)+A_3(t)}{2k_{13}}\right)^{-k_{13}} \left(1+\frac{A_2(t)+A_3(t)}{2k_{23}}\right)^{-k_{23}} \\
S_{12}(t) &= \left(1+\frac{A_1(t)+A_2(t)}{2k_{12}}\right)^{-k_{12}} \left(1+\frac{A_1(t)}{2k_{13}}\right)^{-k_{13}} \left(1+\frac{A_2(t)}{2k_{23}}\right)^{-k_{23}} \\
S_1(t) &= \left(1+\frac{A_1(t)}{2k_{12}}\right)^{-k_{12}} \left(1+\frac{A_1(t)}{2k_{13}}\right)^{-k_{13}},
\end{aligned} \tag{7.20}$$

with the bivariate survivor functions for pairs 1-3 and 2-3 and univariate functions for infections 2 and 3 defined similarly. The survivor functions for the variable weight model in equation 7.14 reduce to

$$\begin{aligned}
S_{123}(t) &= \left(1+\frac{A_1(t)}{k_{12}+k_{13}}+\frac{A_2(t)}{k_{12}+k_{23}}\right)^{-k_{12}} \\
&\quad \left(1+\frac{A_1(t)}{k_{12}+k_{13}}+\frac{A_3(t)}{k_{13}+k_{23}}\right)^{-k_{13}} \left(1+\frac{A_2(t)}{k_{12}+k_{23}}+\frac{A_3(t)}{k_{13}+k_{23}}\right)^{-k_{23}} \\
S_{12}(t) &= \left(1+\frac{A_1(t)}{k_{12}+k_{13}}+\frac{A_2(t)}{k_{12}+k_{23}}\right)^{-k_{12}} \\
&\quad \left(1+\frac{A_1(t)}{k_{12}+k_{13}}\right)^{-k_{13}} \left(1+\frac{A_2(t)}{k_{12}+k_{23}}\right)^{-k_{23}} \\
S_1(t) &= \left(1+\frac{A_1(t)}{k_{12}+k_{13}}\right)^{-(k_{12}+k_{13})},
\end{aligned} \tag{7.21}$$

with the bivariate survivor functions for pairs 1-3 and 2-3 and univariate functions for infections 2 and 3 defined similarly.

The multiplicative model is also simplified: two-dimensional integration is still required, which is still slow, but feasible for relatively small datasets (in terms of the number of binomial data points).

Pairwise models are much less likely to suffer from convergence issues, as each component $k_{ij}$ is informed exclusively by the pairwise correlation

between infection $i$ and $j$, rather than differences between correlations that inform $k_{123}$ and $k_{ij}$ in saturated models. However, this also severely restricts the range of variance estimates that can be obtained from the model, as if pairs 1-2 and 1-3 are positively correlated, 2-3 must be also. As with many of these models, the pattern of parameters and hypothesis tests against simpler models may be more informative than attempts at direct interpretation of the estimates. However, this is something of a problem with the pairwise-only model as it does not include the shared frailty model as a special case, either in the additive or multiplicative form.

### 7.3.2 Asymmetric models

An alternative approach to simplifying the saturated shared component model is to add just one (or two) of the three possible pairwise terms to the shared frailty model. Although there may still be difficulties in distinguishing the 3-way and pairwise component(s), this simplified version is less likely to be problematic than the full flexibility of the saturated form, and can be used to test specific hypotheses about certain infection pairs.

The model is asymmetric in the sense that different infections may have different numbers of frailty components. For an additive component model, this requires that a constant is added to the overall frailty for one or more infections. This is the same issue as the limited support of additive component frailty distributions described in section 7.2.4, as the variances of some of the pairwise components are effectively set to zero. For instance, the single pair asymmetric model with the additional component for pair 1-2 may be specified as

$$\begin{aligned} Z_1 &= (Y_{123} + Y_{12})/2 \\ Z_2 &= (Y_{123} + Y_{12})/2 \\ Z_3 &= (Y_{123} + 1)/2, \end{aligned} \tag{7.22}$$

where $Y_{123}$ and $Y_{12}$ are gamma distributions with unit mean. For a two-pair asymmetric model that also includes a component for pair 1-3 the model

may be specified as

$$\begin{aligned} Z_1 &= (Y_{123} + Y_{12} + Y_{13})/3 \\ Z_2 &= (Y_{123} + Y_{12} + 1)/3 \\ Z_3 &= (Y_{123} + Y_{13} + 1)/3. \end{aligned} \tag{7.23}$$

Again, the multiplicative model more naturally handles the asymmetric specification as pairwise components with mean 1 are simply multiplied to produce the overall frailty density, which also has mean 1. The multiplicative version of equation 7.23 for instance is

$$\begin{aligned} Z_1 &= Y_{123}Y_{12}Y_{13} \\ Z_2 &= Y_{123}Y_{12} \\ Z_3 &= Y_{123}Y_{13}, \end{aligned} \tag{7.24}$$

and the model may be simplified to a single component by omitting $Y_{12}$ or $Y_{13}$, or the shared frailty model if both are omitted. The shared frailty model is therefore nested in multiplicative asymmetric models and hypothesis tests can be conducted to determine whether the additional complexity is warranted.

## 7.4 Epidemiological considerations

Two approaches for additive gamma components have been proposed with slightly different properties: the fixed weight model and the varying weight model. The varying weight model scales the weights of each component according to their variances, while the fixed weight model assumes that the weights are fixed, regardless of the variance of each component. The latter results in a potentially limited support, which may or may not be appropriate for the data. However, the variable weight model has restrictions on the range of possible correlations between the overall frailties for each infection. The alternative is to specify frailty components that combine multiplicatively, although these models are more challenging to work with, requiring approximation of the necessary integrals.

It is worth considering the potential scenarios in which the shared frailty

model requires extension, i.e., where individual susceptibility to infection is not determined by some factor that has an identical effect on the risk of all infections. This may occur where there is more than one route of infection, and the relative importance of these routes differs between infections. In the context of blood-borne viruses in PWID, HCV, HBV and HIV may all be transmitted via sharing of injecting paraphenalia, but there is also a risk of sexual aquisition which is comparatively low for HCV (Balogun et al., 2003), while HIV is transmitted sexually more easily than via blood-to-blood contact and therefore would have a smaller component of injecting risk.

In the case of two infection routes $A$ and $B$, each of which have an associated frailty distribution $Y_A$ and $Y_B$ with unit means, the overall hazard at time $t$ may be written as

$$\lambda(t|Y_A, Y_B) = \lambda_A(t)Y_A + \lambda_B(t)Y_B, \tag{7.25}$$

where $\lambda_A(t)$ and $\lambda_B(t)$ are the route-specific baseline forces of infection. Writing the relative contribution of infection $A$ as $w_A(t) = \frac{\lambda_A(t)}{\lambda_A(t)+\lambda_B(t)}$ and $w_B(t) = 1 - w_A(t)$, the frailties combine additively in a similar way to that of the general additive components model in equation 7.6, acting on the overall hazard. However, there would likely be insufficient information to estimate both the weight term $w_A(t)$ and the variances. Some progress might be made if the FOIs could be considered proportional, although strong assumptions are required in order to consider this model as directly representing route-specific frailty.

Considering further the route-specific risks of infection through sexual and injecting risk, there is in fact some information to investigate their relative contributions in the UAM data. Assuming an average age of sexual maturity, the data consist of time at sexual risk only (pre-injecting) and time at risk of both injecting and sexual risk, plus infection status.

Denoting $A_i(t)$ the cumulative hazard due to injecting, equal to zero for $t < t_i$, the age at start of injecting, and $A_s(t)$ the cumulative hazard due to sexual transmission, equal to zero for $t < t_s$, the age at sexual maturity, the survivor function is defined as

$$S(t) = \exp\left(-A_i(t) - A_s(t)\right). \tag{7.26}$$

This model may therefore estimate age-specific variation in the hazard for sexual transmission within the age range of age at first use available in the data (before the two hazards become confounded) and injecting-duration specific risk, *provided* the risk of sexual infection is not modified by the onset of injecting. The latter might be a strong assumption, as both overall risk and heterogeneity in risk may be strongly influenced by injecting drug use. This model may therefore be of limited usefulness in determining the injecting and sexual components of risk in PWID, although could provide a basis for exploring alternative weight terms for components of injecting and sexual risk in the model described in section 7.2.4.

The multiplicative frailty model is a less natural representation for route-specific frailties, which would not generally have a multiplicative effect: in the UAM example, a doubling of risk in sexual and injecting risks would not result in a 4-fold increase in overall risk. However, this may approximate the route-specific components themselves being positively correlated (without this correlation being explicitly modelled).

One scenario in which the multiplicative model does directly represent component frailties is when overall frailty is the product of behavioural heterogeneity and biological susceptibility, in which case the two components naturally have a multiplicative effect on the overall hazard. Some biological knowledge may be required to specify a model representing these effects, as again the frailty components must be specified in terms of components shared by all three infections or pairwise components. Again, the model might be considered more generally as representing components of risk that act on different infection pairs and overall, rather than an attempt to directly quantify variability in a particular aspect of the risk of infection.

## 7.5 Fitting trivariate frailty models to the UAM data

In this section trivariate frailty models are fitted to the UAM data on people who inject drugs (PWID). The fixed weight additive components model (with equal weights), variable weight model and multiplicative components models are fitted to the data, assuming pairwise-only shared components as

described in section 7.3.1. These models are used to examine general patterns in pairwise correlations and compared with the shared frailty model, which assumes a single common frailty component that acts equally on all infections. Model fit statistics may then be used to determine if the additional complexity of the component models is necessary, although the shared frailty is not a nested model for either the additive or multiplicative models. Asymmetric models are then considered, primarily to test the hypothesis that additional variability in risk for HIV and HBV may arise through sexual transmission. These models are fitted in their multiplicative form such that the nested structure can be explicitly tested, and likelihood ratio statistics used to compare against the shared frailty model.

The dataset considered is that used in section 6.2, which includes data from 2000 onward with complete data for HCV, HBV and HIV tests. Although the likelihood can incorporate partial data where one or more tests are missing, the number of incomplete observations is small and these are discarded. Those injecting for greater than 35 years are also excluded, leaving 24977 observations for analysis. In order to simplify modelling, injecting duration-specific baseline hazards are estimated but no other covariates, including time, are included in the model. The data therefore take the form of the standard age-specific seroprevalence data used in many epidemiological studies (see for example Farrington et al. (2001)). Imperfect sensitivity and specifity are accounted for in the observed data as described previously (section 3.1.2). The trivariate data are summarised in Table 7.1.

Piecewise constant hazards are fitted with injecting duration bands of $\leq 1$, 1-3, 3-5, 5-7, 7-10, 10-15, 15-20, 20-25 and $> 25$ years. Models were fitted in R using bespoke code (see appendix section 9.2.4), and numerical integration for the multiplicative models performed using `integrate` and `adaptIntegrate` from the cubature package for univariate and bivariate integrations respectively as required. Deviances were calculated based on observed and predicted numbers in each trivariate cell, grouped by injecting durations of $\leq 1$, 2, 3-4, 5-7, 8-10, 11-14, 15-19, 20-24, 25-29 and 30-35 years in order to reduce zero cell counts.

Table 7.1: Trivariate infection status data by injecting duration ($t$) from the UAM survey, 2000-2014. Cell counts are denoted $n_{ijk}$ with 0=uninfected and 1=infected for HCV ($i$), HBV ($j$) and HIV ($k$).

| $t$ | $n_{000}$ | $n_{100}$ | $n_{010}$ | $n_{110}$ | $n_{001}$ | $n_{101}$ | $n_{011}$ | $n_{111}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1946 | 395 | 65 | 64 | 7 | 5 | 1 | 3 |
| 2 | 1222 | 287 | 55 | 55 | 7 | 3 | 0 | 3 |
| 3 | 978 | 266 | 57 | 56 | 4 | 3 | 2 | 3 |
| 4 | 870 | 338 | 67 | 63 | 3 | 1 | 0 | 0 |
| 5 | 864 | 360 | 69 | 68 | 5 | 1 | 0 | 0 |
| 6 | 821 | 382 | 68 | 67 | 4 | 3 | 0 | 2 |
| 7 | 685 | 371 | 47 | 77 | 3 | 5 | 0 | 2 |
| 8 | 649 | 371 | 47 | 87 | 5 | 2 | 0 | 5 |
| 9 | 554 | 325 | 50 | 97 | 0 | 2 | 0 | 2 |
| 10 | 595 | 441 | 76 | 109 | 1 | 4 | 1 | 3 |
| 11 | 483 | 345 | 58 | 111 | 3 | 1 | 0 | 5 |
| 12 | 451 | 322 | 36 | 125 | 5 | 3 | 0 | 4 |
| 13 | 414 | 381 | 40 | 118 | 1 | 8 | 1 | 6 |
| 14 | 364 | 314 | 44 | 109 | 5 | 6 | 0 | 5 |
| 15 | 298 | 335 | 36 | 141 | 5 | 5 | 0 | 4 |
| 16 | 274 | 285 | 37 | 137 | 2 | 3 | 1 | 3 |
| 17 | 247 | 277 | 38 | 130 | 3 | 8 | 2 | 4 |
| 18 | 199 | 267 | 41 | 114 | 4 | 2 | 0 | 2 |
| 19 | 196 | 234 | 28 | 108 | 4 | 4 | 2 | 6 |
| 20 | 152 | 245 | 34 | 145 | 3 | 6 | 4 | 2 |
| 21 | 107 | 184 | 26 | 114 | 1 | 5 | 2 | 1 |
| 22 | 97 | 166 | 25 | 100 | 0 | 6 | 0 | 3 |
| 23 | 97 | 147 | 16 | 85 | 1 | 5 | 0 | 6 |
| 24 | 69 | 138 | 29 | 74 | 2 | 1 | 0 | 3 |
| 25 | 55 | 111 | 18 | 70 | 0 | 1 | 0 | 8 |
| 26 | 41 | 103 | 5 | 65 | 0 | 3 | 0 | 2 |
| 27 | 39 | 54 | 10 | 66 | 0 | 3 | 1 | 1 |
| 28 | 28 | 63 | 6 | 58 | 0 | 0 | 1 | 0 |
| 29 | 18 | 54 | 16 | 54 | 0 | 1 | 0 | 1 |
| 30 | 23 | 49 | 8 | 48 | 1 | 1 | 1 | 1 |
| 31 | 6 | 38 | 3 | 35 | 0 | 2 | 0 | 0 |
| 32 | 13 | 28 | 10 | 49 | 0 | 1 | 1 | 0 |
| 33 | 15 | 26 | 4 | 46 | 0 | 0 | 0 | 1 |
| 34 | 11 | 20 | 2 | 32 | 0 | 0 | 0 | 4 |
| 35 | 4 | 15 | 3 | 26 | 0 | 0 | 0 | 0 |

### 7.5.1 Pairwise component model results

Estimated baseline hazards from the four models are shown in Table 7.2. The estimates show a similar pattern to those in chapters 4 and 6, with a far higher risk in the first year of injecting and constant, or no particular pattern thereafter; although the estimated hazards increase again somewhat at longer injecting durations. This is partly due to sparsity of data, but also likely due to not adjusting for variation in calendar time, with those with longer injecting durations having been at risk during the peak of the epidemic in the 1980s. Nevertheless, both of the additive component models produced implausible estimates for the hazard for HCV infection in the 20-25 and 25+ groups, and for the HBV hazard in the 25+ group.

Table 7.3 shows the estimated frailty parameters from the pairwise component models. The components themselves have different interpretations for each model: the fixed weight additive components are scaled by a factor of $\frac{1}{2}$ whereas the variances of the multiplicative components produce a higher overall frailty variance. However, the patterns of these two models are comparable and suggest that the HCV-HBV and HBV-HIV pairs have a stronger component (correlation) than that of the HCV-HIV pairs. This is reflected in the overall frailty variance, which is highest for HBV, as the overall frailty does not include the weaker HCV-HIV component shared by the other two infections. This is not the case for the variable weight additive model however, which has the smallest overall variance for HBV, due to overall frailties being based on inverses of components (equation 7.15). The HCV-HIV component is the largest, which is somewhat implausible as HCV and HIV are expected to have the weakest correlation from an epidemiological perspective and would therefore be expected to have the smallest variance component.

Table 7.2: Estimates of injecting duration-specific baseline hazards for HCV, HBV and HIV infection from trivariate frailty models: shared frailty model, pairwise-only additive models (fixed weight, FW; variable weight, VW) and pairwise-only multiplicative component model.

| Injecting | Shared | | | Additive (FW) | | | Additive (VW) | | | Multiplicative | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| duration | HCV | HBV | HIV | HCV | HBV | HIV | HCV | HBV | HIV | HCV | HBV | HIV |
| $\leq 1$ | 0.233 | 0.075 | 0.0048 | 0.317 | 0.113 | 0.0027 | 0.287 | 0.074 | 0.0061 | 0.265 | 0.081 | 0.0044 |
| 1-3 | 0.048 | 0.024 | 0.0008 | 0.063 | 0.000 | 0.0014 | 0.077 | 0.024 | 0.0000 | 0.049 | 0.022 | 0.0008 |
| 3-5 | 0.088 | 0.008 | 0.0001 | 0.230 | 0.041 | 0.0009 | 0.152 | 0.017 | 0.0001 | 0.104 | 0.030 | 0.0004 |
| 5-7 | 0.034 | 0.006 | 0.0002 | 0.000 | 0.012 | 0.0007 | 0.147 | 0.004 | 0.0010 | 0.084 | 0.014 | 0.0011 |
| 7-10 | 0.088 | 0.031 | 0.0006 | 0.357 | 0.067 | 0.0014 | 0.206 | 0.040 | 0.0017 | 0.141 | 0.049 | 0.0010 |
| 10-15 | 0.144 | 0.030 | 0.0033 | 0.501 | 0.089 | 0.0025 | 0.472 | 0.032 | 0.0000 | 0.203 | 0.055 | 0.0020 |
| 15-20 | 0.136 | 0.036 | 0.0004 | 0.875 | 0.137 | 0.0009 | 0.910 | 0.063 | 0.0028 | 0.247 | 0.089 | 0.0015 |
| 20-25 | 0.118 | 0.027 | 0.0010 | 2.513 | 0.457 | 0.0017 | 1.492 | 0.044 | 0.0000 | 0.292 | 0.120 | 0.0016 |
| 25+ | 0.227 | 0.103 | 0.0004 | 1.644 | 2.141 | 0.0011 | 3.982 | 0.202 | 0.0003 | 0.579 | 0.469 | 0.0016 |

Interestingly however, the correlations between overall frailties, $\rho_{ij}$, are very similar under the three models, and support the hypothesis that HCV and HBV are likely to have a strong correlation due to the importance of the shared route of injecting transmission, and HBV and HIV will have a strong correlation (compared to HCV-HIV) due to the shared sexual component.

The shared frailty model estimated the overall frailty variance to be 0.90, less than any of the infection pairs under any of the component frailty models. This might be due in part to the way that the relative frailty variance evolves under the different models: the component models result in declining relative frailty variance, while it is constant for the shared frailty model. A similar pattern was observed under an inverse Gaussian model compared to gamma frailties in chapter 6.

Table 7.3: Estimates of frailty parameters from trivariate frailty models: pairwise-only additive models (fixed weight, FW; variable weight, VW) and pairwise-only multiplicative component model. $\delta$ terms are component variances, var($Z$) the overall frailty variances and $\rho$ the correlations for infection 1, 2 and 3 (HCV, HBV and HIV).

| Parameter | Infection(s) | Additive (FW) | Additive (VW) | Multiplicative |
|---|---|---|---|---|
| $\delta_{12}$ | HCV-HBV | 10.75 | 3.70 | 1.22 |
| $\delta_{13}$ | HCV-HIV | 3.70 | 12.50 | 0.36 |
| $\delta_{23}$ | HBV-HIV | 11.36 | 4.55 | 1.79 |
| var($Z_1$) | HCV | 3.63 | 2.82 | 2.01 |
| var($Z_2$) | HBV | 5.60 | 2.02 | 5.16 |
| var($Z_3$) | HIV | 3.82 | 3.29 | 2.79 |
| $\rho_{12}$ | HCV-HBV | 0.60 | 0.65 | 0.60 |
| $\rho_{13}$ | HCV-HIV | 0.25 | 0.24 | 0.25 |
| $\rho_{23}$ | HBV-HIV | 0.62 | 0.58 | 0.63 |

Predicted trivariate infection status under the fitted models are shown in Figure 7.1. In general, the model fit appears similar, and reasonably good, for infection combinations for which there are sufficient data (uninfected, HCV only, HCV-HBV coinfection) but with greater differences between models where data are sparse (HBV without HCV, any combination with HIV). In terms of formal model comparison, the multiplicative model provides an improvement over the shared frailty model, with a difference in AIC scores of 10.6. However, the fixed weight additive model provided a substantially worse fit, with an *increase* in AIC of 49.0, which was particularly poor for

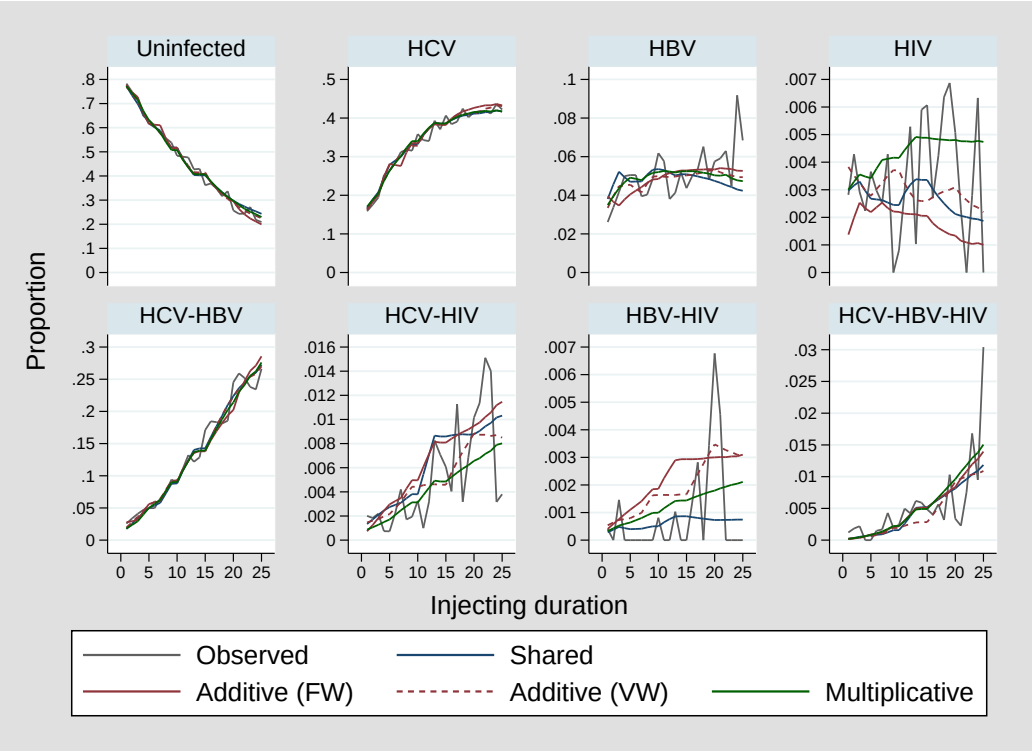


Figure 7.1: Observed and predicted trivariate infection status by injecting duration under a shared frailty model and pairwise-only additive (fixed weight, FW; variable weight, VW) and multiplicative component models.

the HIV data. Despite its questionable interpretation, the variable weight additive model provided the best fit to the data, with an improvement in AIC score of 16.5 compared to the multiplicative model.

These results indicate that there is additional complexity in the data that warrant a different model to the shared frailty model, but the additive pairwise component model with fixed weights is not the most appropriate choice. The overall model fits were passable but not ideal, with a deviance of 366.7 for the shared frailty model on 35 octonomial observations (multinomial data points with 8 possible categories) with 28 parameters. For the pairwise models (with 30 parameters) deviances were 410.9 for the fixed weight additive model and 355.4 for the multiplicative model. Although the model fit was best under the variable weight additive model, the multiplicative model has a more natural interpretation and is likely to be a more suitable choice.

### 7.5.2 Asymmetric model results

Multiplicative asymmetric models were fitted to the UAM data, primarily with the purpose of testing for additional variability for HBV-HIV due to sexual transmission, but also to examine any pairwise differences for the infection pairs HCV-HBV and HCV-HIV beyond the shared frailty model. Even these simple models showed some issues with identifiabilty, and frequently converged to a lower log-likelihood than the shared frailty model. Starting values were therefore based on estimates from the shared frailty model, with a low initial variance assigned to pairwise components. Resulting estimates of frailty variances and AIC scores are shown in Table 7.4.

The HBV-HIV model provided a significant improvement in model fit, with a reduction in AIC score of 9.6 compared to the shared frailty model, and estimated an additional variance of 0.373 for HBV and HIV. The estimate of the shared frailty component was very similar to the model including a shared frailty only, indicating a similar variability for HCV but a marked increase for HBV and HIV.

The HCV-HBV model provided a greater improvement in model fit, with a reduction in AIC score of 23.0 compared to the shared frailty model. Interestingly, the variance of the shared frailty component was estimated to be lower, but with a substantial HCV-HBV component. This results in the overall frailty variance for HCV and HBV being comparable to the shared frailty model, but a lower variance for HIV, in contrast to the HBV-HIV component model. The HCV-HIV model provided a negligible improvement in model fit, with a reduction in AIC score of 1.8, and the variance of the HCV-HIV component estimated to be practically zero.

The results of the single pairwise component models indicated that a model with both HCV-HBV and HBV-HIV components warranted investigation. This model was fitted to the data using model estimates as initial values from (1) the HCV-HBV model, and (2) the HBV-HIV model, with low initial value for the added components. Different estimates were obtained from the two starting points, with the estimated variance of the added component remaining low. In both cases there were only modest improvements in model fit, with a change in AIC score of 2.2 and 3.2 from the first and second sets of initial values.

Table 7.4: Estimates of frailty parameters from shared frailty and asymmetric trivariate frailty models, with additional pairwise components for HCV-HBV (var($Y_{CB}$)), HBV-HIV (var($Y_{BI}$)) and HCV-HIV (var($Y_{CB}$)). The model with both HCV-HBV and HBV-HIV components (C-B B-I) is started from two sets of initial values: (1) estimates from the HCV-HBV model, and (2) from the HBV-HIV model.

| Model | var($Y_{CBI}$) | var($Y_{CB}$) | var($Y_{BI}$) | var($Y_{CI}$) | AIC |
|---|---|---|---|---|---|
| Shared frailty | 0.896 | | | | 54675.2 |
| HBV-HIV | 0.911 | | 0.373 | | 54665.6 |
| HCV-HBV | 0.638 | 0.318 | | | 54652.2 |
| HCV-HIV | 0.898 | | | 0.050 | 54673.4 |
| C-B B-I (1) | 0.642 | 0.319 | 0.016 | | 54650.0 |
| C-B B-I (2) | 0.893 | 0.037 | 0.352 | | 54662.4 |

Deviances for the HBV-HIV and HCV-HBV models were 357.0 and 337.4 for 35 octonomial observations and 29 parameters; and for the best-fitting HCV-HBV and HBV-HIV model 333.5 with 30 parameters. These results indicate that there is a substantial improvement in model fit with the addition of HCV-HBV or HBV-HIV parameters, but the model fit is still not ideal.

## 7.6 Conclusions

The models proposed in this chapter extend the available models for current status data. Nevertheless, even with trivariate data there are limits to what can be understood about aspects of frailty that are not shared by all infections, due to the restriction that any component of frailty must be common to at least two infections to be identifiable. Key areas of interest in this topic, namely the contributions of different routes of infection and behavioural heterogeneity vs. biological susceptibility, therefore have limited scope for investigation.

The results obtained in section 7.5.1 are suggestive of a plausible epidemiological explanation, in that both HCV and HBV are commonly transmitted in PWID through the sharing of needles and syringes, whereas HIV transmission is relatively rare due to its inherently lower infectiousness through blood-borne routes. However, HIV is readily transmitted sexually, and this is the component that is shared by HBV, while sexual transmission of HCV is

relatively rare. This may therefore have produced the observed pattern here, with HBV having the highest frailty variance and (equivalently) a weaker correlation between HCV and HIV.

This is a plausible explanation, but is only a tentative hypothesis. One argument against it is that for HBV to have a markedly higher frailty variance than HCV, the component of sexual transmission (or $w_i$ parameter, to use the form of equation 7.6) would need to be of comparable magnitude to that of injecting risk in order to have any substantial contribution to overall heterogeneity. Although it is difficult to unpick the relative contributions of injecting and sexual risk in this population, unless sexual risk is markedly higher than that of the general population, injecting transmission risk would far outweigh that of sexual risk.

The asymmetric models fitted in section 7.5.2 support the sexual risk hypothesis for HBV-HIV, but the HCV-HBV model provided a better fit still. The two models provided different estimates of frailty variance for HIV; it is possible that both models have converged to local optima in the likelihood and did not find the overall optimum. Some care is required in the interpretation of these results, as the HIV data are very sparse and likely outweighed by any potential improvements in model fit to the other data. Differences in model fits, although substantial in a statistical sense, do not markedly improve the absolute fit of the models, which remains adequate rather than good.

# Chapter 8

# Conclusions

## 8.1 Summary of findings

In this thesis models for current status data on people who inject drugs (PWID) have been studied, using data from the Unlinked Anonymous Monitoring (UAM) survey of PWID. Key themes have been the investigation of risk factors and patterns of unmeasured heterogeneity in different groups, which are examined in the framework of force of infection (FOI) models. This thesis has explored technical developments of the FOI model, such as the inclusion of covariate information and extension to trivariate frailty models; and addressed practical questions on patterns of risk that can be used to develop public health policy.

FOI models provide a natural framework for investigating age-specific current status data, which in this thesis are the injecting duration-specific prevalence of HCV, HBV and HIV. The relationship between time at risk and infection status can be economically modelled by the FOI. If it is constant, a single parameter can describe how prevalence increases with time at risk, and the rate of infection is also the most meaningful parameter when considering public health policy. A key result of this thesis is that risk of infection with HCV (and other viruses) is very high in the first year, but largely constant thereafter; thus the risks of infection in PWID can be summarised by the two parameters for the 1st year and subsequent FOI.

Fitting FOI models requires some technical expertise, but under the assumption of additive risk this can be done using generalised linear model

routines with standard software. This approach can therefore be used to quickly investigate different models, in particular when considering a number of covariates and potential interactions. Additive covariate models were fitted to the UAM data and revealed striking regional patterns in risk during the first year of injecting, and a more homogenous risk across regions for the remainder of injecting career. Differences in temporal trends were also identified, although model comparison scores indicated that a three-way interaction between region, injecting duration and time was not needed, leading to a more parsimonious model than fitting models to each region separately.

Covariate effects can be fixed, having the same effect on the FOI at all injecting durations, or varying according to injecting duration, calendar time, or both. Time-varying covariates can also be included, where effects change according to exposure time, as in survival analysis. This approach was used to determine the risk of infection before and after starting to use needle exchange services, based on the reported age at first use of such services. Results were somewhat difficult to interpret, indicating a decrease in the risk of infection after starting to use needle exchange services, but only in the post-2000 period (a time-varying covariate with changing effect over calendar time). Nevertheless, this analysis shows that information on the timing of a certain risk factor can be incorporated in force of infection models, and could be considered when designing cross-sectional surveys.

Multiplicative risks are more complicated to handle, as the link between the covariates and the outcome variable(s) is non-linear. In this thesis bespoke R code was developed to fit such models. Covariates are still somewhat problematic, as the model code requires adapting for different covariate structures. A "multirow" formulation of the data was developed, in which the injecting duration and calendar time bands are split within each individual, allowing an arbitrary matrix of covariate information to be included easily. This has the downside of increasing the size of the dataset and number of computations required, and therefore the time and computing resources needed to fit models. It is likely that the routine could be made more efficient, or somehow "compiled" (as in computer programming) to minimise the number of calculations performed, although this was not pursued.

Shared frailty models identify the degree of unobserved heterogeneity via

the correlation between infections in bivariate data. Patterns of bivariate association were therefore examined. These revealed that the association between infections declined with injecting duration for all infection pairs, although data on HIV were sparse. For HCV-HBV, there was evidence of a potential change in correlation patterns according to survey year, with a broadly constant association in more recent data, but a greater decline with increasing injecting duration in earlier years. Some potential differences according to age and reported risk behaviour were also noted, but of particular interest are the patterns by gender, with a higher level of variability in males compared to females. Females have a higher average risk, leading to the conclusion that females are generally at high risk of infection, whereas males are a mixture of lower and higher risks.

The declining association between infections with time at risk indicates that the gamma frailty model, which has constant relative frailty variance (RFV), will not capture the patterns of correlation in the data. The inverse Gaussian distribution induces a selection effect that results in a declining RFV, and indeed this model provided a better fit to the data. A better fit still was obtained under a time-varying frailty (TVF) model, which explicitly models a decline in frailty variance. This may point to a homogenisation in risk behaviour throughout injecting career, which appears plausible as PWID may commence injecting under a variety of circumstances, but become more homogeneous in their risk patterns as time passes (as well as being at generally lower risk). However, selection effects cannot be distinguished from genuine reductions in underlying variability, and the TVF model may simply fit better due to the additional model parameter. In particular, the TVF model indicates a faster decline in RFV following initiation of injecting than the inverse Gaussian model, although other frailty distributions that better reflect variability in risk are of course possible, and need not necessarily be time-varying to capture the observed correlation pattern.

Covariates were then added to the different bivariate frailty models, focussing on HCV-HBV, which have the most data. Interestingly, inclusion of covariates resulted in substantially improved model fit, but only modest reductions in estimates of frailty variances, which may be interpreted as residual heterogeneity. Overall estimates of frailty variances were 1.94 in the unadjusted and 1.44 for the fully adjusted inverse Gaussion model. To

interpret these results the ratio between the 75th and 25th percentiles of the frailty distribution can be considered: this is 4.5 for the unadjusted model, and 3.9 for the adjusted model, indicating substantial differences in risk even after accounting for a number of risk factors.

Also of note is that the estimated excess risk in first year injectors persisted in frailty models. A well-known effect of frailty is that high-risk individuals tend to experience the event of interest earlier on, which may have partly explained the observed high prevalence in 1st year injectors. Unless the underlying frailty distribution has a very different form to the standard distributions considered here, this appears not to be the case.

The final chapter investigated trivariate frailty models. Trivariate (2x2x2) data have seven degrees of freedom at each time point; after estimating the three baseline hazards, this leaves four degrees of freedom to estimate the correlation structure between infections, and hence components of frailty. However, the same limitation exists as for bivariate data: information on individual variability is only available at a single time point via the correlation between infections, and frailty components relating to a single infection cannot be identified. The "saturated" model therefore consists of a component relating to all three infections, and three pairwise components. There is some choice in the manner in which additive components can be combined to form a single frailty that has unit mean, which broadly fall into *fixed weight* and *variable weight* models; the latter has a similar form to the correlated frailty model. Alternatively, the components may combine multiplicatively. Multiplicative models have attractive properties, but require numerical integration and are more difficult to work with.

The information available to estimate *differences* in variances is very low, and in practice simpler models had to be fitted to the data. This allowed for the investigation of a specific hypothesis for blood-borne infections in PWID, that HBV and HIV would have an additional frailty component pertaining broadly to sexual risk, which is generally low for HCV. While this was found to provide a better fit compared to a model with a single overall frailty parameter, a better fit still was obtained when including an HCV-HBV component rather than the HBV-HIV component. Including more than one pairwise component resulted in model instability, with the model estimating that the pairwise variability (in addition to the overall

variability) was wholly attributable to HCV-HBV or HBV-HIV, depending on the initial values used when starting the model. Ultimately, the HCV-HBV and HBV-HIV correlations are the strongest, whereas the correlation between HCV and HIV is weakest. This may point to sexual risk being a greater component for HIV, and this risk is shared for HBV. However, the majority of risk for HCV and HBV is through injecting, and hence the correlation between these infections is strong.

## 8.2 Public health implications

The analyses presented here confirm the high excess risk in recent initiates to injecting drug use, which persisted after accounting for individual variability and indicated that the FOI for HCV in 1st year injectors has increased in recent years (2010-2015). The implication is that harm reduction efforts need to focus on recent initiates (in addition to discouraging individuals to begin injecting in the first place). High rates of infection in 1st year injectors were observed in Yorkshire and the Humber, London, and in particular the North West. The latter two are well-known areas of high prevalence, although these analyses suggest that minimising risk in new initiates is of utmost importance, with generally comparable levels of risk across regions following the first year of injecting.

Females appear to be particularly vulnerable to infection from HCV, with a higher risk and less variability than men, some of whom appear to be at relatively low risk. There is less of a gender difference for HBV and HIV. It is not clear under what circumstances females tend to start injecting, but efforts clearly need to be made for vulnerable risk groups. Those that initiate injecting at an older age (25 or more) appear to be at particularly high risk of HCV and HBV, although for HIV those initiating injecting at younger ages ($< 18$) appeared to be at higher risk. Ever being in prison is a major risk factor for HCV and HBV, and the majority of PWID in the survey (69%) have been imprisoned at least once. Potential explanations are the lack of sterile injecting equipment when in prison, the disruption of networks and routines, and post-imprisonment binge use when released.

It is interesting to consider the hypothesis that risk of infection is generally low in those with established injecting patterns and access to needle

exchanges, with infections occurring during periods of disruption. The high initial risk might (in part) be due to all PWID being in a "disrupted" state at initiation, rather than something intrinsic about initiation risks. If such a hypothesis holds, this would indicate that harm reductions should minimise the impact of disruptive states, such as homelessness and imprisonment.

This thesis has focussed on HCV, which has the highest prevalence in PWID and comparatively high risks of developing severe disease. Results for HBV are somewhat more difficult to interpret, due to a proportion of individuals being vaccinated. However, HBV prevalence has fallen consistently throughout the survey. HIV infection has remained low, but shows no sign of decreasing. It is quite possible that sexual transmission is a major route of infection in this population, rather than injecting. Patterns of correlation indicate that this may be the case, in addition to the markedly higher HIV prevalence in men who have sex with men, who are a key risk group for sexually acquired HIV.

Of particular interest at the current time is the potential impact that new treatments for HCV will have on the prevalence and resulting transmission of infection in PWID. Incident infections are difficult to observe or estimate directly, requiring large cohort studies or relying on elusively small numbers of individuals with markers of recent infection. FOI models can provide estimates of infection rates through the differences in injecting duration-specific prevalence from one year to the next, and have the potential to detect changes in incidence that might be associated with the impact of treatment.

The FOI model is also superior to examining trends in observed prevalence across survey years, which does not account for potential changes in factors such as injecting duration, either through sampling variability or a systematic shift. This has become increasingly important as the average age, and injecting duration, of PWID sampled in the UAM has increased over time.

## 8.3 Further work

The methodology in this thesis has focussed on piecewise constant FOI models and standard frailty distributions, fitted within a classical framework.

Parametric models may lead to more efficient estimates, which could allow for more complexity in other areas of the model, such as local-level estimates. Flexible, non-parametric models such as splines and their multidimensional extensions may also be used to investigate the relationship between calendar time, injecting duration and risk of infection. However, these analyses appeared to indicate that the only important differences in risk occur in the first year of injecting, although there was evidence of temporal changes.

Current modelling of HCV in England aims to generate prevalence estimates for the 22 geographic areas of Operational Delivery Networks (ODN) for HCV treatment, using a FOI model to estimate HCV incidence and prevalence in PWID (Public Health England, 2018*b*). Due to the sparse data at this level a relatively simple model is required, such as a constant FOI with excess risk in the first year of injecting, and broad, piecewise constant time intervals. The modelling conducted as part of this thesis is therefore a valuable input for this work, guiding the formulation of ODN-level modelling.

Further investigation of alternative frailty distributions could be undertaken, in particular the positive stable distribution. Of interest would be to use information on contact patterns, or numbers of equipment sharing partners, in PWID to determine an appropriate distribution. However, such information is likely to be very difficult to obtain.

Fitting models within a Bayesian framework would allow a number of possible extensions. Variance components may be assigned semi-informative priors, which may improve identifiability of parameters, in particular frailty components. However, if there is insufficient information to reliably estimate a parameter in the classical framework, estimates will likely be highly sensitive to prior assumptions. More usefully, the marginal distribution of the survivor function need not be explicitly defined: the frailty distribution may be assigned a prior distribution with hyper-parameters, and frailties for each individual simulated with a fully Bayesian Markov Chain Monte Carlo framework. This would allow the use of frailty distributions that are not analytically tractable, and obviate the need for numerical integration of multivariate functions such as that of the multiplicative component model.

Other possibilities in a Bayesian framework are to incorporate external information, such as data on mortality rates (which could incorporate uncertainty, rather than being fixed) in order to account for differential mor-

tality in infected individuals. Another question around the UAM data is whether self-reported injecting duration is accurate. The misreporting process could be modelled within a Bayesian framework, although this would either require some assigned distribution, parametric model assumptions for the injecting-duration specific FOI, or external information on this process to be identifiable. Finally, Bayesian models are a natural environment for non-parametric functions, which may be used for the baseline injecting duration- and time-specific FOI as described above.

Markers of recent infection have been tested for in the UAM data since 2011, including those RNA positive and antibody negative, and avidity levels, although the latter has been discontinued. In principle, such data could be incorporated as part of a multi-state FOI model, in which susceptible individuals move to a recently-infected state initially, then to established infection. This would provide an assessment of the consistency of estimates generated under a standard FOI model and the data on recent infections, and how modelling assumptions need to be adjusted to incorporate the data on recent infections. For instance, the assumed window period may be too long or short for a marker of recent infection to be consistent with the injecting duration-specific seroprevalence data, and some flexibility needs incorporating in order to accommodate this. This may again naturally be accomplished in a Bayesian framework, in which the window period could be assigned an informative, but flexible prior distribution.

The modelling and approaches developed in this thesis can be applied to other datasets. In particular Scotland's NESI programme is very similar to the UAM, consisting of time at risk, serological status and questionnaire-based risk factor information. Routine surveillance data may include little in the way of covariate information, but geographic differences and temporal changes may be considered, as has been here. Other seroprevalence surveys that include questionnaire data may also make use of the methods for analysing covariates in a FOI model. One potential area of application is sexual health, which might also include testing for multiple sexually acquired infections, which would allow the type of frailty models considered in this thesis. Of particular interest would be further investigation of trivariate models, which was hampered in this thesis by sparse data on HIV.

Trivariate frailty models are a novel development, and further investiga-

tion may be worthwhile. In practice, the model is limited by the ability to estimate only components that are shared by two or more infections, rather than infection-specific effects. Further, even with large amounts of data there are likely to be identifiability issues, although progress might be made by combining data from different countries. Much of the interpretation of the trivariate model amounts to the investigation of the correlation structure of trivariate infection status, and it may be more worthwhile to pursue methods of association for trivariate data, extending the definition of Unkel and Farrington (2012) beyond the bivariate case.

There is currently great interest in reducing the disease burden of viral hepatitis and HIV, with new treatments for HCV, effective HBV vaccines and high standards of HIV care making elimination of these infections as major public health threats a real prospect in the UK and elsewhere. Injecting drug use is the cause of a substantial proportion of blood-borne virus infections, and nearly all HCV infections in England. Monitoring the population of people who inject drugs, and understanding patterns of risk will be a crucial part of developing policy to tackle blood-borne viruses. The methods and analyses developed in this thesis may be used to this end. Force of infection models can be traced right back to Daniel Bernoulli's seminal 1766 paper on smallpox (Dietz, 2002). More than 250 years later, these methods are still proving useful. And they are far from being exhausted.

# Chapter 9

# Appendices

## 9.1 Unlinked Anonymous Monitoring survey of people who inject drugs: questionnaire

The Unlinked Anonymous Monitoring survey of people who inject drugs questionnaire is displayed in full below (2015 version). The document is printed double-sided on a single sheet of landscape A3 paper and folded in half to produce a 4-page A4 leaflet. Hence reading of the questionnaire as shown below begins on the right hand side of the first page, follows on the left then right sides of the second page and ends on the left side of the first page.

**NOTE: Sample questionnaire below has been reduced to A4 size**

15. Which of these health services have you used in the **last year** (12 months)?(*Tick **all** that apply*)

Sexual Health, GUM or STI Clinic ☐2 Family Planning Clinic ☐2 NHS Walk-in Clinic ☐2
A&E or Casualty Department ☐2 GP or Family Doctor ☐2 None of these services ☐2

16. Have you **ever** had a blood test for HIV?

Yes ☐2 No ☐1

*If yes:* In which year did you **last** have an HIV test? _____
What was the result of your **last** test?
Positive ☐1 Negative ☐2 Awaiting result ☐3

17. Have you **ever** had a blood test for hepatitis C?

Yes ☐2 No ☐1

*If yes:* In which year did you **last** have a hepatitis C test? _____
What was the result of your **last** test?
Positive ☐1 Negative ☐2 Awaiting result ☐3

*If tested positive,* have you **ever** seen a specialist nurse or doctor (e.g. a hepatologist) about your hepatitis C?
No ☐1 Yes, but **not** given any medicine for hepatitis C ☐2
Yes, and been **given** medicine for hepatitis C ☐3

18. Have you **ever** been in prison (or a young offenders institution)?

Yes ☐2 No ☐1

*If yes:* How many times have you been in prison? _____
How old were you when you **first** went to prison? _____
Have you **ever** injected drugs whilst in prison? Yes ☐2 No ☐1

19. Have you **ever** been homeless - that is living in a hostel, having no fixed abode, or living on the streets?

Yes ☐2 No ☐1

*If yes:* Have you been homeless during the **last 12 months**?
Yes, currently ☐3 Yes, but not currently ☐2 No ☐1

20. Were you born in the United Kingdom? (England, Scotland, Wales or Northern Ireland)

Yes ☐2 No ☐1

*If no:* What country were you born in? _____

**Thank you for completing this questionnaire.**

Please fold the questionnaire back up and then place it inside the brown envelope.

DO NOT SEAL THE ENVELOPE.

Type 15 (2014)

UAM Survey of PWID

# CONFIDENTIAL

Survey of People Who Inject Drugs

Centre: _____

Lab use barcode

**This questionnaire is completely anonymous: we do not wish to know your name or any other form of identification.**

The information from this study can help to develop better services for people who inject drugs.

Please ☑ the box or write in your answer.

Please try to answer all the questions.

1a. Have you **ever** injected any drug? No ☐1 Yes ☐2

1b. How old were you when you **first** injected? _____

2. How old are you now? _____

3. Are you? Male ☐1 Female ☐2

4. Have you done this survey before? No ☐1 Yes ☐2 → *If yes,* in which year did you last take part? _____

5. Have you **ever** used a Needle Exchange (*including a pharmacy exchange*)?

Yes ☐2 No ☐1

*If yes:* How old were you when you **first** used a Needle Exchange? _____
Have you used a Needle Exchange in the **last 12 months**? Yes ☐2 No ☐1

*If yes:* How many *individual* needles (including ones attached to syringes) did you get from Needle Exchanges during the **last month** (28 days)? _____
How many of these needles were **already attached** to syringes (barrels)? _____

6. Have you **ever** been prescribed a detox or maintenance drug regime?

Yes ☐2 No ☐1

*If yes:* How old were you when you were **first** prescribed a detox or maintenance drug regime (script)? _____
Are you currently being prescribed a detox or maintenance drug regime?
Yes ☐2 No ☐1

*If yes,* how long have you been on your **current** regime (script)?
Less than a month ☐1 1 to 6 months ☐2 Over 6 months ☐3

***Please turn over***

**NOTE: Sample questionnaire below has been reduced to A4 size**

7. Have you injected drugs in the **last year** (12 months)?

Yes ☐2 No ☐1 →

*If no*: In which year did you last inject drugs? _____
How old were you when you last injected drugs? _____
Which drug were you injecting most often? _____
*If you have **not** injected in the last year, please go to Question 9.*

*If yes:*

In the **last year**, which of the following drugs have you injected? (*Tick **all** that apply*)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Heroin | ☐2 | Crack | ☐2 | Amphetamine (speed) | ☐2 | Ketamine | ☐2 |
| Methadone | ☐2 | Cocaine | ☐2 | Mephedrone (m-cat) | ☐2 | Other Drugs | ☐2 |

In the **last year**, have you had a swelling containing pus (abscess), sore, or open wound at an injection site? Yes ☐2 No ☐1

In the **last year**, did you inject with a needle or syringe that had already been used by someone else (*including your partner*)? Yes ☐2 No ☐1

8. Have you injected drugs in the **last month** (28 days)?

Yes ☐2 No ☐1 → *If you have **not** injected in the last month please go to Question 9.*

*If yes:*

In the **last month** (28 days), on how many days have you injected drugs? _____

On the **last full day** that you injected, how many times did you inject drugs? _____

In the **last month** (28 days), into which parts of your body did you inject drugs? (*Tick **all** that apply*)

Arms ☐2 Hands ☐2 Groin ☐2 Legs ☐2 Feet ☐2 Neck ☐2 Other ☐2

In the **last month** (28 days), which **drug** have you injected most often? _____

In the **last month** (28 days), which of these drugs have you injected? (*Tick **all** that apply*)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Heroin | ☐2 | Crack | ☐2 | Amphetamine (speed) | ☐2 | Ketamine | ☐2 |
| Methadone | ☐2 | Cocaine | ☐2 | Mephedrone (m-cat) | ☐2 | Other Drugs | ☐2 |

| In the **last month** (28 days): | None | 1 | 2 or more |
|---|---|---|---|
| **To** how many people have you passed on used needles or syringes (*including your partner*)? | ☐0 | ☐1 | ☐2 |
| **From** how many people have you received used needles or syringes (*including your partner*)? | ☐0 | ☐1 | ☐2 |

In the **last month** (28 days), did you use spoons or other containers for mixing which had previously been used by someone else (*including your partner*)? Yes ☐2 No ☐1

In the **last month** (28 days), did you use filters which had previously been used by someone else (*including your partner*)? Yes ☐2 No ☐1

In the **last month** (28 days), did you inject with a needle or syringe after it had been cleaned (e.g. with water, bleach or detergent)? Yes ☐2 No ☐1

***Please go to top of the next page***

9. Have you **ever** received used needles or syringes from anyone?

Yes ☐2 No ☐1

10. In the **last month** (28 days), have you? (*Tick **all** that apply*)

| | | | |
|---|---|---|---|
| Snorted Cocaine | ☐2 | Drunk, snorted or swallowed Amphetamine (speed) | ☐2 |
| Smoked Crack | ☐2 | Smoked, chased or snorted Heroin | ☐2 |
| Smoked or swallowed Cannabis | ☐2 | Snorted, smoked or swallowed Mephedrone (m-cat) | ☐2 |
| Inhaled Solvents or Glue | ☐2 | Drunk, snorted or swallowed Ketamine | ☐2 |
| Taken Ecstasy/'E' (MDMA) | ☐2 | Swallowed non-prescribed Benzodiazepines | ☐2 |
| Done none of these | ☐2 | | |

11. In the **last year** (12 months), have you overdosed (OD-ed, gone-over, gone-under) to the point where you have lost consciousness?

Yes ☐2 No ☐1

*If yes:* How many times in the **last 12 months** have you overdosed?

1 ☐1 2-4 ☐2 5-9 ☐3 10 or more ☐4

In the **last 12 months,** did you receive naloxone (the heroin overdose antidote) when you overdosed?

Yes ☐2 No ☐1 Not Sure ☐3

12. Have you had sex (vaginal or anal) in the **last year** (12 months)?

Yes ☐2 →

*If yes:* With how many **men** in the last year?

None ☐0 1 ☐1 2-4 ☐4 5-9 ☐5 10 or more ☐3

With how many **women** in the last year?

None ☐0 1 ☐1 2-4 ☐4 5-9 ☐5 10 or more ☐3

Did you use a condom? Always ☐1 Sometimes ☐2 Never ☐3

No ☐1

13. Have you **ever** received money, goods or drugs in exchange for sex?

Yes, in the last year ☐4 Yes, but **not** in the last year ☐3 Never ☐1

14. Have you **ever** been vaccinated for hepatitis B (hep B jab)?

Yes ☐2 No ☐1 Not sure ☐3

*If yes:* How many hep B jabs have you had?

1 ☐1 2 ☐2 3+ ☐3 Not sure ☐4

At which of the following services did you receive a hep B jab? (*Tick **all** that apply*)

| | | | |
|---|---|---|---|
| Needle Exchange | ☐2 | Drug Treatment Service | ☐2 |
| Sexual Health, GUM or STI Clinic | ☐2 | In Prison | ☐2 |
| Hostel or Homeless Service | ☐2 | GP or Family Doctor | ☐2 |
| A&E or Casualty Department | ☐2 | Elsewhere | ☐2 |

***Please turn over***

## 9.2 R code

### 9.2.1 Univariate multiplicative model code

Shown below is the R code for a univariate, piecewise constant force of infection model. Injecting duration ("age") and time combine multiplicatively, and the model includes *injecting duration* X *time* interaction terms. The model uses a dataset that includes the time at risk contributions to each piecewise constant segment; for example, 2 years exposure in the 5-10 year injecting duration category, 3 years exposure in the 1995-1999 time category and so on, which contribute to the cumulative hazard. The code itself for forming the cumulative hazard is not very concise: the sums of parameters multiplied by time at risk contributions is written out term-by-term.

```
## y=outcome, n=denominator, s=sensitivity of test,
## data=exposure time contributions, inits=list of parameters
univarAT <- function(y,n,s,data,inits){

  sumht <- array(0,c(1,length(y)))
  llc <- rep(NA,length(y))

  ## 8 time periods: pre-1980 then 5 year
  ## first four have same injdur effect
  ## 7 injdur periods, baseline injdur is 3 (4-5 yrs)
  ## decompose init list
  Bt <- array(0,8)
  Ba <- array(0,7)
  Bta <- array(0,c(8,7))

  Bt <- inits[1:8]
  Ba[1:2] <- inits[9:10]
  Ba[4:7] <- inits[11:14]
  for(i in 5:8){
    Bta[i,1] <- inits[15+(i-5)*6]
    Bta[i,2] <- inits[16+(i-5)*6]
```

```
  Bta[i,4] <- inits[17+(i-5)*6]
  Bta[i,5] <- inits[18+(i-5)*6]
  Bta[i,6] <- inits[19+(i-5)*6]
  Bta[i,7] <- inits[20+(i-5)*6]
}

## hazards for each infection
## age and time parameters combine multiplicativel
ht <- array(0,c(8,7))
for(t in 1:8){
  for(a in 1:7){
    ht[t,a] <- exp(Bt[t] + Ba[a] + Bta[t,a])
  }
}


## contributions to cumulative hazard
## could be written more efficiently!
sumht[] <- (
  ht[1,1]*data$x1_1+  ht[1,2]*data$x1_2+  ht[1,3]*data$x1_3+
    ht[1,4]*data$x1_4+  ht[1,5]*data$x1_5+  ht[1,6]*data$x1_6+
    ht[1,7]*data$x1_7+
    ht[2,1]*data$x2_1+  ht[2,2]*data$x2_2+  ht[2,3]*data$x2_3+
    ht[2,4]*data$x2_4+  ht[2,5]*data$x2_5+  ht[2,6]*data$x2_6+
    ht[2,7]*data$x2_7+
    ht[3,1]*data$x3_1+  ht[3,2]*data$x3_2+  ht[3,3]*data$x3_3+
    ht[3,4]*data$x3_4+  ht[3,5]*data$x3_5+  ht[3,6]*data$x3_6+
    ht[3,7]*data$x3_7+
    ht[4,1]*data$x4_1+  ht[4,2]*data$x4_2+  ht[4,3]*data$x4_3+
    ht[4,4]*data$x4_4+  ht[4,5]*data$x4_5+  ht[4,6]*data$x4_6+
    ht[4,7]*data$x4_7+
    ht[5,1]*data$x5_1+  ht[5,2]*data$x5_2+  ht[5,3]*data$x5_3+
    ht[5,4]*data$x5_4+  ht[5,5]*data$x5_5+  ht[5,6]*data$x5_6+
    ht[5,7]*data$x5_7+
    ht[6,1]*data$x6_1+  ht[6,2]*data$x6_2+  ht[6,3]*data$x6_3+
```

```
      ht[6,4]*data$x6_4+  ht[6,5]*data$x6_5+  ht[6,6]*data$x6_6+
      ht[6,7]*data$x6_7+
      ht[7,1]*data$x7_1+  ht[7,2]*data$x7_2+  ht[7,3]*data$x7_3+
      ht[7,4]*data$x7_4+  ht[7,5]*data$x7_5+  ht[7,6]*data$x7_6+
      ht[7,7]*data$x7_7+
      ht[8,1]*data$x8_1+  ht[8,2]*data$x8_2+  ht[8,3]*data$x8_3+
      ht[8,4]*data$x8_4+  ht[8,5]*data$x8_5+  ht[8,6]*data$x8_6+
      ht[8,7]*data$x8_7)


  ################################
  ## probabilities and likelihood

  pi_tru0 <- exp(-sumht[])
  pi_tru1 <- 1 - pi_tru0

  ## relate true to observed via sensitivity
  pi1 <- s*pi_tru1
  pi0 <- 1-pi1

  ## likelihood - negative as optim minimises function
  llc <- log(pi0)*(n-y) + log(pi1)*(y)
  loglik <- -sum(llc)
  return(loglik)

}
```

### 9.2.2 Bivariate shared frailty model code

Shown below is the R code used to fit bivariate frailty models. The code is more general-purpose than the univariate model in section 9.2.1 and can be used to fit bivariate models under independence, gamma and inverse Gaussian frailty distributions, and also allows for different frailty variances across subgroups (strata). The code also allows for an arbitrary covariate

matrix, which can also be specified differently for each infection. In order to achieve this flexibility a "multi-row" formulation of the data, in which the time at risk for an individual is split according to changing covariate combinations (described subsequently).

```
# fast tapply method
library("ecoreg",lib="C:/My programs/R library")

#################################################################
## multirow model
## - split age/time covariate levels into repeat observations
## y1, y2: infection status for 1 & 2
## s1, s2: sensitivity of tests for infection 1&2
## X1, X2: covariate matrix for infection risk
## frstr: frailty variance strata (constant for single frailty)
## inits: parameter list
## frdist: specified distribution
## (indep, gamma, inverse Gaussian)
## ret: log-likelihood for optim, or predictions/AIC

modelMR <- function(y1,y2,s1,s2,X1,X2,frstr,inits,frdist,ret){

  llc <- rep(NA,length(y1))

  ## X - covar dataset consisting of ID, j (sub ID)
  ## , and T (time within band)
  ## subset inits - based on width of covar mat
  clen1 <- dim(X1)[2]-3 ## width minus references and T
  beta1 <- inits[1:clen1]
  clen2 <- dim(X2)[2]-3 ## width minus references and T
  beta2 <- inits[(clen1+1):(clen1+clen2)]
  ## remaining portion of inits are frailty variance(s)
  theta <- inits[(clen1+clen2+1):length(inits)]
  thstr <- theta[frstr]
  table(thstr)
```

```
## exposure X beta coefs
betalist1 <- t(t(X1[,4:(clen1+3)])*beta1)
betalist2 <- t(t(X2[,4:(clen2+3)])*beta2)
## sum of beta coefs
betasum1<-rowSums(betalist1)
betasum2<-rowSums(betalist2)
## combine with ID var and exponentiate, multiply by T
htx1<-cbind(X1[,1],exp(betasum1)*X1[,3])
htx2<-cbind(X2[,1],exp(betasum2)*X2[,3])
## sum over IDs - aggregate and tapply both slow
## - this is a special version where groups are ordered
## - much faster
At1<-tapplysum.fast(htx1[,2],htx1[,1])
At2<-tapplysum.fast(htx2[,2],htx2[,1])

#########################################################
## multinomial probabilities with different frailties

## No frailty
if(frdist=="INDEP"){
  pi_tru00 <- exp(-At1-At2)
  pi_tru10 <- exp(-At2) - pi_tru00
  pi_tru01 <- exp(-At1) - pi_tru00
  pi_tru11 <- 1-pi_tru00-pi_tru10-pi_tru01
}
## GAMMA
if(frdist=="GAM"){
  pi_tru00 <- (1+(At1+At2)/thstr)^-thstr
  pi_tru10 <- (1+At2/thstr)^-thstr - pi_tru00
  pi_tru01 <- (1+At1/thstr)^-thstr - pi_tru00
  pi_tru11 <- 1-pi_tru00-pi_tru10-pi_tru01
}
## INVERSE GAUSSIAN
if(frdist=="INVG"){
```

```
  pi_tru00 <- exp(thstr*(1- sqrt(1+2*(At1+At2)/thstr) ))
  pi_tru10 <- exp(thstr*(1- sqrt(1+2*(At2)/thstr) )) - pi_tru00
  pi_tru01 <- exp(thstr*(1- sqrt(1+2*(At1)/thstr) )) - pi_tru00
  pi_tru11 <- 1-pi_tru00-pi_tru10-pi_tru01
}

## Relate true proportions to observed through sensitivity of test
## Sutton BMC 2006, p3
pi00 <- pi_tru00 + (1-s1)*pi_tru10 + (1-s2)*pi_tru01
+ (1-s1)*(1-s2)*pi_tru11
pi10 <- pi_tru10*s1 + s1*(1-s2)*pi_tru11
pi01 <- pi_tru01*s2 + s2*(1-s1)*pi_tru11
pi11 <- pi_tru11*s1*s2

pi1_0 <- pi00+pi01
pi1_1 <- pi10+pi11
pi2_0 <- pi00+pi10
pi2_1 <- pi01+pi11

## full likelihood where both infections present
llc <- ifelse(is.na(y1) | is.na(y2), 0, log(pi00)*((1-y1)*(1-y2)) +
log(pi10)*(y1*(1-y2)) + log(pi01)*((1-y1)*y2) + log(pi11)*(y1*y2) )
## partial likelihood where one infection missing - y1 present only
llc <- ifelse(is.na(y2),
  llc+log(pi1_0)*(1-y1) + log(pi1_1)*(y1), llc)
## y2 only
llc <- ifelse(is.na(y1),
  llc+log(pi2_0)*(1-y2) + log(pi2_1)*(y2), llc)

loglik <- -sum(llc)
## Minus LL for optimisation
if(ret=="LL"){
  return(loglik)
}
## return predicted probabilities at fixed parameters
```

```
  if(ret=="pred"){
    return(cbind(pi00,pi10,pi01,pi11))
  }
  ## or AIC
  if(ret=="AIC"){
    npar<-length(inits)
    if(frdist=="INDEP"){
      npar<-npar-1
    }
    AIC<- 2*loglik+2*npar
    return(AIC)
  }

}
```

Shown below is an excerpt of the covariate data used for the "multi-row" formulation. The first column, **id**, is an identifier for the individual. **j** is a counter for the time interval, and **t** is the amount of time spent by the individual in a particular covariate combination. The **a_** and **t_** are indicator variables specifying which age (injecting duration) and calendar time effects the individual was subjected to within the time interval; the time terms governing the log baseline FOI and the age/injecting duration effects are log hazard ratios compared to the baseline category, 3.

For example, **id**=1 is subjected to the baseline hazard for the time interval 4 (**t_4**) with the change in hazard associated with **a_1** for one year (**j**=1), then spends 2 years in time interval 4 with hazard associated with **a_2** (**j**=2). At (**j**=3) the individual is in the baseline age/injecting duration category for 1 year, and at (**j**=4) the individual moves to the next time interval (**t_5**) for one year.

Additional covariate effects are age at first use groups (**fage1** and **fage3**, vs. a baseline group 2) and **female** (vs. male). Parameters for the log FOI and log hazard ratios are then multiplied by the indicators, and the sum of the products is thus the log FOI for that covariate combination. The log FOI is exponentiated for the multiplicative effects model. The resulting FOIs are multiplied by the time at risk in that covariate combination (**t**) and the sum of the FOI contributions across the individual is equal to the cumulative hazard.

Additional covariates are easy to include, and in particular interaction effects may be obtained by multiplying columns (for instance, an interaction between **female** and the age at first use variables).

| id | j | t | a_1 | a_2 | a_4 | a_5 | a_6 | a_7 | t_1 | t_2 | t_3 | t_4 | t_5 | t_6 | t_7 | t_8 | fage1 | fage3 | female |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 5 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 5 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

### 9.2.3 Bivariate time-varying frailty code

Shown below is the R code for the time-varying frailty model. This has some additional complications, as the calculation of the cumulative hazards for each infection cannot be performed separately, as in the simpler bivariate frailty model (section 9.2.2) and also involves the frailty parameters.

```
#################################################################
## takes additional argument frstrC for
## frailty strata within covariate data
modelMRTVF <- function(y1,y2,s1,s2,X1,X2,frstr,frstrC,inits,ret){

  llc <- rep(NA,length(y1))

  ## X - covar dataset consisting of ID, j (sub ID)
  ##, and T (time within band)
  clen1 <- dim(X1)[2]-5 ## width minus references and T
  beta1 <- inits[1:clen1]
  clen2 <- dim(X2)[2]-5 ## width minus references and T
  beta2 <- inits[(clen1+1):(clen1+clen2)]
  ## remaining portion of inits assigned to frailty variance
  ## 3 parts: theta, rho, q
  frlen <- length(inits)-clen1-clen2
  frN <- frlen/3
  theta <- inits[(clen1+clen2+1):(clen1+clen2+frN)]
  rho <- inits[(clen1+clen2+frN+1):(clen1+clen2+frN*2)]
  q <- inits[(clen1+clen2+frN*2+1):(length(inits))]
  ## theta is needed for both outcome and multirow form
  thstr <- theta[frstr]
  thstrC <- theta[frstrC]
  ## rho and q are at at the multirow level
  rhostr <- rho[frstrC]
  qstr <- q[frstrC]

  ## time and other vars - note that the covariate matrices
  ## must have matching time cutpoints
```

```
id <- X1[,1]
t <- X1[,3]
c_i <- X1[,4]
v_i <- X1[,5]

## exposure X beta coefs
betalist1 <- t(t(X1[,6:(clen1+5)])*beta1)
betalist2 <- t(t(X2[,6:(clen2+5)])*beta2)
## sum of beta coefs
betasum1<-rowSums(betalist1)
betasum2<-rowSums(betalist2)
## ht vars - used alone with t in places
ht<-array(NA,dim=c(dim(X1)[1],3))
ht[,1]<-exp(betasum1)
ht[,2]<-exp(betasum2)
ht[,3]<-ht[,1]+ht[,2]

## Derive the other variables
eqq <- exp(qstr)/(1+exp(qstr))
epv <- exp(-rhostr*v_i)
epc <- exp(-rhostr*c_i)

## loop over FOIs for 1, 2, combined (3)
## setup arrays
S<-array(NA,dim=c(dim(X1)[1],3,4))
S.sum<-array(NA,dim=c(length(y1),3,4))
St<-array(NA,dim=c(length(y1),3))

for(i in 1:3){
  S[,i,1] <- ht[,i]*t*(1-eqq)
  S[,i,2] <- ht[,i]*(epv-epc)/(rhostr*(1+exp(qstr)))
  S[,i,3] <- ht[,i]*(epc-epv)/(rhostr*thstrC*(1+exp(qstr)))
  S[,i,4] <- ht[,i]*t*eqq/thstrC
  ## summations over multirow ID
  S.sum[,i,1] <- tapplysum.fast(S[,i,1],id)
```

```
  S.sum[,i,2] <- tapplysum.fast(S[,i,2],id)
  S.sum[,i,3] <- tapplysum.fast(S[,i,3],id)
  S.sum[,i,4] <- tapplysum.fast(S[,i,4],id)
  St[,i] <- exp(-S.sum[,i,1]-S.sum[,i,2])
      * (1+S.sum[,i,3]+S.sum[,i,4])^(-thstr)
}
#########################################################
## multinomial probabilities
pi_tru00 <- St[,3]
pi_tru10 <- St[,2] - pi_tru00
pi_tru01 <- St[,1] - pi_tru00
pi_tru11 <- 1-pi_tru00-pi_tru10-pi_tru01

## sensitivity - Sutton BMC 2006, p3
pi00 <- pi_tru00 + (1-s1)*pi_tru10 + (1-s2)*pi_tru01
  + (1-s1)*(1-s2)*pi_tru11
pi10 <- pi_tru10*s1 + s1*(1-s2)*pi_tru11
pi01 <- pi_tru01*s2 + s2*(1-s1)*pi_tru11
pi11 <- pi_tru11*s1*s2

pi1_0 <- pi00+pi01
pi1_1 <- pi10+pi11
pi2_0 <- pi00+pi10
pi2_1 <- pi01+pi11

## full likelihood where both infections present
llc <- ifelse(is.na(y1) | is.na(y2), 0, log(pi00)*((1-y1)*(1-y2)) +
log(pi10)*(y1*(1-y2)) + log(pi01)*((1-y1)*y2) + log(pi11)*(y1*y2) )
## partial likelihood where one infection missing - y1 present only
llc <- ifelse(is.na(y2),
  llc+log(pi1_0)*(1-y1) + log(pi1_1)*(y1), llc )
## y2 only
llc <- ifelse(is.na(y1),
  llc+log(pi2_0)*(1-y2) + log(pi2_1)*(y2), llc )
```

```
  loglik <- -sum(llc)

  if(ret=="LL"){
    return(loglik)
  }
  if(ret=="pred"){
    return(cbind(pi00,pi10,pi01,pi11))
  }
  if(ret=="AIC"){
    npar<-length(inits)
    AIC<- 2*loglik+2*npar
    return(AIC)
  }

}
```

### 9.2.4 Trivariate frailty model code

Shown below is R code for the asymmetric trivariate frailty model, with multiplicative components for frailty specific to all infections, and a component relating to HBV-HIV only (or other pair). The model requires numerical integration of the survivor functions. Model fitting is computationally intensive, due to numerical integration being carried out at each step of the numerical optimisation routine. Data are aggregated by injecting duration (age) only, unlike earlier models which examined individual level data including time and other covariates.

```
## numerical integration package
library(cubature)

## survivor functions for numerical integration
## trivariate
St123 <- function(x, A1,A2,A3,k1,k2){
  exp(-k2*x)*(k1^k1)*(k2^k2)*(x^(k2-1))
   / (gamma(k2)*(A1+A2*x+A3*x+k1)^k1)
```

```
}
## bivariate - same form for s13 and s23 (HBV & HIV)
## but spell out in case of any confusion
St12 <- function(x, A1,A2,k1,k2){
  exp(-k2*x)*(k1^k1)*(k2^k2)*(x^(k2-1))
    / (gamma(k2)*(A1+A2*x+k1)^k1)
}
St13 <- function(x, A1,A3,k1,k2){
  exp(-k2*x)*(k1^k1)*(k2^k2)*(x^(k2-1))
    / (gamma(k2)*(A1+A3*x+k1)^k1)
}
St23 <- function(x, A2,A3,k1,k2){
  exp(-k2*x)*(k1^k1)*(k2^k2)*(x^(k2-1))
    / (gamma(k2)*((A2+A3)*x+k1)^k1)
}
## univariate - HCV does not require integration
St1 <- function(A1,k1){
  (1+A1/k1)^(-k1)
}
St2 <- function(x, A2,k1,k2){
  exp(-k2*x)*(k1^k1)*(k2^k2)*(x^(k2-1))
    / (gamma(k2)*(A2*x+k1)^k1)
}
St3 <- function(x, A3,k1,k2){
  exp(-k2*x)*(k1^k1)*(k2^k2)*(x^(k2-1))
    / (gamma(k2)*(A3*x+k1)^k1)
}

## y is the dataset and includes time at risk, infection status
## (s000, s001, s010 etc.) and test sensitivities
trivarM4 <- function(y,inits,ret){

  ## inits and quantities for calculation
  N <- dim(y)[1]
  llc <- rep(NA,N)
```

```
len<-(length(inits)-2)/3
ht <- matrix(data=NA, nrow=3, ncol=len)
ht[1,] <-exp(inits[1:len])
ht[2,] <-exp(inits[(len+1):(2*len)])
ht[3,] <-exp(inits[(2*len+1):(3*len)])
k1<-exp(inits[(3*len+1)])  ## exp. to ensure positive
k2<-exp(inits[(3*len+2)])

## cumulative hazard functions
At<-matrix(data=NA, nrow=N, ncol=3)
for(j in 1:3){
  At[,j]<-ht[j,1]+ht[j,2]*(y$t2)+ht[j,3]*(y$t3)+ht[j,4]*(y$t4)
      + ht[j,5]*(y$t5)+ht[j,6]*(y$t6)+ht[j,7]*(y$t7)
      + ht[j,8]*(y$t8)+ht[j,9]*(y$t9)
}

## survivor functions
## upper limits of integrals are 1+ 10 times gamma variance
## or 5, whichever is greater
ilim <- 1+10/c(k1,k2)
ilim <- pmax(ilim,5)
S123<-rep(NA,N)
S12<-rep(NA,N)
S13<-rep(NA,N)
S23<-rep(NA,N)
S1<-rep(NA,N)
S2<-rep(NA,N)
S3<-rep(NA,N)

## numerical integration of survivor functions
for(i in 1:N){
  S123[i] <-integrate(St123, 0,ilim[2],
      At[i,1],At[i,2],At[i,3], k1,k2)$val
  S12[i] <-integrate(St12, 0,ilim[2],
      At[i,1],At[i,2], k1,k2)$val
```

```
  S13[i] <-integrate(St13, 0,ilim[2],
      At[i,1],At[i,3], k1,k2)$val
  S23[i] <-integrate(St23, 0,ilim[2],
      At[i,2],At[i,3], k1,k2)$val
  S1[i] <- St1(At[i,1], k1)
  S2[i] <- integrate(St2, 0,ilim[2], At[i,2], k1,k2)$val
  S3[i] <- integrate(St3, 0,ilim[2], At[i,3], k1,k2)$val
}
pi_tru000 <- pmax(S123,0.0000001)
pi_tru100 <- pmax((S23-pi_tru000) ,0.0000001)
pi_tru010 <- pmax((S13-pi_tru000) ,0.0000001)
pi_tru001 <- pmax((S12-pi_tru000) ,0.0000001)

pi_tru110 <- pmax((S3-pi_tru000-pi_tru100-pi_tru010) ,0.0000001)
pi_tru101 <- pmax((S2-pi_tru000-pi_tru100-pi_tru001) ,0.0000001)
pi_tru011 <- pmax((S1-pi_tru000-pi_tru001-pi_tru010) ,0.0000001)
pi_tru111 <- pmax((1-pi_tru000-pi_tru100-pi_tru010-pi_tru001
-pi_tru110-pi_tru101-pi_tru011) ,0.0000001)

## sensitivity of tests
pi000 <- pi_tru000 + (1-y$senshcv)*pi_tru100
 + (1-y$senshbv)*pi_tru010 + (1-y$senshiv)*pi_tru001
 + (1-y$senshcv)*(1-y$senshbv)*pi_tru110
 + (1-y$senshcv)*(1-y$senshiv)*pi_tru101
 + (1-y$senshbv)*(1-y$senshiv)*pi_tru011
 + (1-y$senshcv)*(1-y$senshbv)*(1-y$senshiv)*pi_tru111

pi100 <- pi_tru100*y$senshcv
  + y$senshcv*(1-y$senshbv)*pi_tru110
  + y$senshcv*(1-y$senshiv)*pi_tru101
  + y$senshcv*(1-y$senshbv)*(1-y$senshiv)*pi_tru111
pi010 <- pi_tru010*y$senshbv
  + y$senshbv*(1-y$senshcv)*pi_tru110
  + y$senshbv*(1-y$senshiv)*pi_tru011
  + y$senshbv*(1-y$senshcv)*(1-y$senshiv)*pi_tru111
```

```
  pi001 <- pi_tru001*y$senshiv
    + y$senshiv*(1-y$senshcv)*pi_tru101
    + y$senshiv*(1-y$senshbv)*pi_tru011
    + y$senshiv*(1-y$senshcv)*(1-y$senshbv)*pi_tru111

  pi110 <- pi_tru110*y$senshcv*y$senshbv
    + y$senshcv*y$senshbv*(1-y$senshiv)*pi_tru111
  pi101 <- pi_tru101*y$senshcv*y$senshiv
    + y$senshcv*y$senshiv*(1-y$senshbv)*pi_tru111
  pi011 <- pi_tru011*y$senshbv*y$senshiv
    + y$senshbv*y$senshiv*(1-y$senshcv)*pi_tru111

  pi111 <- pi_tru111*y$senshcv*y$senshbv*y$senshiv

  if(ret=="pred"){
    pred<-cbind(pi000,pi100,pi010,pi001,pi110,pi101,pi011,pi111)
    return(pred)
  }
  if(ret=="LL"){
    llc <- log(pi000)*y$s000 +log(pi001)*y$s001
        +log(pi010)*y$s010 +log(pi100)*y$s100
        +log(pi011)*y$s011 +log(pi101)*y$s101
        +log(pi110)*y$s110 +log(pi111)*y$s111
    loglik <- -sum(llc)
    return(loglik)
  }

}
```

## 9.3 Conference presentations

Ross Harris, Daniela De Angelis, Vivian Hope and Paddy Farrington. Analysis of serial cross-sectional data: Risk patterns of blood-borne viruses in people who inject drugs. Applied Epidemiology Scientific Meeting 2015, Warwick University.

Ross Harris, Daniela De Angelis, Vivian Hope and Paddy Farrington. Estimating the force of infection via generalised linear models: hepatitis C infection in people who inject drugs. Public Health Research and Science Conference 2016, Warwick University.

Ross Harris, Daniela De Angelis, Vivian Hope and Paddy Farrington. Estimating the force of infection of blood-borne viruses in people who inject drugs: risk factors and frailty. Royal Statistical Society Conference 2017, Glasgow.

Ross Harris, Daniela De Angelis, Helen Harris and Sema Mandal. Monitoring progress toward elimination of hepatitis C in England using routinely collected data. Public Health Research and Science Conference 2018, Warwick University.

# Bibliography


Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008), *Survival and Event History Analysis: A Process Point of View*, Springer.

Ades, A. E. and Medley, G. F. (1994), 'Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: HIV in New York City', *Statistics in Medicine* **13**(18), 1881–1894.

Ades, A. E. and Nokes, D. J. (1993), 'Modeling age- and time-specific incidence from seroprevalence: toxoplasmosis', *American Journal of Epidemiology* **137**(9), 1022–1034.

Afdhal, N., Zeuzem, S., Kwo, P., Chojkier, M., Gitlin, N., Puoti, M., Romero-Gomez, M., Zarski, J.-P., Agarwal, K., Buggisch, P., Foster, G. R., Bräu, N., Buti, M., Jacobson, I. M., Subramanian, G. M., Ding, X., Mo, H., Yang, J. C., Pang, P. S., Symonds, W. T., McHutchison, J. G., Muir, A. J., Mangia, A. and Marcellin, P. (2014), 'Ledipasvir and Sofosbuvir for untreated HCV denotype 1 infection', *New England Journal of Medicine* **370**(20), 1889–1898.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–724.

Allen, E., Taylor, A., Rees, C., Palmateer, N., Hutchinson, S., Mathieson, A., Stewart, G., Cameron, S. and Goldberg, D. (2010), 'Needle Exchange Surveillance Initiative (NESI): Prevalence of HCV and injecting risk behaviours among people who inject drugs attending injecting equipment provision services in Scotland, 2008/2009 & 2010'.
**URL:** *http://www.hepatitisscotlandc.org.uk/media/51339/nesi report booklet web.pdf*

Alliance - CAHR (2013), 'Reaching Drug Users: A toolkit for outreach services'.
**URL:** *https://www.aidsalliance.org/assets/000/000/380/90629-Reaching-drug-users-A-toolkit-for-outreach-services_original.pdf?1405520658*

Altman, D. and Bland, M. (1994), 'Statistics Notes: Diagnostic tests I: sensitivity and specificity', *BMJ* **308**(March 2009), 1552.

Amin, J., Gidding, H., Gilbert, G., Backhouse, J., Kaldor, J., Dore, G. and Burgess, M. (2004), 'Hepatitis C prevalence - a nationwide serosurvey', *Communicable Diseases Intelligence* **28**(4), 517–521.

Aranda-Ordaz, F. J. (1981), 'On two families of transformations to additivity for binary response data', *Biometrika* **68**(2), 357–363.

Baggaley, R. F., Boily, M.-C., White, R. G. and Alary, M. (2006), 'Risk of HIV-1 transmission for parenteral exposure and blood transfusion: a systematic review and meta-analysis', *AIDS* **20**(6), 805–812.

Balogun, M. A., Ramsay, M. E., Parry, J. V., Donovan, L., Andrews, N. J., Newham, J. A., Cliffe, S., Harris, K. A. and Teo, C. G. (2000), 'The prevalence and genetic diversity of hepatitis C infection in antenatal clinic attenders in two regions of England', *Epidemiology and Infection* **125**(3), 705–712.

Balogun, M., Ramsay, M. E., Parry, J. V., Donovan, L., Andrews, N. J., Newham, J., McGarrigle, C., Harris, K. and Teo, C. (2003), 'A national survey of genitourinary medicine clinic attenders provides little evidence of sexual transmission of hepatitis C virus infection', *Sexually Transmitted Infections* **79**(4), 301–306.

Bargagli, A. M., Hickman, M., Davoli, M., Perucci, C. A., Schifano, P., Buster, M., Brugal, T. and Vicente, J. (2006), 'Drug-related mortality and its impact on adult mortality in eight European countries', *European Journal of Public Health* **16**(2), 198–202.

Besag, J., York, J. and Mollie, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics* **43**(1), 1–20.

Birrell, P. J., Gill, O. N., Delpech, V. C., Brown, A. E., Desai, S., Chadborn, T. R., Rice, B. D. and De Angelis, D. (2013), 'HIV incidence in men who have sex with men in England and Wales 2001-10: A nationwide population study', *The Lancet Infectious Diseases* **13**(4), 313–318.

BMA Board of Science (2013), *Drugs of dependence: the role of medical professionals*, BMA publications unit, UK.

**URL:** *https://www.bma.org.uk/collective-voice/policy-and-research/public-and-population-health/illicit-drug-use/full-report*

Bollaerts, K., Aerts, M., Shkedy, Z., Faes, C., Van der Stede, Y., Beutels, P. and Hens, N. (2012), 'Estimating the population prevalence and force of infection directly from antibody titres', *Statistical Modelling* **12**(5), 441–462.

Brewer, D. D. and Garrett, S. B. (2001), 'Evaluation of interviewing techniques to enhance recall of sexual and drug injection partners', *Sexually Transmitted Diseases* **28**(11), 666–677.

Brizzi, F. (2018), Estimating HIV incidence from multiple sources of data, PhD thesis, University of Cambridge.

Brookmeyer, R. and Gail, M. H. (1988), 'A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic', *Journal of the American Statistical Association* **83**(402), 301–308.

Burnham, K. and Anderson, D. (2002), *Model selection and multimodel inference: a practical information-theoretic approach*, Vol. 172, Springer.
**URL:** *http://linkinghub.elsevier.com/retrieve/pii/S0304380003004526*

Clayton, D. G. (1978), 'A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence', *Biometrika* **65**(1), 141–151.

Collett, D. (2002), *Modelling Binary Data, Second Edition*, 2nd edn, CRC Press.

Coppola, N., Pisapia, R., Tonziello, G., Masiello, A., Martini, S., Pisaturo, M., Messina, V., Sagnelli, C., Macera, M., Signoriello, G. and Sagnelli, E. (2009), 'Improvement in the aetiological diagnosis of acute hepatitis C: a diagnostic protocol based on the anti-HCV-IgM titre and IgG Avidity Index.', *Journal of Clinical Virology* **46**(3), 222–229.

Cox, D. R. (1972), 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society . Series B* **file:///C:**(2), 187–220.

Crofts, N., Jolley, D., Kaldor, J., van Beek, I. and Wodak, A. (1997), 'Epidemiology of hepatitis C virus infection among injecting drug users in Australia', *Journal of Epidemiology & Community Health* **51**(6), 692–697.

Cullen, K. J., Hope, V. D., Croxford, S., Shute, J., Ncube, F. and Parry, J. V. (2015), 'Factors associated with recently acquired hepatitis C virus infection in people who inject drugs in England, Wales and Northern Ireland: New findings from an unlinked anonymous monitoring survey', *Epidemiology and Infection* **143**(7), 1398–1407.

De Angelis, D., Hickman, M. and Yang, S. (2004), 'Estimating long-term trends in the incidence and prevalence of opiate use/injecting drug use and the number of former users: back-calculation methods and opiate overdose deaths', *American Journal of Epidemiology* **160**(10), 994–1004.

De Angelis, D., Presanis, A. M., Conti, S. and Ades, a. E. (2014), 'Estimation of HIV burden through Bayesian evidence synthesis', *Statistical Science* **29**(1), 9–17.

De Angelis, D., Sweeting, M., Ades, A., Hickman, M., Hope, V. and Ramsay, M. (2009), 'An evidence synthesis approach to estimating Hepatitis C prevalence in England and Wales', *Statistical Methods in Medical Research* **18**(4), 361–379.

de Vos, A. S., van der Helm, J. J., Prins, M. and Kretzschmar, M. E. (2012), 'Determinants of persistent spread of HIV in HCV-infected populations of injecting drug users.', *Epidemics* **4**(2), 57–67.

Del Fava, E. (2012), Statistical methods for modeling of drug-related and close-contact infections, PhD thesis, Universiteit Hasselt.

Del Fava, E., Kasim, A., Usman, M., Shkedy, Z., Hens, N., Aerts, M., Bollaerts, K., Scalia Tomba, G., Vickerman, P., Sutton, A. J., Wiessing, L. and Kretzschmar, M. (2011a), 'Joint modeling of HCV and HIV infections among injecting drug users in Italy using repeated cross-Sectional prevalence data', *Statistical Communications in Infectious Diseases* **3**(1).

Del Fava, E., Shkedy, Z., Hens, N., Aerts, M., Suligoi, B., Camoni, L., Vallejo, F., Wiessing, L. and Kretzschmar, M. (2011b), 'Joint modeling

of HCV and HIV co-infection among injecting drug users in Italy and Spain using individual cross-sectional data', *Statistical Communications in Infectious Diseases* **3**(1).

Denniston, M. M., Jiles, R. B., Drobeniuc, J., Klevens, M., Ward, J. W., Mcquillan, G. M. and Scott, D. (2014), 'Chronic hepatitis C virus infection in the United States, National Health and Nutrition Examination Survey 2003 to 2010', *Annals of Internal Medicine* **160**(5), 293–300.

Deuffic-Burban, S., Mathurin, P., Pol, S., Larsen, C., Roudot-Thoraval, F., Desenclos, J. C., Dhumeaux, D. and Yazdanpanah, Y. (2012), 'Impact of hepatitis C triple therapy availability upon the number of patients to be treated and associated costs in France: a model-based analysis.', *Gut* **61**(2), 290–296.

Dietz, K. (2002), 'Daniel Bernoulli's epidemiological model revisited', *Mathematical Biosciences* **180**, 1–21.

El-Serag, H. B. (2012), 'Epidemiology of viral hepatitis and hepatocellular carcinoma', *Gastroenterology* **142**(6), 1264–1273.

Enki, D. G., Noufaily, A. and Farrington, C. P. (2014), 'Time-varying shared frailty model with application to infectious diseases', *Annals of Applied Statistics* **8**(1), 430–447.

European Association for the Study of the Liver (2015), 'EASL recommendations on treatment of hepatitis C 2014', *Journal of Hepatology* **63**, 199–236.

Farrington, C. P. (1990), 'Modelling forces of infection for measles, mumps and rubella', *Statistics in Medicine* **9**(8), 953–967.

Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2001), 'Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data', *Journal of Applied Statistics* **50**(3), 251–292.

Farrington, C. P., Unkel, S. and Anaya-Izquierdo, K. (2012), 'The relative frailty variance and shared frailty models', *Journal of the Royal Statistical Society: Series B* **74**(4), 673–696.

Farrington, C. P., Whitaker, H. J., Unkel, S. and Pebody, R. (2013), 'Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases', *American Journal of Epidemiology* **177**(5), 474–486.

Fettig, J., Swaminathan, M., Murrill, C. S., Kaplan, J. E., Branch, S. I. and Branch, T. (2016), 'Global epidemiology of HIV', *Infectious Disease Clinics of North America* **28**(3), 323–337.

Fu, R., Gutfraind, A. and Brandeau, M. L. (2016), 'Modeling a dynamic bilayer contact network of injection drug users and the spread of blood-borne infections', *Mathematical Biosciences* **273**, 102–113.

Gaudy-Graffin, C., Lesage, G., Kousignian, I., Laperche, S., Girault, A., Dubois, F., Goudeau, A. and Barin, F. (2010), 'Use of an anti-hepatitis C virus (HCV) IgG avidity assay to identify recent HCV infection', *Journal of Clinical Microbiology* **48**(9), 3281–3287.

Genest, C. and Rivest, L. (1993), 'Statistical Inference Procedures for Bivariate Archimedean Copulas', *Journal of the American Statistical Association* **88**(423), 1034–1043.

Greenland, S. (2001), 'Putting background information about relative risks into conjugate prior distributions', *Biometrics* **57**(3), 663–670.

Harris, R. J., Hope, V. D., Morongiu, A., Hickman, M., Ncube, F. and De Angelis, D. (2012a), 'Spatial mapping of hepatitis C prevalence in recent injecting drug users in contact with services', *Epidemiology and Infection* **140**(6), 1054–1063.

Harris, R. J., Ramsay, M., Hope, V. D., Brant, L., Hickman, M., Foster, G. R. and De Angelis, D. (2012b), 'Hepatitis C prevalence in England remains low and varies by ethnicity: an updated evidence synthesis.', *European Journal of Public Health* **22**(2), 187–192.

Harris, R. J., Thomas, B., Griffiths, J., Costella, A., Chapman, R., Ramsay, M. E., De Angelis, D. and Harris, H. E. (2014), 'Increased uptake and new therapies are needed to avert rising hepatitis C-related end stage liver disease in England: modelling the predicted impact of treatment under different scenarios', *Journal of Hepatology* **61**, 530–537.

Hastie, T. J. and Tibshirani, R. (1986), 'Generalized additive models', *Statistical Science* **1**(3), 297–318.

Hay, G., dos Santos, A. R. and Worsley, J. (2012), 'Estimates of the prevalence of opiate use and/or crack cocaine use, 2011/12: Sweep 8 summary report'.
**URL:** *http://www.nta.nhs.uk/uploads/estimates-of-the-prevalence-of-opiate-use-and-or-crack-cocaine-use-2011-12-summary-report-v2.pdf*

Health Protection Scotland (2017), 'Needle Exchange Surveillance Initiative (NESI)'.
**URL:** *https://www.hps.scot.nhs.uk/bbvsti/resourcedetail.aspx?id=3186*

Hens, N., Wienke, A., Aerts, M. and Molenberghs, G. (2009), 'The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data', *Statistics in Medicine* **28**(July 2006), 2785–2800.

Hickman, M., Hope, V., Brady, T., Madden, P., Jones, S., Honor, S., Holloway, G., Ncube, F. and Parry, J. (2007), 'Hepatitis C virus (HCV) prevalence, and injecting risk behaviour in multiple sites in England in 2004', *Journal of Viral Hepatitis* **14**(9), 645–652.

Hickman, M., Hope, V., Coleman, B., Parry, J., Telfer, M., Twigger, J., Irish, C., Macleod, J. and Annett, H. (2009), 'Assessing IDU prevalence and health consequences (HCV, overdose and drug-related mortality) in a primary care trust: implications for public health action.', *Journal of Public Health* **31**(3), 374–382.

Hope, V. D., Harris, R. J., De Angelis, D., Croxford, S., Marongui, A., Parry, J. V. and Ncube, F. (2014), 'Two decades of successes and failures in controlling the transmission of HIV through injecting drug use in England and Wales: 1990 to 2011', *Eurosurveillance* **19**(14).

Hope, V. D., Judd, A., Hickman, M., Lamagni, T., Hunter, G., Stimson, G. V., Jones, S., Donovan, L., Parry, J. V. and Gill, O. N. (2001), 'Prevalence of hepatitis C among injection drug users in England and Wales: is harm reduction working?', *American Journal of Public Health* **91**(1), 38–42.

Hope, V. D., Judd, A., Hickman, M., Sutton, A., Stimson, G. V., Parry, J. V. and Gill, O. N. (2005), 'HIV prevalence among injecting drug users in England and Wales 1990 to 2003: evidence for increased transmission in recent years.', *AIDS* **19**(11), 1207–1214.

Hope, V. D., Ncube, F., Hickman, M., Judd, A. and Parry, J. V. (2007), 'Hepatitis B vaccine uptake among injecting drug users in England 1998 to 2004: is the prison vaccination programme driving recent improvements?', *Journal of Viral Hepatitis* **14**(9), 653–660.

Hope, V. D., Rogers, P. A., Jordan, L., Paine, T. C., Barnett, S., Parry, J. V. and Gill, O. N. (2002), 'Sustained increase in the sharing of needles and syringes among drug users in England and Wales', *AIDS* **16**, 2494–2496.

Hope, V., Parry, J. V., Marongui, A. and Ncube, F. (2012), 'Hepatitis C infection among recent initiates to injecting in England 2000-2008: is a national hepatitis C action plan making a difference?', *Journal of Viral Hepatitis* **19**(1), 55–64.

Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, Springer, New York.

Hutchinson, S. J., Bird, S. M. and Goldberg, D. J. (2005), 'Modeling the current and future disease burden of hepatitis C among injection drug users in Scotland.', *Hepatology* **42**(3), 711–723.

Iachine, I. A. (2004), 'Identifiability of bivariate frailty models based on additive independent components'.
**URL:** *https://pdfs.semanticscholar.org/c9d7/e63e186c963d001f0243e6a8b704eab3c0a0.pdf*

Judd, A., Hickman, M., Hope, V. D., Sutton, A. J., Stimson, G. V., Ramsay, M. E., Gill, O. N. and Parry, J. V. (2007), 'Twenty years of selective hepatitis B vaccination: is hepatitis B declining among injecting drug users in England and Wales?', *Journal of Viral Hepatitis* **14**(8), 584–591.

Judd, A., Parry, J., Hickman, M., McDonald, T., Jordan, L., Lewis, K., Contreras, M., Dusheiko, G., Foster, G., Gill, N., Kemp, K., Main, J.,

Murray-Lyon, I. and Nelson, M. (2003), 'Evaluation of a modified commercial assay in detecting antibody to hepatitis C virus in oral fluids and dried blood spots', *Journal of Medical Virology* **71**(1), 49–55.

Kaplan, E. H. (1997), 'Snapshot samples', *Socio-Economic Planning Sciences* **31**(4), 281–291.

Keiding, N. (1991), 'Age-specific incidence and prevalence: a statistical perspective', *Journal of the Royal Statistical Society. Series A* **154**(3), 371–412.

Keiding, N., Begtrup, K., Scheike, T. H. and Hasibeder, G. (1996), 'Estimation from current-status data in continuous time.', *Lifetime Data Analysis* **2**, 119–129.

Keiding, N., Holst, C. and Green, A. (1989), 'Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality', *American Journal of Epidemiology* **130**(3), 588–600.

Kessel, A. S. and Watts, C. J. (2001), 'Evaluation of the unlinked anonymous prevalence monitoring programme for HIV in England and Wales: science, ethics and health policy.', *Medical Science Monitor* **7**(5), 1052–1063.

Kimber, J., Copeland, L., Hickman, M., Macleod, J., McKenzie, J., De Angelis, D. and Robertson, J. R. (2010), 'Survival and cessation in injecting drug users: Prospective observational study of outcomes and effect of opiate substitution treatment', *BMJ* **341**, c3172.

Klimashevskaya, S., Obriadina, A., Ulanova, T., Bochkova, G., Burkov, A., Araujo, A., Stramer, S. L., Tobler, L. H., Busch, M. P. and Fields, H. A. (2007), 'Distinguishing acute from chronic and resolved hepatitis C virus (HCV) infections by measurement of anti-HCV immunoglobulin G avidity index.', *Journal of Clinical Microbiology* **45**(10), 3400–3403.

Kuha, J. (2004), 'AIC and BIC: Comparisons of assumptions and performance', *Sociological Methods and Research* **33**(2), 188–229.

Lemeshow, S. and Hosmer, D. W. (1982), 'A review of goodness of fit statistics for use in the development of logistic regression models.', *American Journal of Epidemiology* **115**(1), 92–106.

Mann, A. G., Ramsay, M. E., Brant, L. J., Balogun, M. A., Costella, A. and Harris, H. E. (2009), 'Diagnoses of, and deaths from, severe liver disease due to hepatitis C in England between 2000 and 2005 estimated using multiple data sources', *Epidemiology and Infection* **137**(4), 513–518.

Marschner, I. C. (1996), 'Fitting a multiplicative incidence model to age- and time-specific prevalence data', *Biometrics* **52**(2), 492–499.

Marschner, I. C. (1997), 'A method for assessing age-time disease incidence using serial prevalence data', *Biometrics* **53**(4), 1384–1398.

Martin, N. K., Foster, G. R., Vilar, J., Ryder, S., Cramp, M. E., Gordon, F., Dillon, J. F., Craine, N., Busse, H., Clements, A., Hutchinson, S. J., Ustianowski, A., Ramsay, M., Goldberg, D. J., Irving, W., Hope, V., De Angelis, D., Lyons, M., Vickerman, P. and Hickman, M. (2015), 'HCV treatment rates and sustained viral response among people who inject drugs in seven UK sites: real world results and modelling of treatment impact', *Journal of Viral Hepatitis* **22**, 399–408.

Martin, N. K., Vickerman, P., Grebely, J., Hellard, M., Hutchinson, S. J., Lima, V. D., Foster, G. R., Dillon, J. F., Goldberg, D. J., Dore, G. J. and Hickman, M. (2013), 'Hepatitis C virus treatment for prevention among people who inject drugs: Modeling treatment scale-up in the age of direct-acting antivirals', *Hepatology* **58**(5), 1598–1609.

Mauss, S., Berg, T., Rockstroh, J., Sarrazin, C. and Wedemeyer, H. (2018), *Hepatology: A clinical textbook*, 9th edn, Hannover.
**URL:** *www.hepatologytextbook.com*

McClelland, D. B. L. (2013), 'UK BTS - Handbook of Transfusion Medicine'.
**URL:** *https://www.transfusionguidelines.org/transfusion-handbook*

Meffre, C., Strat, Y. L., Delarocque-Astagneau, E., Dubois, F., Antona, D., Lemasson, J.-M., Warszawski, J., Steinmetz, J., Coste, D., Meyer, J.-F., Leiser, S., Giordanella, J.-P., Gueguen, R. and Desenclos, J.-C. (2010), 'Prevalence of hepatitis B and hepatitis C virus infections in France in 2004: social factors are important predictors after adjusting for known risk factors', *Journal of Medical Virology* **82**, 546–555.

Micallef, J. M., Kaldor, J. M. and Dore, G. J. (2006), 'Spontaneous viral clearance following acute hepatitis C infection: a systematic review of longitudinal studies', *Journal of Viral Hepatitis* **13**(1), 34–41.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and Edmunds, W. J. (2008), 'Social contacts and mixing patterns relevant to the spread of infectious diseases', *PLoS Medicine* **5**(3), e74.

Mravčík, V., Strada, L., Stolfa, J., Bencko, V., Groshkova, T., Reimer, J. and Schulte, B. (2013), 'Factors associated with uptake, adherence, and efficacy of hepatitis C treatment in people who inject drugs: a literature review', *Patient Preference and Adherence* **7**, 1067–1075.

Muench, H. (1934), 'Derivation of rates from summation data by the catalytic curve', *Journal of the American Statistical Association* **29**(185), 25–38.

Murphy, G. and Parry, J. V. (2008), 'Assays for the detection of recent infections with human immunodeficiency virus type 1', *Eurosurveillance* **13**(36), 4–10.

Nagelkerke, N., Heisterkamp, S., Borgdorff, M., Broekmans, J. and Van Houwelingen, H. (1999), 'Semi-parametric estimation of age-time specific infection incidence from serial prevalence data', *Statistics in Medicine* **18**(3), 307–320.

Namata, H., Shkedy, Z., Faes, C., Aerts, M., Molenberghs, G., Theeten, H., van Damme, P. and Beutels, P. (2007), 'Estimation of the force of infection from current status data using generalized linear mixed models', *Journal of Applied Statistics* **34**(8), 923–939.

National Institute for Health and Care Excellence (2015), 'NICE technology appraisal guidance 330: Sofosbuvir for treating chronic hepatitis C'.
**URL:** *https://www.nice.org.uk/guidance/ta330*

Nelder, J. A. and Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A* **135**(3), 370–384.

NHS Scotland (2005), 'Hepatitis C Proposed Action Plan in Scotland'.
**URL:** *https://www.gov.scot/Publications/2005/06/14134528/45311*

Nocedal, J., Wright, S. J. and Robinson, S. M. (2006), *Numerical Optimization*, 2nd edn, Springer-Verlag, Berlin, New York.

Noone, A., Durante, A. J., Brady, A. R., Majid, F., Swan, A. V., Parry, J. V., Hart, G. J., Connell, J. A., Perry, K. R. and Joce, R. E. (1993), 'HIV infection in injecting drug users attending centres in England and Wales, 1990-1991', *AIDS* **7**(11), 1501–1507.

Oakes, D. (1989), 'Bivariate Survival Models Induced by Frailties', *Journal of the American Statistical Association* **84**(406), 487–493.

Orkin, C., Leach, E., Flanagan, S., Wallis, E., Ruf, M., Foster, G. R. and Tong, C. Y. (2015), 'High prevalence of hepatitis C (HCV) in the emergency department (ED) of a London hospital: Should we be screening for HCV in ED attendees?', *Epidemiology and Infection* **143**(13), 2837–2840.

Paik, M. C., Tsai, W.-Y. and Ottman, R. (1994), 'Multivariate survival analysis using piecewise gamma frailty', *Biometrics* **50**(4), 975–988.

Palmateer, N. E., Taylor, A., Goldberg, D. J., Munro, A., Aitken, C., Shepherd, S. J., McAllister, G., Gunson, R. and Hutchinson, S. J. (2014), 'Rapid decline in HCV incidence among people who inject drugs associated with national scale-up in coverage of a combination of harm reduction interventions', *PLoS ONE* **9**(8).

Prevost, T. C., Presanis, A. M., Taylor, A., Goldberg, D. J., Hutchinson, S. J. and De Angelis, D. (2015), 'Estimating the number of people with hepatitis C virus who have ever injected drugs and have yet to be diagnosed: An evidence synthesis approach for Scotland', *Addiction* **110**(8), 1287–1300.

Public Health England (2014), 'People who inject drugs: HIV and viral hepatitis monitoring'.
**URL:** *https://www.gov.uk/government/publications/people-who-inject-drugs-hiv-and-viral-hepatitis-monitoring*

Public Health England (2015), 'Hepatitis C in the UK 2015 report'.
**URL:** *https://www.gov.uk/government/uploads/system/uploads/ attachment_data/file/448710/NEW_FINAL_HCV_2015_IN_THE_UK_REPORT_28072015_v2.pdf*

Public Health England (2016), 'HIV in the UK: 2016 report'.
**URL:** *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/602942/HIV_in_the_UK_report.pdf*

Public Health England (2017), 'Acute hepatitis B (England): annual report for 2016'.
**URL:** *https://www.gov.uk/government/uploads/system/uploads/ attachment_data/file/643558/hpr3117_hepB_ann.pdf*

Public Health England (2018*a*), 'Hepatitis C in England 2018 report'.
**URL:** *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/693917/HCV_in_England_2018.pdf*

Public Health England (2018*b*), 'Hepatitis C: tools for estimating disease burden'.
**URL:** *https://www.gov.uk/government/publications/hepatitis-c-commissioning-template-for-estimating-disease-prevalence*

Public Health England (2018*c*), 'Quality at the Laboratories of Public Health England, Colindale'.
**URL:** *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/682224/QW0080.23_laboratory_quality_at_PHE.pdf*

Public Health England, Health Protection Scotland, Public Health Wales and Public Health Agency Northern Ireland (2014), 'Shooting up: Infections among people who inject drugs in the United Kingdom 2013'.
**URL:** *https://www.gov.uk/government/uploads/system/uploads/ attachment_data/file/370707/Shooting_Up_2014.pdf*

Reuter, P. and Stevens, A. (2008), 'Assessing UK drug policy from a crime control perspective', *Criminology and Criminal Justice* **8**(4), 461–482.

Rhodes, T., Davis, M. and Judd, A. (2004), 'Hepatitis C and its risk management among drug injectors in London: renewing harm reduction in the context of uncertainty', *Addiction* **99**(5), 621–633.

Rolls, D. A., Daraganova, G., Sacks-Davis, R., Hellard, M., Jenkinson, R., McBryde, E., Pattison, P. E. and Robins, G. L. (2012), 'Modelling hepatitis C transmission over a social network of injecting drug users', *Journal of Theoretical Biology* **297**, 73–87.

Rosenberg, N. E., Pilcher, C. D., Busch, M. P., Cohen, M. S., Hill, C. and Francisco, S. (2016), 'How can we better identify early HIV infections?', *Current Opinion in HIV and AIDS* **10**(1), 61–68.

Royston, P. (2000), 'A strategy for modelling the effect of a continuous covariate in medicine and epidemiology', *Statistics in Medicine* **19**(14), 1831–1847.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Schenzle, D., Dietz, K. and Frösner, G. (1979), 'Antibody against hepatitis A in seven European countries. II. Statistical analysis of cross-sectional surveys', *American Journal of Epidemiology* **110**(1), 70–76.

Shepard, C. W., Simard, E. P., Finelli, L., Fiore, A. E. and Bell, B. P. (2006), 'Hepatitis B virus infection: Epidemiology and vaccination', *Epidemiologic Reviews* **28**(1), 112–125.

Shkedy, Z., Aerts, M. and Molenberghs, G. (2013), 'Modelling forces of infection by using monotone local polynomials', *Journal of the Royal Statistical Society. Series C* **52**(4), 469–485.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2006), 'Modelling age-dependent force of infection from prevalence data using fractional polynomials', *Statistics in Medicine* **25**(9), 1577–1591.

Singal, A. G., Volk, M. L., Jensen, D., Di Bisceglie, A. M. and Schoenfeld, P. S. (2010), 'A sustained viral response is associated with reduced liver-related morbidity and mortality in patients with hepatitis C virus.', *Clinical Gastroenterology and Hepatology* **8**(3), 280–288.e1.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. and Carpenter, J. R. (2009), 'Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls', *BMJ* **338**(July), b2393.

Stewart, D., Gossop, M., Marsden, J. and Rolfe, A. (2000), 'Drug misuse and acquisitive crime among clients recruited to the National Treatment Outcome Research Study (NTORS)', *Criminal Behaviour and Mental Health* **10**(1), 10–20.

Sutton, A. J., Gay, N. J., Edmunds, W. J., Andrews, N. J., Hope, V. D. and Gill, O. N. (2005), 'Modelling the characteristics of the male injecting drug user population in England and Wales', *International Journal of Drug Policy* **16**(3), 176–182.

Sutton, A. J., Gay, N. J., Edmunds, W. J., Hope, V. D., Gill, O. N. and Hickman, M. (2006), 'Modelling the force of infection for hepatitis B and hepatitis C in injecting drug users in England and Wales.', *BMC Infectious Diseases* **6**(93).

Sutton, A. J., Hope, V. D., Mathei, C., Mravcik, V., Sebakova, H., Vallejo, F., Suligoi, B., Brugal, M. T., Ncube, F., Wiessing, L. and Kretzschmar, M. (2008), 'A comparison between the force of infection estimates for blood-borne viruses in injecting drug user populations across the European Union: A modelling study', *Journal of Viral Hepatitis* **15**(11), 809–816.

Swed, F. S. and Eisenhart, C. (1943), 'Tables for testing randomness of grouping in a sequence of alternatives', *The Annals of Mathematical Statistics* **14**(1), 66–87.

Sweeting, M., De Angelis, D., Hickman, M. and Ades, A. E. (2008), 'Estimating hepatitis C prevalence in England and Wales by synthesizing evidence from multiple data sources. Assessing data conflict and model fit', *Biostatistics* **9**(4), 715–734.

Sweeting, M. J., De Angelis, D. and Aalen, O. O. (2005), 'Bayesian back-calculation using a multi-state model with application to HIV', *Statistics in Medicine* **24**(24), 3991–4007.

Sweeting, M. J., De Angelis, D., Ades, A. and Hickman, M. (2009a), 'Estimating the prevalence of ex-injecting drug use in the population', *Statistical Methods in Medical Research* **18**(4), 381–395.

Sweeting, M. J., De Angelis, D., Brant, L. J., Harris, H. E., Mann, A. G. and Ramsay, M. E. (2007), 'The burden of hepatitis C in England.', *Journal of Viral Hepatitis* **14**(8), 570–576.

Sweeting, M. J., De Angelis, D., Neal, K. R., Ramsay, M. E., Irving, W. L., Wright, M., Brant, L. and Harris, H. E. (2006), 'Estimated progression rates in three United Kingdom hepatitis C cohorts differed according to method of recruitment', *Journal of Clinical Epidemiology* **59**(2), 144–152.

Sweeting, M. J., Hope, V. D., Hickman, M., Parry, J. V., Ncube, F., Ramsay, M. E. and De Angelis, D. (2009b), 'Hepatitis C infection among injecting drug users in England and Wales (1992-2006): there and back again?', *American Journal of Epidemiology* **170**(3), 352–360.

Taylor, A., Goldberg, D., Hutchinson, S., Cameron, S., Gore, S. M., McMenamin, J., Green, S., Pithie, A. and Fox, R. (2000), 'Prevalence of hepatitis C virus infection among injecting drug users in Glasgow 1990-1996: Are current harm reduction strategies working?', *Journal of Infection* **40**(2), 176–183.

The Antiretroviral Therapy Cohort Collaboration (2009), 'Variable impact on mortality of AIDS-defining events diagnosed during combination antiretroviral therapy: not all AIDS defining conditions are created equal', *Clinical Infectious Diseases* **48**(8), 1138–1151.

The Antiretroviral Therapy Cohort Collaboration (2017), 'Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies', *The Lancet HIV* **4**, e349–e356.

The Health Protection Agency (2009), 'Hepatitis C in the UK 2009'.
**URL:** *http://webarchive.nationalarchives.gov.uk/20140712181932/ http://www.hpa.org.uk/Publications/InfectiousDiseases/ BloodBorneInfections/HepatitisCInTheUK/0912HepatitisC/*

The Health Protection Agency (2012*a*), 'HIV in the United Kingdom : 2012 Report'.
**URL:** *http://webarchive.nationalarchives.gov.uk/20140711233213/ http://www.hpa.org.uk/webw/HPAweb&Page&HPAwebAutoListNameDesc/ Page/1249920575999*

The Health Protection Agency (2012*b*), 'Shooting Up: Infections among people who inject drugs in the United Kingdom 2011'.
**URL:** *http://webarchive.nationalarchives.gov.uk/20140711235006/ http://www.hpa.org.uk/Publications/InfectiousDiseases/ BloodBorneInfections/ShootingUp/1211Shootingup2012/*

The Polaris Observatory Collaborators (2017), 'Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study', *The Lancet Gastroenterology & Hepatology* **2**(3), 161–176.

The Polaris Observatory Collaborators (2018), 'Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: a modelling study', *The Lancet Gastroenterology & Hepatology* **1253**(18), 1–21.

Thomson, B. J., Kwong, G., Ratib, S., Sweeting, M., Ryder, S. D., De Angelis, D., Grieve, R. and Irving, W. L. (2008), 'Response rates to combination therapy for chronic HCV infection in a clinical setting and derivation of probability tables for individual patient management', *Journal of Viral Hepatitis* **15**(4), 271–278.

Thuluvath, P. J., Krok, K. L., Segev, D. L. and Yoo., H. Y. (2007), 'Trends in post-liver transplant survival in patients with hepatitis C between 1991 and 2001 in the United States', *Liver Transplantation* **13**, 719–724.

Turner, K. M. E., Hutchinson, S., Vickerman, P., Hope, V., Craine, N., Palmateer, N., May, M., Taylor, A., De Angelis, D., Cameron, S., Parry, J., Lyons, M., Goldberg, D., Allen, E. and Hickman, M. (2011), 'The impact of needle and syringe provision and opiate substitution therapy on the incidence of hepatitis C virus in injecting drug users: pooling of UK evidence', *Addiction* **106**(11), 1978–1988.

UNAIDS (2016), 'Global AIDS Update 2016'.
**URL:** *http://www.unaids.org/sites/default/files/media_asset/ UNAIDS_Gap_report_en.pdf*

United Nations (2015), 'Transforming our world: the 2030 Agenda for Sustainable Development'.
**URL:** *https://sustainabledevelopment.un.org/post2015/ transformingourworld*

Unkel, S. and Farrington, C. P. (2012), 'A new measure of time-varying association for shared frailty models with bivariate current status data.', *Biostatistics* **13**(4), 665–679.

van de Laar, T. J. W., van der Bij, A. K., Prins, M., Bruisten, S. M., Brinkman, K., Ruys, T. A., van der Meer, J. T. M., de Vries, H. J. C., Mulder, J.-W., van Agtmael, M., Jurriaans, S., Wolthers, K. C. and Coutinho, R. A. (2007), 'Increase in HCV incidence among men who have sex with men in Amsterdam most likely caused by sexual transmission', *The Journal of Infectious Diseases* **196**(2), 230–238.

van den Berg, C. H. S. B., Smit, C., Bakker, M., Geskus, R. B., Berkhout, B., Jurriaans, S., Coutinho, R. A., Wolthers, K. C. and Prins, M. (2007), 'Major decline of hepatitis C virus incidence rate over two decades in a cohort of drug users', *European Journal of Epidemiology* **22**(3), 183–193.

Vickerman, P., Martin, N., Turner, K. and Hickman, M. (2012), 'Can needle and syringe programmes and opiate substitution therapy achieve substantial reductions in hepatitis C virus prevalence? Model projections for different epidemic settings', *Addiction* **107**(11), 1984–1995.

Wald, A. and Wolfowitz, J. (1940), 'On a test whether two samples are from the same source', *The Annals of Mathematical Statistics* **11**(2), 147–162.

WHO (2012*a*), 'WHO Hepatitis B Fact Sheet No. 204'.
**URL:** *http://www.who.int/mediacentre/factsheets/fs204/en/index.html*

WHO (2012*b*), 'WHO Hepatitis C Fact Sheet No. 164'.
**URL:** *http://www.who.int/mediacentre/factsheets/fs164/en/index.html*

WHO (2012*c*), 'WHO HIV/AIDS Fact Sheet No. 360'.
**URL:** *http://www.who.int/mediacentre/factsheets/fs360/en/index.html*

Wienke, A., Arbeev, K. G., Locatelli, I. and Yashin, A. I. (2005), 'A comparison of different bivariate correlated frailty models and estimation strategies', *Mathematical Biosciences* **198**(1), 1–13.

Wood, S. N. (2003), 'Thin plate regression splines', *Journal of the Royal Statistical Society: Series B* **65**(1), 95–114.

World Health Organization (2016), 'Global health sector strategy on viral hepatitis 2016-2021.'.
**URL:** *http://www.who.int/hepatitis/strategy2016-2021/ghss-hep/en/*

World Health Organization (2017), 'Global hepatitis report, 2017'.
**URL:** *http://www.who.int/hepatitis/publications/global-hepatitis-report2017/en/*

Xia, Y., Seaman, S., Hickman, M., Macleod, J., Robertson, R., Copeland, L., McKenzie, J. and De Angelis, D. (2015), 'Factors affecting repeated cessations of injecting drug use and relapses during the entire injecting career among the Edinburgh Addiction Cohort', *Drug and Alcohol Dependence* **151**, 76–83.

Yashin, A. I., Vaupel, J. W. and Iachine, I. A. (1995), 'Correlated individual frailty: an advantageous approach to survival analysis of bivariate data', *Mathematical Population Studies* **5**(2), 145–159.