# Statistical Applications in the Analysis of Vaccine Preventable Diseases

Alan Yeung

Department of Mathematics and Statistics

University of Strathclyde

Glasgow, UK

2016

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science.

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree

**Signed:**

S. Yeung

**Date:** July 24, 2016

# Acknowledgements

I have found the PhD to be a difficult but rewarding experience and so it is necessary to fully express my gratitude to those that have helped me to get through it. First and foremost, I would like to give a huge thanks to my supervisor, Professor Chris Robertson, for the invaluable support and guidance offered to me throughout my PhD. I feel that I have come a long way since I started and have tried to get through many of the challenges on my own, but there is no doubt that Chris has been pivotal in getting me to where I am today.

I would also like to thank the University of Strathclyde for financial support, facilities and providing a warm and friendly atmosphere for me to undertake my research. Finally a mention goes out to my fellow postgraduate students in the maths and stats department, my family and any staff that helped me along the way for ensuring my enthusiasm remained high. I believe that you should try to enjoy yourself at your place of work and these people have certainly done their part in that respect.

# Abstract

Disease outbreaks are a constant threat to public health and so effective management of these outbreaks is vital. By using statistical methods, we can better understand how a disease is affecting populations and monitor the progression of diseases over time. This thesis applies and develops statistical methods to studies of vaccine-preventable disease outbreaks in Scotland and aims to aid in the detection and management of outbreaks.

For detecting outbreaks, a system was designed for Health Protection Scotland (HPS) to link cases with incomplete genetic typing data to other cases to form potential clusters that may be worthy of further investigation. A novel influenza strain spread worldwide in 2009 and this work helps in the understanding and monitoring of that outbreak. A key parameter is the reproductive number and this was monitored for pandemic influenza using routinely collected data. Then postcode data was added to develop a spatial model for estimating the rate of spread. For vaccine-preventable diseases, the primary intervention strategy is vaccination but their effectiveness must be assessed. Vaccine effectiveness (VE) was assessed against various clinical outcomes which were associated with influenza to differing extents. Additionally, different methods were employed, including attempts to correct for biases.

The main findings of this work have important implications. The system to identify linked TB cases helps to ensure more links between cases are found, preventing further disease spread. The spatial method for estimating reproductive numbers offers improved parameter estimates. The VE study found that estimates differ more by the outcomes it was measured against than by the methods employed. Moreover, estimates found using outcomes with low specificity for influenza can be unreliable. Therefore, the recommendation for future studies is to focus on using outcomes with higher specificity for influenza.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Modern society is certainly faced with no shortage of arduous problems which must be tackled diligently. These problems are diverse and include those relating to politics, economics, the environment, social issues and health issues. Ultimately, these issues all intermingle with one another to some extent but in the field of epidemiology, we are primarily interested in health issues. With regards to these, diseases pose one of the most serious problems facing us today. Diseases can be broadly categorised into two forms – *chronic degenerative diseases* such as cancer and diabetes, and *infectious diseases* such as influenza and malaria. Both of these are of massive concern to public health bodies around the world but our main focus here is on infectious diseases and more precisely, on a subgroup of these which we can entitle as *vaccine preventable diseases.* These are perhaps, the group of infectious diseases that we are best equipped to deal with and we will discuss them in more detail later in this introduction.

In the past, infectious disease was undoubtedly the number one health issue. In more recent times, developed countries have redirected a lot of their attention towards chronic degenerative diseases [385]. In particular, chronic diseases which are currently being given a great deal of attention include cancer [178, 230], obesity [376], diabetes [7] and oral health problems [194, 231] which are increasingly being recognised as having associations with numerous other health problems including diabetes and chronic heart disease [29]. One of the primary reasons for the increased attention being paid to chronic disease is that we have become much better at dealing with infectious disease outbreaks as many of them can now be controlled with vaccines and some can be effectively treated with antibiotics [375].

However, infectious diseases remain as the main focus in developing countries and despite the increased focus on chronic degenerative diseases in current times, the world is always wary of infectious diseases as they can have *explosive* characteristics and can also be very unpredictable.

There are plenty of classical examples to illustrate the devastation caused on populations by infectious diseases throughout history. Some of the more documented ones include cholera [105], plagues [87, 400] and influenza [192]. A more comprehensive list of infectious disease threats arising from 2000–2011 can be found in a list created by the Centers for Disease Control and Prevention (CDC) [61]. Each of the diseases we just mentioned has killed millions of people over a relatively short space of time and they continue to lurk in the background to varying extents. Although our main focus is on diseases that directly affect humans, it would be inappropriate to not mention here that infectious diseases that spread in animals are also hugely important public health issues. Disease outbreaks in animal populations can have detrimental impacts on the economy and some of those diseases can also spread from animals to humans. Some of the largest infectious disease threats to animals in Europe have been suggested to come from influenza and foot-and-mouth disease [201].

Advances in medicine such as antibiotics and vaccines have given us plenty of tools to fight back against diseases. However, new issues continually arise such as antibiotic resistance [380] and the fight is a constant back and forth between human endeavour versus nature's extraordinary evolutionary capability. For these reasons, infectious diseases remain perfectly capable of causing substantial fear and panic at any moment as evidenced fairly recently by the likes of severe acute respiratory syndrome (SARS) in 2002 [344], novel forms of influenza such as pandemic influenza in 2009 [63], new coronaviruses like middle east respiratory syndrome (MERS) in 2012 [78], and Ebola virus disease (EVD) in 2014 [125].

The use of statistics in epidemiological investigations goes back a long time in history. Some noted early examples come from Florence Nightingale's employment of statistics to help in understanding that a major cause of death for British soldiers in the 19th century came from unsanitary conditions in military hospitals [210], and also John Snow's investigation to find the source of the 1854 cholera outbreak in the Soho neighbourhood of London (now well known as the Broad Street pump outbreak) [279]. It is now widely recognised that understanding and employing mathematical and statistical methods is crucial for assessing epidemiological data in order to make decisions regarding public health [246, 347]. Since those early days, a vast array of statistical methods relevant for epidemiologists have been developed and a good overview of some of the most commonly utilised methods has been compiled by Jewell [189].

In this thesis, we apply and develop statistical methods and models in order to study vaccine-preventable disease outbreaks in Scotland. The goal is to assist in the management of the disease outbreaks considered in this thesis as well as aiding in the management of outbreaks on a more general level. Discussions on infectious diseases can contain a great deal of domain-specific jargon and a basic understanding of some of the main aspects of infectious diseases can be very beneficial for the reader. Hence, here we give a short background to infectious disease characteristics, provide some definitions to the basic terms used in discussions of infectious diseases, and briefly describe some of the most relevant steps involved in managing infectious diseases, paying particular attention to the role of the statistician within this process.

## 1.1 Definitions

Various terms related to infectious diseases will be used throughout this thesis and we define and explain some of them here. The list is non-exhaustive but larger glossaries are available in for instance, Last et al. [212] or Barreto et al. [20], and more information about infectious diseases in general can be found in Nelson and Williams [260].

### Infectious Agent

An infectious agent is any micro or macro organism that has the ability to produce infection in a human or animal. This includes bacteria, viruses, fungi and parasites. Not all infectious agents entering into a body will actually develop into disease.

### Infectious Disease

The exact definition of an infectious disease has been debated throughout history [104], particularly as technological advances in our ability to detect different types of diseases have been made. One of the earliest definitions of infectious disease was coined by Jakob Henle and Robert Koch in the nineteenth century. The requirements for an infectious disease to exist under their causal criteria have been named as the *Henle-Koch Postulates*. Although these are still used by some researchers today [121], the general consensus seems to be that they are sufficient but not strictly necessary for establishing causation [103]. Hence, we shall use the definition as given in the Dictionary of Epidemiology [212] rather than getting too concerned with the intricacies of causality criteria.

They state that an infectious disease is an illness caused by an infectious agent or the toxic product of an agent that results from transmission of that agent (or its toxic product) from an infected person, animal or reservoir (for example, plants, soil or substances) to a susceptible host, either directly or indirectly through an intermediate animal host, vector, plant or inanimate environment.

### Vaccine Preventable Disease

When an effective and preventative vaccine exists for an infectious disease, then we can refer to that disease as a vaccine preventable disease. Coincidentally, any deaths resulting from a disease for which a vaccine exists is termed a *vaccine preventable death*. Effective vaccines can induce immunity from an infectious disease and they provide a way of potentially eradicating infectious diseases. Thus, these diseases should be of significant interest to public health bodies as vaccines have the power to save many lives through preventative action rather than corrective action.

As an example of how successful vaccines have been, Roush et al. [316] gives a comparison of the numbers of cases and deaths for a number of vaccine preventable diseases in the US in the pre-vaccine era against the post-vaccine era. In summary, it shows that vaccines have been instrumental in decreasing numbers of cases and deaths from many diseases by over 90%. There have even been exactly 100% or very close to 100% decreases in the US for some diseases such as diphtheria, polio and smallpox. However, there are plenty of issues to consider with regards to the use of vaccines – they vary in their efficacy in preventing disease; they may only induce immunity for some limited period of time; some vaccines induce harmful side effects upon certain groups of individuals; some diseases evolve over time; vaccines can be costly to manufacture on a global scale; and there can be logistical issues in administering them to people, particularly in politically unstable regions.

Some vaccines against human diseases have been around for a long time as shown in Table 1.1 but at present, only two infectious diseases have been classified as being eradicated worldwide as a result of human intervention and only one of these is a human disease. Those two diseases are *smallpox* (a human disease) and *rinderpest* (a disease affecting cattle) [247]. The smallpox vaccine introduced by Edward Jenner in 1798 was groundbreaking in that it was the first successful vaccine ever developed, but it took serious concerted effort to eradicate smallpox globally and this was finally achieved in 1979 [24, 42]. Today, the elimination of smallpox has saved countless millions of lives and it is considered as one of the greatest feats of humanity. However, the fact that only two diseases have ever been eradicated despite so many effective vaccines being available clearly demonstrates it is still a massive challenge to completely eliminate any infectious disease.

### Emerging Infectious Disease

A number of the vaccine preventable diseases we look at in this thesis can also be described as emerging infectious diseases (EIDs). The exact definition of an EID may be fairly subjective but one definition is an infectious disease whose incidence

| | | | |
|---|---|---|---|
| Smallpox | 1798 | Mumps | 1967 |
| Rabies | 1885 | Rubella | 1969 |
| Typhoid | 1896 | Anthrax | 1970 |
| Cholera | 1896 | Meningitis | 1975 |
| Plague | 1897 | Streptococcus pneumoniae | 1977 |
| Diphtheria | 1923 | Adenovirus | 1980 |
| Pertussis | 1926 | Hepatitis B | 1981 |
| Tetanus | 1927 | *Hemophilus influenzae* type b | 1985 |
| Tuberculosis | 1927 | Japanese encephalitis | 1992 |
| Influenza | 1945 | Hepatitis A | 1995 |
| Yellow fever | 1953 | Varicella | 1995 |
| Poliomyelitis | 1955 | Lyme disease | 1998 |
| Measles | 1963 | Rotavirus | 1998 |

**Table 1.1:** Vaccine preventable human diseases alongside the years the vaccines were developed. The source of this information is Nelson and Williams [260].

has been increasing in the past two decades and is likely to increase in the future [214]. An attempt has been made to add quantitative rigour to the definition of an EID by Funk et al. [126] but as yet, a clear consensus still does not exist. Some of the main reasons for the appearance of EIDs include genetic drift in disease strains (for example, in influenza strains), alterations in climate, changes in human susceptibility, and antibiotic resistance. Something which should be mentioned alongside EIDs are *reemerging infections* such as drug resistant tuberculosis, which has reemerged in recent times due to antibiotic resistance.

### Herd Immunity

Herd immunity is a term which may convey a number of meanings in the context of epidemiology [117]. One use for the term is to describe the proportion of people in a population that have become immune (either naturally or through the use of vaccines) to a particular infectious disease. However, the term may also be used to refer to some threshold proportion of a population which has to become immune in order to lead to a decline in the incidence of infection. Although there may be different specific meanings for the term, the general implication is always similar – the overall risk of infection among susceptible individuals in a population is greatly reduced if a sufficient proportion of the population is immune. This is because infectious individuals are much less likely to come into contact with susceptible individuals (see Figure 1.1 for a depiction of this concept). The protection conferred to susceptible individuals in this scenario is sometimes also known as *indirect protection*.

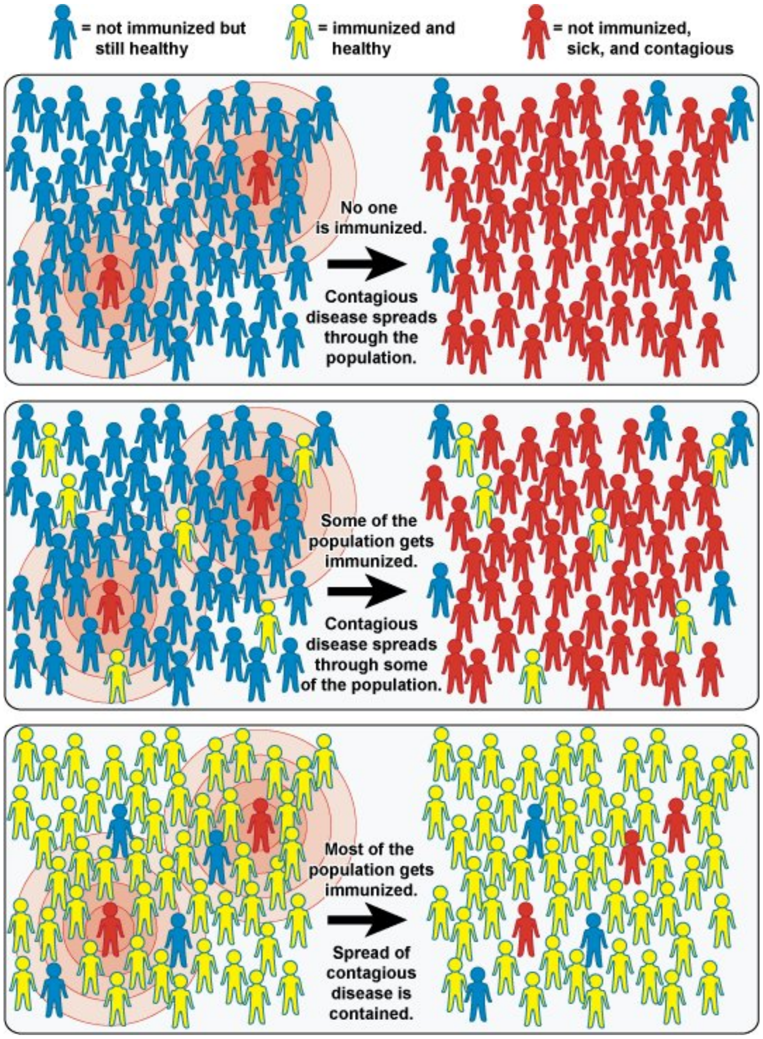

= not immunized but still healthy
= immunized and healthy
= not immunized, sick, and contagious
No one is immunized.
Contagious disease spreads through the population.
Some of the population gets immunized.
Contagious disease spreads through some of the population.
Most of the population gets immunized.
Spread of contagious disease is contained.

**Figure 1.1:** Depiction of the herd immunity concept. The source of the image is the National Institute of Allergy and Infectious Disease (NIAID) [254].

### Endemic

A disease which continues to occur at some expected frequency in a particular geographical location over a certain period of time is known as being endemic. In other words, the rate of infection is in a *steady state* or is consistently maintained in a population and therefore, the rate is not noticeably increasing or decreasing over time. As an example, malaria is endemic to parts of Africa [259] and varicella (chickenpox) remains endemic in a number of European countries [229].

### Epidemic

When a disease occurs at a greater rate than expected in a particular geographical location, then we have an epidemic. Some examples of well-known and relatively recent epidemics include SARS which occurred in Asia and Ebola in Africa. We can also say that there is an AIDS (acquired immunodeficiency syndrome) epidemic in Sub-Saharan Africa where the prevalence of the disease in adults is still much higher than in other parts of the world [360].

### Pandemic

If an epidemic has spread to different countries and a large population, then it becomes a pandemic. In 2009, we had a pandemic strain of influenza as it rapidly spread from Mexico to the US and then all over the world [63]. Novel forms of influenza have great potential for causing pandemics as most populations will be susceptible to them and the disease is easily transmitted. This is especially true in the modern world due to increasing population sizes, urbanisation and better transportation systems that allow individuals to readily move to other countries.

### Outbreak

An outbreak is the term given to an epidemic that is contained within a small population or small geographical area. Public Health England (PHE) defines it in broad terms as an incident in which two or more people experience similar illness and can be linked in time or location [297]. Of course, the usual expected background rate of a disease must always be taken into consideration with respect to the previous statement. An outbreak can also be an event that involves microbial or chemical contamination of food or water. In 2012, there was an outbreak of Legionnaires' disease in Edinburgh, Scotland [238]. Once the source of the infection was isolated, this disease could be confined to a small area as it is not contagious (cannot be spread from person to person) and is spread through breathing in droplets of contaminated water.

## 1.2 Disease Transmission

A number of different factors play vital roles in disease transmission. Firstly, there is the infectivity or virulence of the disease itself. Some diseases are naturally more potent and aggressive than others and these will have a greater ability to evade a host's defenses. Then there is also the susceptibility of the host which is affected by, for instance, an individuals genetics as well as their current state of health. Finally, the environment plays a role as well and this can include factors such as the climate, sanitation and access to health care. A particularly worrying scenario plays out when a highly virulent disease agent has been introduced into a susceptible population and there is an environment that brings these two together. For instance, Ebola was able to spread rapidly throughout Africa as it is a highly virulent disease and some countries there also do not have the adequate medical, communications and transport infrastructure in place to quickly respond to the outbreak [384]. There are numerous ways that a disease can be spread, some of which involve direct physical contact and some which are transmitted in a more indirect manner.

### Physical Contact

The most obvious route of infection is through direct physical contact. Exposure to the skin, body secretions or infected wounds of an individual or animal carrying a disease can cause transmission to another susceptible individual. This can occur in everything from shaking hands to sexual contact. Examples of diseases that can be transmitted in this way include the common cold and sexually transmitted diseases such as HIV/AIDS.

### Droplet

Diseases can be transmitted through droplets in the air, usually coming from respiratory secretions such as a person coughing, sneezing, talking or even simply breathing. Influenza – one of the main diseases we consider in this thesis – is transmitted in this way. Other examples of diseases spread by droplets are measles, mumps and rubella (MMR). There tends to be a higher incidence of diseases like influenza in the winter months and one reason for this may be that the droplets can survive better when there are lower indoor humidity levels [209].

### Vehicle

Inanimate objects or substances can host infectious agents and in doing so, is converted into a vehicle that can disseminate the disease. Common sources of diseases transmitted in this way include water, air, food and soil. For diseases transmitted in this fashion, it is critical to find out the source of disease transmission as quickly as possible to prevent further spread. For instance, if it is a food-borne outbreak, then the restaurants or food suppliers providing that food should be isolated. This exact situation arose in 2011 when an outbreak of E. Coli was linked firstly, to several restaurants and then later to the source which was a German bean-sprout farm [46]. Related to vehicular transmission is *iatrogenic transmission*. This is the specific case of where the transmission is related to medical equipment or medical procedures. Methicillin-resistant Staphylococcus aureus (MRSA) is one such disease that fits this category as it is usually acquired in hospital.

### Vector-borne

When an insect (or more generally, an arthropod) carries a disease which can be transmitted to humans or animals, it is called a vector-borne disease. One of the most widely known vector-borne diseases is malaria which is carried by mosquitoes. Almost half the world's population is at risk of this disease but the worst affected are those in the poorest countries [392]. Annually, there are hundreds of millions of cases of malaria and hundreds of thousands die from the disease. As well as malaria, mosquitoes also carry many other viral diseases including yellow fever, dengue fever and Japanese encephalitis. This has led to the term *mosquito-borne disease* which is sometimes used to refer to the group of diseases carried by mosquitoes.

#### Vertical

Vertical transmission occurs when a disease is passed down from parent to child. Examples of diseases which can be transmitted in this way include rubella [82], toxoplasmosis [169] and sexually transmitted diseases like HIV [220] and chlamydia [396]. Vertical transmission can be prevented by vaccinating prior to pregnancy (if a vaccine exists for the disease) or for diseases such as toxoplasmosis, antibiotics can be given to the mother if she is diagnosed early.

## 1.3 Managing Disease Outbreaks

There are a multitude of procedures involved in the management of communicable diseases which vary dependent upon many factors – severity of illness, transmissibility, and whether it affects humans, animals or both, to name just a few. A vague overview of the key stages involved can be seen in Figure 1.2 which is taken from the Communicable Disease Outbreak Plan by PHE [297]. It concisely summarises the entire process, right from the beginning, which is identification of disease outbreaks up to an official declaration of the end of an outbreak. Rough guides similar to this one for investigating outbreaks are commonly established by public health bodies around the world as it is essential for them to have some sort of protocol in place for investigators to follow when the need arises. This increases the chance of mounting a swift and effective response. For instance, the CDC in America also has a set of steps for outbreak investigations [62].

### 1.3.1 Identification

The first step in outbreak management is to establish that any suspected disease outbreak is actually a real outbreak. To come to this conclusion, certain pieces of information must be known in advance – namely, the expected number of cases in some geographical region over some defined period of time. If the observed

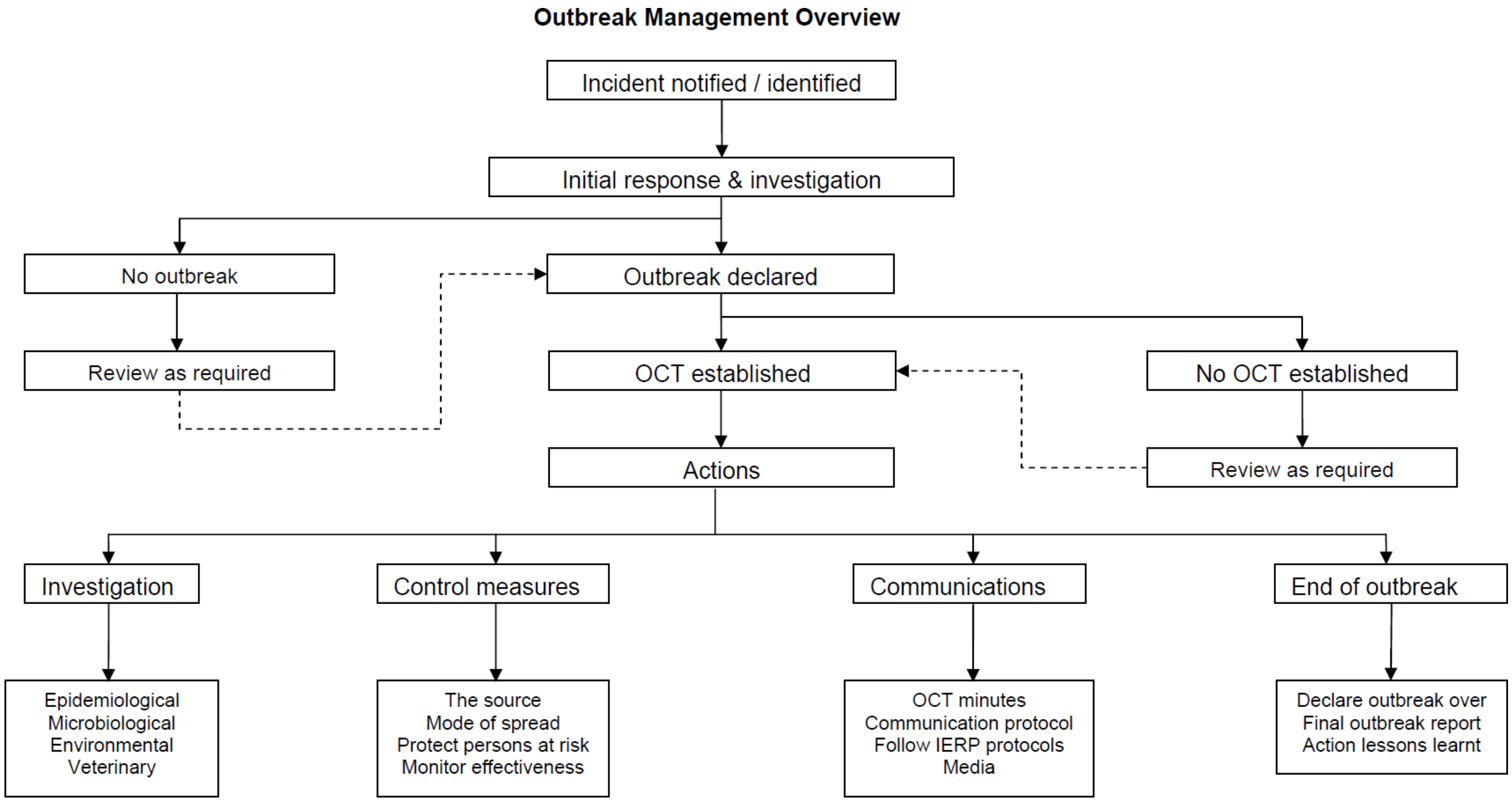


**Figure 1.2:** Flow diagram showing key stages of outbreak management in investigations by PHE as given in their outbreak plan [297]. Abbreviations – **OCT**: Outbreak Control Team; **IERP**: Incident and Emergency Response Plan.

number of cases exceeds this expected number, then this meets one of the fundamental criteria for an outbreak. Furthermore, if there is a connection between two or more people experiencing similar illness, either linked in time or location, then this strengthens the case for an actual outbreak. Links between cases can be found by speaking to the cases to see if they may have been in contact with each other recently or by using more sophisticated methods such as DNA analysis to find identical (or at least, highly similar) strains of a disease; in Chapter 2, we use genetic data to try to identify cases that may be part of the same transmission chain. Note that the conversations with cases is important as the link may not be from a person but could be from food, water, environment, animal or vector and these conversations could help in confirming a source.

The expected number of cases is usually found by looking at data on the numbers of cases over a comparable time period in the past. Often this kind of data can be found from looking at primary care data – for example, numbers of GP consultations for specific diseases. For diseases that appear extremely rarely in developed countries such as rabies and polio [322], a single case may be sufficient to meet the requirements for an outbreak investigation as this would be outside the norm. If it is a novel disease (or new strain of a disease), then it can be compared to numbers of cases from similar types of diseases in the past. However, since it is novel, it would naturally be of great interest and this would warrant an investigation, regardless of it's transmissibility and severity.

Another important point to take into consideration is that if the observed number of cases increases at a particular time, it may not necessarily be due to an outbreak, but could be due to a number of other factors. For instance, rises in reporting frequency or improvements in our ability to diagnose a disease would cause an increase in observed case numbers. As well as these, there could be an increased awareness about a particular disease that could result from media attention or publicity campaigns and this could have an impact on the number of case reports [127]. One more factor which could affect an increase in cases, but probably much more slowly over time, is changes in population in an area. If a population increases, perhaps due to increased migration, then this obviously increases the potential for more cases, at least in terms of the absolute numbers.

### 1.3.2 Epidemiological Investigation

To undertake an epidemiological investigation, it is first necessary to set definitions for classifying individuals as having the disease under consideration. The criteria assigned will usually be clear and objective but may vary in their ability to truly determine if the disease is present. For instance, one could be the presence of certain symptoms like fever or an increased level of antibody against a disease; the latter example here is more likely to be indicative of a true case as it is probably more directly attributable to a specific disease, while the former considers a symptom which appears in numerous illnesses. Since there may be some uncertainty here, cases are often categorised into *confirmed*, *probable* and *possible* cases. Laboratory-confirmed cases often constitute the confirmed cases; probable and possible cases will likely differ in the types of symptoms exhibited by individuals or may differ in the place or time where their illness took place.

Once case definitions have been set, detailed information is required from all cases. This includes demographic information such as age, gender, race and possibly geographic information as well, which is useful for finding out if the disease spread is confined to particular regions. Generally, demographic information is important for exploring whether or not certain groups of people are affected differently by a disease, or even if some groups are more likely to contract the disease. These differences can be found using simple statistical methods, beginning with tabulations, and then moving onto methods such as $t$-tests and $\chi^2$ tests to find differences and associations respectively. Adjusted estimates can be found using methods like linear regression, logistic regression or multinomial regression [108]. Additionally, it is usually desirable to have risk factor information on cases as those with, for example, chronic conditions are often more severely affected by diseases. It is crucial to have clinical information on cases. In particular, the date of onset of symptoms or illness and some idea of recent exposures are highly useful as is information on the severity of illness experienced by cases. Marrying the demographic, risk factor and clinical information together allows a fairly detailed epidemiological investigation of cases to be undertaken. The exploratory process of identifying patterns among cases of a disease and developing hypotheses based on any identified patterns is called *descriptive epidemiology*. Following from this, the testing of those hypotheses in a more formal statistical framework is called *analytical epidemiology*. We carry out these kind of analyses in Chapter 3.

### 1.3.3 Surveillance

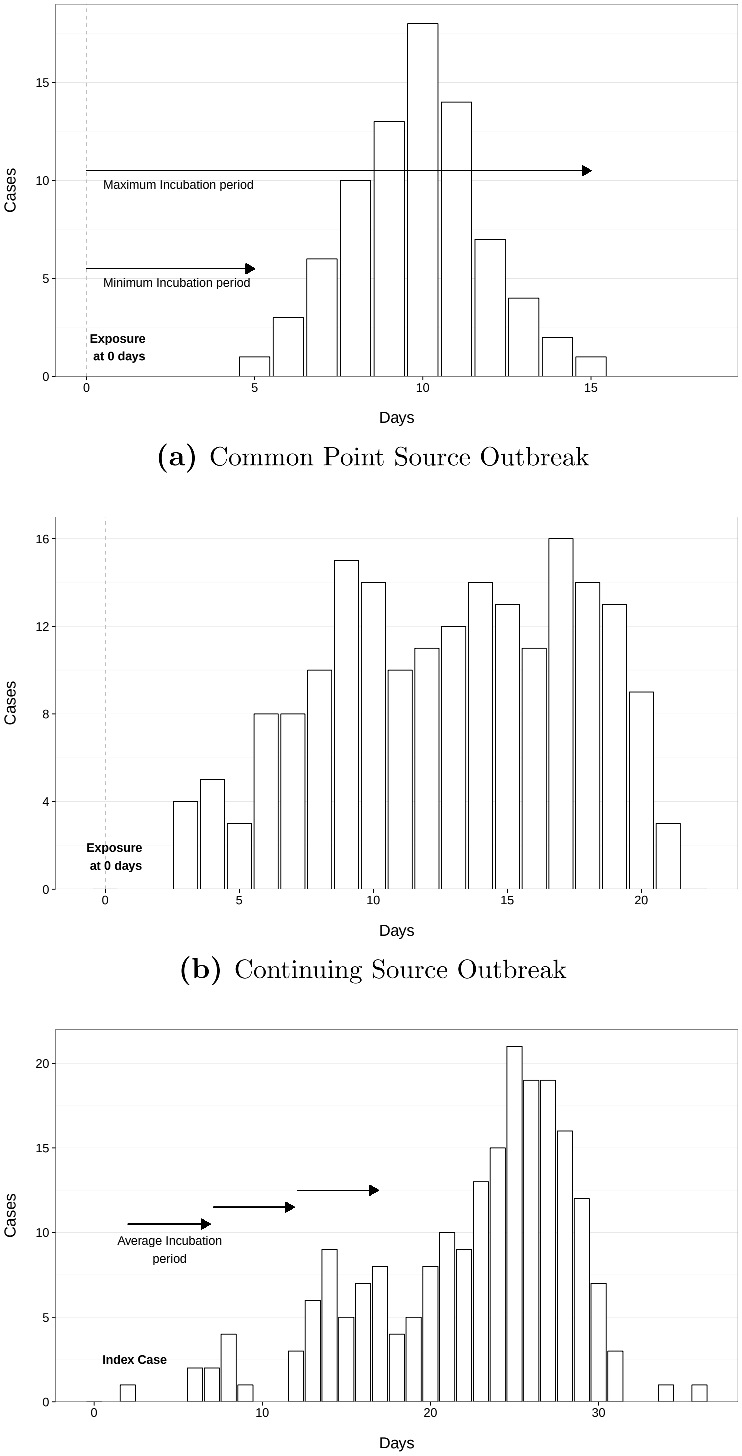
Disease outbreaks need to be continually monitored as they are unfolding and surveillance, in the context of pandemic influenza, is explored in Chapter 5. Surveillance is used to provide useful management information to help inform decisions on whether or not interventions are needed to bring an outbreak under control. A simple and crucial tool for surveillance is to produce the *epidemic curve* for a disease outbreak, which is a visual display of the number of cases over time. For instance, in Chapter 5, we use the epidemic curve as one of the initial steps on the surveillance of pandemic influenza in Scotland during 2009. To produce an epidemic curve, we require from the epidemiological investigation, the dates of onset of symptoms or something similar such as the dates that cases reported their illness. This can be updated as new data on case reports becomes available.

The epidemic curve can help in showing us a number of things. Most obviously, it illustrates the times at which the outbreak is growing or when the numbers of cases are decreasing, and we can also see the epidemic peak (or peaks if there are multiple peaks). If the peak is yet to occur, this should also be noticeable in the epidemic curve. Some examples of fairly typical epidemic curves created using fictitious data are shown in Figure 1.3.

The *shape* of an epidemic curve can give insights into the pattern of the epidemic such as whether or not it is likely that there has been person-to-person transmission of the disease or if it is more likely that a common source such as food or water has been responsible for the disease transmission. For example, if there is a sudden rapid rise in the number of cases followed by gentle decreases in the number of cases (see Figure 1.3a), then this is more indicative of a common source being responsible for the outbreak. This is because people are often exposed to a common source over a short period and then as awareness of illness increases, people will probably take protective measures. For person-to-person transmission, there are more likely to be a series of taller peaks over time until the point at which the outbreak begins to wane (see Figure 1.3c). This is due to the nature of person-to-person transmission, where we should be able to observe points in time where cases are *reproducing*. Note that taller peaks will only occur if each person, on average, transmits the infection to more than one other person. The eventual waning occurs as there will be less people who can be infected over time. Usually this is either due to people gaining immunity naturally after being infected or due to successful implementation of control measures.

The information given in the epidemic curve can also give us ideas about the *incubation period* (time between exposure to disease and symptoms first appearing) of a disease. If the outbreak occurs from a common source as in Figure 1.3a, then all of the cases will usually appear within the duration of one incubation period. However, if people are exposed to the source of disease for a prolonged period, then we will not observe a peak but it should look more like a plateau instead; this is the pattern shown in Figure 1.3b. In person-to-person transmission diseases, the peaks in the epidemic curve should initially be approximately one incubation period apart but as more cases arise, the peaks usually tend to merge into wave patterns (Figure 1.3c). If there are any outliers in an epidemic curve, they should be considered carefully. Those that appear early in the outbreak may point to a possible source of an outbreak or they may simply be an early case unrelated to the current outbreak. Similarly, late outliers may also be unrelated to the main outbreak or could be cases with unusually long incubation periods.

Another useful surveillance tool is to map the geographic locations of the cases, provided this data is available. For this we could use the locations of the cases' place of residence, where they work, or locations they commonly visit – in Chapter 5, we use postcode data related to where cases live to create a map. Producing



**(a)** Common Point Source Outbreak

**(b)** Continuing Source Outbreak

**(c)** Propagated Source Outbreak

**Figure 1.3:** Common examples of epidemic curves for different types of infectious disease outbreaks. Note that the data used to create the graphs is entirely fictitious.

a map like this can help us to identify the geographical extent of the outbreak as well as possible areas where people were exposed to the disease. As the outbreak progresses, the map should show particular areas where clusters of cases are appearing. One thing to always be aware of when mapping cases in this fashion is that more cases will nearly always appear in areas that are densely populated so sometimes it may be better to also consider the proportions of populations affected by a disease in particular regions. Conversely, if many cases appear in places that are not densely populated, then this immediately points to a likely hotbed for the spread of disease.

For a disease that transmits from person-to-person, we will also want a system for monitoring the rate that the disease is spreading over time. This is commonly characterised by the *reproductive number*, which tells us the average number of secondary cases that each primary case produces. There are a vast number of ways to calculate this parameter, depending on the disease characteristics as well as the available data and this is explained in much more detail in Chapter 4. Tools that monitor the reproductive number over time can help us to know when interventions should be put in place and once they have been put in place, they can be used again to assess the effectiveness of these control measures.

### 1.3.4 Control Measures

Control measures should always be implemented as early as possible in an outbreak. This should be the point where a sufficient amount is known regarding the characteristics of the disease and we can be fairly certain that particular actions will be effective in interrupting further spread. To achieve maximum effect, the control measures should be targeted at the source of the outbreak or those most likely to transmit infection [374]. Moreover, protective measures should be aimed at those most susceptible to the disease and most likely to suffer severe outcomes from the disease. With any control measure there are many considerations to be taken into account including practicality, cost-effectiveness of the strategy, side effects on health and impacts on the natural environment. As well as this, the characteristics of the disease play a central role in deciding which control measures are best to use.

There are many fairly routine ways of limiting a person's exposure to a disease and thus, decreasing their likelihood of contracting it. A typical measure which individuals can take which is usually highly beneficial is to practice good hygiene. An obvious action for this is more thorough and frequent hand-washing and efforts can be made by health services to promote this behaviour. Studies have already shown that improved hand hygiene can help in reducing some healthcare associated infections in hospitals [346]. Another measure to limit exposure include actions which isolate people from the general population. For instance, when people have a disease, they should not go to work and children should not go to school. In the case of children, it is particularly an issue as they tend to come into contact with greater numbers of people than adults [248]. If the situation becomes more severe, then schools can be closed for a period of time. However, when making this kind of decision, we must weigh the benefits of reduced disease transmission against disruption of normal day-to-day activities and services [58].

If the disease affects an animal population, there is an additional control strategy that has a similar effect to isolation but can, for obvious reasons, never be used in humans – culling. Of course, with a strategy like culling, there are many ethical considerations which must be taken into account [74] and it should only be used under extreme circumstances as a last resort. Examples of where a decision to cull animals has been made in the past include culling badgers to prevent the spread of tuberculosis in badgers and culling cattle to prevent the spread of the foot-and-mouth epidemic [115]. Since culling as a control measure is so drastic, it will inevitably attract a lot of public attention and there must be confidence that it will definitely be effective and there were no other eligible options. In the case of culling badgers, there have been contrasting opinions regarding the effect that culling badgers has had on preventing tuberculosis [92]. This kind of controversy emphasises the point that culling should only ever be be considered after all other possible control measures have been categorically ruled out.

For people that have already contracted the disease, there are still measures which can be taken to limit their transmissibility. To achieve this, sometimes medicines or antimicrobials can be used that either relieves a patient's symptoms or shortens the period that they can transmit the disease. This can ultimately have a large impact on the eventual size of the outbreak [150].

### Vaccination as a Control Measure

The control measure that has arguably greatest impact is vaccination as it reduces susceptibility by inducing long-term protective effects against a disease. However, it is often impractical and expensive to immunise entire populations against a disease. Hence, after some epidemiological investigation has been carried out to acquire a deeper understanding on a particular infectious disease, it is usually more appropriate to target specific groups of people who are more at risk of adverse outcomes in vaccination campaigns. For instance, in the UK, a targeted vaccination approach has been employed for the pneumococcal vaccine. Since 2003, PPV23 (pneumococcal polysaccharide vaccine containing polysaccharide from 23 capsular types) has been targeted at those aged 65 and over, and since 2010, PCV13 (pneumococcal conjugate vaccine containing polysaccharide from 13 common capsular types) has been aimed at children and is part of the routine childhood immunisation programme in the UK [294]. Note that before 2010, PCV7 was aimed at children instead of PCV13 and a further notable point is that children and adults that belong to a clinical risk group are also targeted for vaccination – for example, those with immunosuppression or chronic heart disease. Similarly, in Scotland, the influenza vaccine is targeted at the elderly, those with long-term health conditions, healthcare workers and also primary school children who get the vaccine via a nasal spray [298].

Another consideration with vaccines is that they are not always completely effective at preventing disease and their effectiveness can vary depending on the characteristics of individuals. There are numerous different methods for estimating the effectiveness of vaccines which are discussed more widely in Chapter 6. Commonly, the effect of vaccines are estimated from case-control and cohort studies. Regardless of the method used, the basic principle is always to compare the rate of disease among vaccinated individuals against the rate among unvaccinated individuals. An example of a vaccine with high efficacy is the measles vaccine [206] but other vaccines such as the pneumococcal vaccine (PPV23) and influenza vaccines have more moderate and varying levels of effectiveness [11, 116] (also see Chapter 7 for more on the effectiveness of influenza vaccines). In particular, influenza vaccines may not always be highly effective as the strains that circulate each season always vary.

Some vaccines may also have side-effects on the health of some groups of people. If the side-effects are mild such as a mild fever or a rash, it is generally still advisable to get vaccinated if at risk from a disease. However, some people would not be advised to get the vaccine such as those who have an allergy to certain ingredients used to synthesise the vaccine. At the most extreme, these people could suffer an anaphylactic reaction [97]. This problem may be counteracted to an extent when there are multiple vaccines available for the same disease. For instance, a variety of influenza vaccines are available for patients; most of the vaccines are inactivated and do not contain live viruses but some may contain viruses that have been attenuated. In addition to this, many influenza vaccines contain detectable amounts of egg protein so consideration must be given to individuals with egg allergies [86, 98, 359]. In the UK, those with egg allergies are recommended to get an inactivated influenza vaccine that is either egg-free or has a low amount of ovalbumin (the main protein found in egg white) [298].

### 1.3.5 Outbreak Conclusions

Every disease outbreak presents an opportunity to accrue additional knowledge on diseases and every opportunity must be fully seized as future outbreaks will inevitably occur. Hence, after the declaration that an outbreak is over, everything concerning the outbreak should be evaluated. For example, it is important to ascertain whether or not surveillance measures were good enough or if control measures were effective and implemented early enough. All of these findings have to be communicated to the relevant people and organisations to ensure we have learnt as much as possible from the outbreak and can mount more effective responses in future. Eventually the lessons learnt from the outbreak will appear in the scientific literature as a contribution to our overall knowledge on public health and epidemiology (see, for example, Leung and Nicoll [218] for a discussion on lessons learnt following 2009 pandemic influenza). The ultimate goal is to be able to react to these outbreaks in an optimal manner that will minimise the burden on public health.

## 1.4 Aims and Objectives

The general aims of the work in this thesis are to explore how statistical methods can be applied to assist in each of the main stages involved in managing vaccine-preventable infectious disease outbreaks – detecting outbreaks, descriptive and analytical epidemiology, surveillance of a disease outbreak, and evaluation of control measures. More specific aims include the following:

- Applying existing statistical methods to aid in the understanding of important epidemics that have affected Scotland in recent years.
- Creating novel systems that utilise data available from Scottish healthcare systems that can be used identify potential clusters or linked cases of a disease.
- Investigating how data routinely collected in Scotland can be used to extend current methods employed to assist with surveillance of infectious disease outbreaks.
- Exploring different methods of assessing the impact of control measures and providing recommendations on the best ways to conduct analyses to obtain reliable results.

## 1.5 Thesis Outline

This thesis essentially follows the structure just outlined for the management of infectious diseases, with the chapters looking at the individual parts involved.

Moreover, the thesis progresses following the stages involved in managing disease outbreaks and it has also been structured in a way such that large sections of it are relatively self-contained. For each of the analysis chapters, we had clear goals on what we wished to achieve – for instance, identifying disease outbreaks, surveillance of outbreaks, evaluation of intervention measures – but the decisions on exactly which methods and models to apply and develop to achieve these aims have been largely guided by the data that was made available by Health Protection Scotland (HPS).

In Chapter 2, we look at data on reported cases of tuberculosis in Scotland from 2000–2013, with a particular interest in the genetic strain typing data. This chapter fits well with the first stage in managing infectious disease outbreaks as we aim to identify clusters of tuberculosis cases. To meet this aim, we design a bespoke system for HPS to use which can help ensure that they do not miss potential strain clusters of cases. This is important for limiting the spread of the disease in the population. Furthermore, we also explore the use of multivariate methods such as multidimensional scaling for finding clusters. Chapter 2 stands well on its own and can be read independently of the rest of the thesis without much loss in understanding.

The 2009 influenza pandemic provided a great research opportunity as it was an immediate and pressing public health priority. We decided to take advantage of the wealth of data that was being generated on the pandemic – some of which are only really generated during public health emergencies – and begin to look at the pandemic in Scotland in Chapter 3. There we analyse a rich but hastily assembled dataset on early case reports of pandemic influenza as fully as possible to better understand the characteristics of those being infected. To achieve this, we use several logistic regression models to test whether or not different symptoms were being reported more frequently between males and females as well as people within different age groups. Since numerous statistical tests are conducted, we take into consideration the issue of multiple testing when making conclusions. By linking the case data to data on deprivation status for residents in Scotland, we also look at the effects of deprivation on treatment and severity of illness as described by the reported numbers of symptoms. Furthermore, a separate linked dataset on contacts of those cases was kept by HPS, which allows us to explore case-contact relationships at the start of an epidemic.

In Chapter 4, we conduct a literature review on methods of estimating the reproductive number for infectious diseases – arguably the most crucial parameter that has to be found during an outbreak. Here, the emphasis is on infectious diseases which can be contracted by humans. The review is intended to be general to give an idea of the vast range of methods out there for achieving this single task and does not go into great detail on particular methods. This is why the review is broad and describes mathematical modelling methods and statistical estimation methods for deriving the reproductive number. The main purpose of the chapter is to give background on some of the methods which can be used in Chapter 5.

In that chapter, we look at the issue of surveillance for pandemic influenza in Scotland. We utilise some of the statistical estimation methods introduced in the literature review chapter to find the reproductive number during the exponential growth phase (early phase) of the pandemic. The main motivation for the estimation methods we decided to use were based upon the data that was made available by HPS. Since postcode data relating to the place of residence of cases is routinely collected in Scotland when they visit healthcare providers, we saw this as an opportunity to develop a method for estimating reproductive numbers which incorporates a spatial element. Hence, we attempt to extend one of the existing methods for tracking the reproductive number over time and theorise that using the spatial data can result in improved estimates; this is as opposed to just using data on the time of onset of symptoms.

We finally turn our attention to control measures in Chapter 6. Similarly to before, we conduct a literature review on methods of estimating the effect of vaccines to serve as reference material for the chapter immediately following the review. The reason why so many methods exist is that it can be tricky to obtain reliable vaccine effect estimates as it is often unethical to deny individuals a vaccine in practice and thus, we have to use observational studies rather than the superior design – randomised controlled studies. In the literature review, we consider the advantages and disadvantages of using different study designs and different clinical outcomes on which, vaccine effect is measured against. As well as this, we also look at methods of reducing bias in observational studies. We then apply some of these methods in Chapter 7 where we obtain estimates of the effect of seasonal influenza vaccine for seasons 2011–12 and 2012–13 in Scotland. Again, the methods that we decided to employ for estimating vaccine effect were driven by the data that HPS was able to provide.

# Chapter 2

# Tuberculosis in Scotland

The first stage in managing an infectious disease is to try to detect the main sources of an outbreak. In this chapter, we consider how to detect outbreaks and more specifically, look at how applying statistical methods to genetic typing data can assist with the detection of clusters of linked cases which can be instrumental in causing outbreaks. The focus of our attention here is on a vaccine-preventable infectious disease which can be spread via the respiratory route – tuberculosis (TB).

TB is an ancient disease that is believed to have co-evolved with its hosts over several millennia. It was initially a disease that affected lower mammals but has also affected humans for thousands of years – in fact, the earliest evidence of the disease found in humans so far dates back to around 9,000 years ago [167]. The disease really came into prominence in Europe when cities became larger and more crowded. Subsequently, a major factor that allowed the disease to spread to other regions of the world was the empire-building tendencies of powerful European nations [22]. The precision of data on TB cases and deaths prior to the 20th century is questionable. However, this data exists for the UK from the early 20th century when TB was an urgent public health issue, and in some years there were over 100,000 cases reported with around 40,000 deaths and a death rate of over 100 per 100,000 population (Figure 2.1).

From the early part of the 20th century, the numbers of cases and deaths declined sharply, with the exception of the two world war periods where there were increases in deaths related to TB (centre plot of Figure 2.1). The reductions in TB cases and deaths can probably be attributed to a combination of a number of

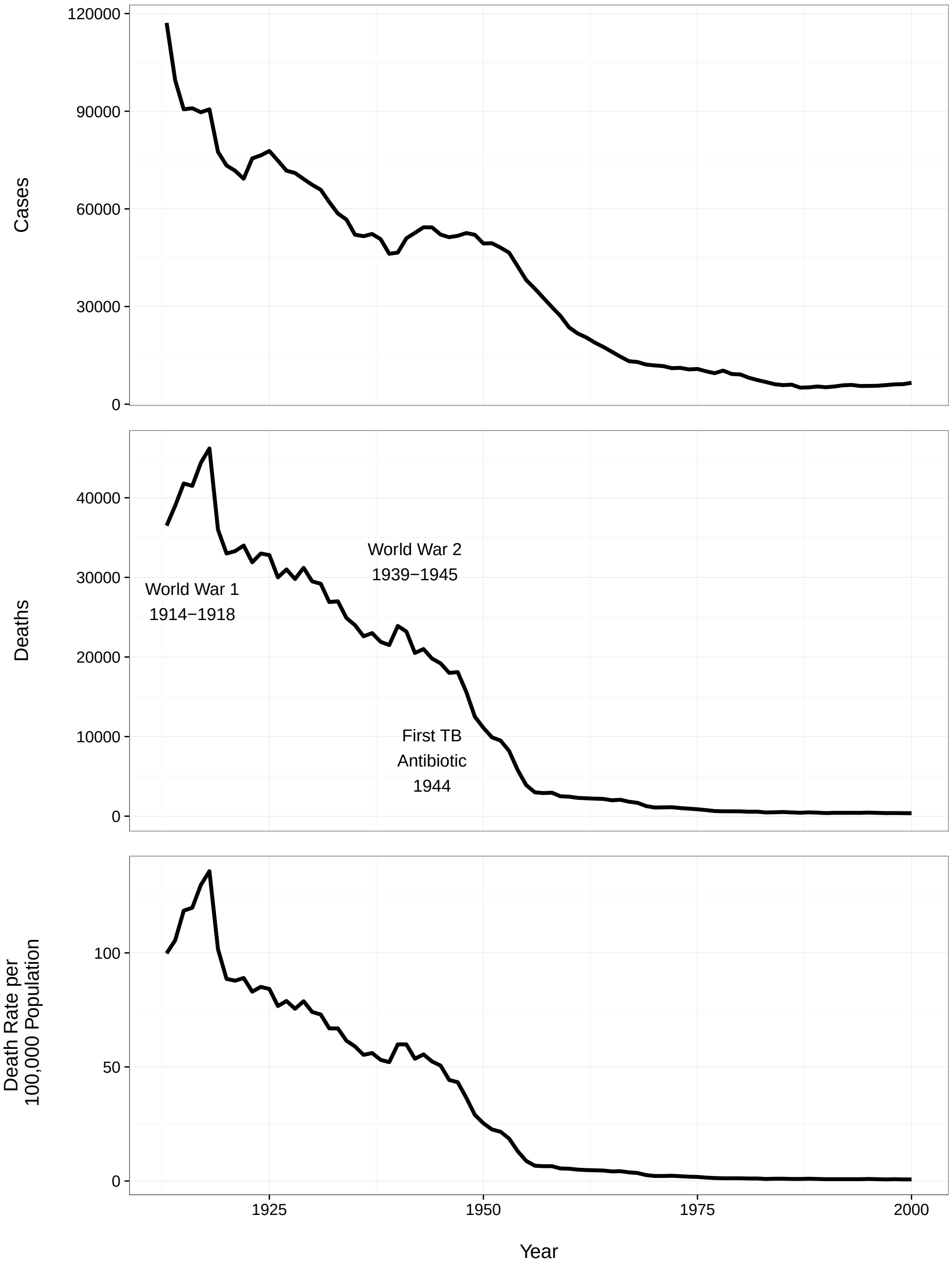


**Figure 2.1:** Tuberculosis cases, deaths and death rates per 100,000 population in the UK from 1913 to 2000. The data used to create the plot is publicly available from Public Health England [295]. Spikes in cases and deaths were observed during the world wars. The first antibiotic (streptomycin) against TB was used in 1944 and this was one of the major developments in the fight against TB.

changes which took place during this period. Some of the major factors include improved living conditions, reduction of poverty, better nutrition and advances in public health such as sanatoria and pasteurisation of milk [107]. Scientific and medical breakthroughs also played their part in reducing TB cases. One of the first breakthroughs came in 1895 with the discovery of the X-ray by Wilhelm Roentgen [215] as this allowed the disease to be diagnosed and tracked. Effective medical treatments for TB came much later and caused the numbers of cases and deaths to drop even more. For example, the first effective antibiotic against TB, streptomycin, was isolated in 1944 [325], and then isoniazid was developed in 1952 (an effective oral medication against TB) [36]. Despite work on a TB vaccine starting as far back as 1900, it took a long time to prove it was safe and development was also disrupted by the first world war. The vaccine (now known as a BCG vaccine – Bacille Calmette Guérin) only began to be used more widely in the UK following the second world war [333].

Although the disease is now much less common in the UK since the introduction of improved public health measures, effective medical treatments and screening, numbers of cases have been gradually increasing in the last 20 years and so it is important to continue monitoring TB cases to see how the situation develops here. One of the main reasons for the increasing numbers of cases is because of immigration to the UK from people originally living in countries where TB is more prevalent. Often these cases arrive with latent TB infection (disease is present but person is not symptomatic) which later reactivates into active TB infection [280]. A further reason for the increase in attention being paid to TB is the fairly recent rise in multi-drug resistant TB (MDR-TB) and the even more serious, extensively drug resistant TB (XDR-TB) cases around the world and in the UK. These more worrying forms of TB can cause a great deal of burden on any health care system and this makes it a matter of great concern.

Here we introduce some bespoke systems implemented in the R project for statistical computing [302] that can aid in the identification of potential clusters of TB cases in Scotland. This tool can be used by Health Protection Scotland (HPS) as part of their surveillance efforts and the system is important as missing clusters could lead to further spread of the disease throughout the population. Rapid detection of potential clusters of TB cases allows resources to be better directed and aids in overall efforts to control TB outbreaks in Scotland. We also provide some general exploratory statistical analysis on these cases to give an overview of TB trends in Scotland in recent years.

## 2.1 Background on Tuberculosis

Tuberculosis is caused by bacteria belonging to the *Mycobacterium tuberculosis* complex (MTBC) (see Section 2.1.1 later for more details on this and other mycobacteria organisms) which in most cases, will affect the lungs. When it does affect the lungs, we would have *pulmonary cases* and only these cases are infectious. If TB spreads outside of the lung causing other forms of TB, then the case is said to be *extrapulmonary*. Pulmonary cases can spread the disease to others through the air as they cough and sneeze. Inhaling a small amount of germs can cause a person to become infected, but usually only prolonged exposure to someone with the illness will cause the disease to spread (for instance, sharing a bedroom with someone infected with TB). Another way that TB can be spread is through contact from open wounds and ulcers but that is much less common.

The symptoms are unpleasant and at times, coughing will become violent, producing spatum (mucus from the lower airways) and blood. Other symptoms of active TB cases include fever, night sweats, chest pain, fatigue, loss of appetite and unexplained weight loss. One way of diagnosing TB is to check for the presence of TB bacteria in mucus samples using a microscope. Another way is to use X-rays to examine if there is scarring in the lungs. It is considerably more difficult to detect latent TB cases, but it is possible via skin testing or blood testing methods.

TB is a serious condition which can be fatal but can be cured with antibiotic drugs and is ultimately preventable. Usually a six-month course of antibiotics will be sufficient to cure the disease. To prevent TB, a BCG vaccine is available in the

UK and offered to babies who are at an increased risk of coming into contact with someone infected with TB. The effectiveness of the vaccine has been estimated to vary anywhere between 0 to 80% for up to 15 years but on average, it has been found to reduce the risk of TB in newborns and infants by over 50% in a meta-analysis of published literature [68]. Even though TB can be treated with antibiotics, there are forms of the disease that are drug resistant which are termed MDR-TB and XDR-TB. A major reason that these forms of TB have arisen is because of inadequate drug regimens being taken by patients; for instance, drug regimens not being taken for the appropriate duration [304]. These forms of TB are much more difficult and costly to treat [276] and they are of massive concern as the numbers of drug resistant TB cases have been rising [293].

The disease remains as the leading cause of death among those infectious diseases which can be cured, with around 1.3 million deaths worldwide from it in 2012 out of 8.6 million falling ill with TB in that year [383]. The disease still occurs in every part of the world and around one-third of the world's population has latent TB. The bacteria can remain latent in the body for a long time, possibly even life-long. Most cases are from developing countries in South-East Asia and Africa as shown in Figure 2.2 which was produced by the World Health Organization (WHO) for a report in 2011 [394].

It is known that many new cases of TB in the UK are people of ethnic minorities coming from countries with higher incidences of TB [281]. Hence, individuals arriving to the UK from countries with an incidence of 40 or more cases per 100,000 population per annum must be screened upon arrival [152]. In the UK, there are currently (in 2013) around 9,000 cases reported per year and most of these come from the largest and most population-dense cities such as London [300]. The main risk groups for TB are those who are co-infected with the Human Immunodeficiency Virus (HIV). These people are 21–34 times more likely to contract TB [383]. People with weakened immune systems (immunosuppression) through other routes such as those with drug or alcohol misuse problems also have an increased risk of TB. Furthermore, it has also been linked to overcrowding and malnutrition

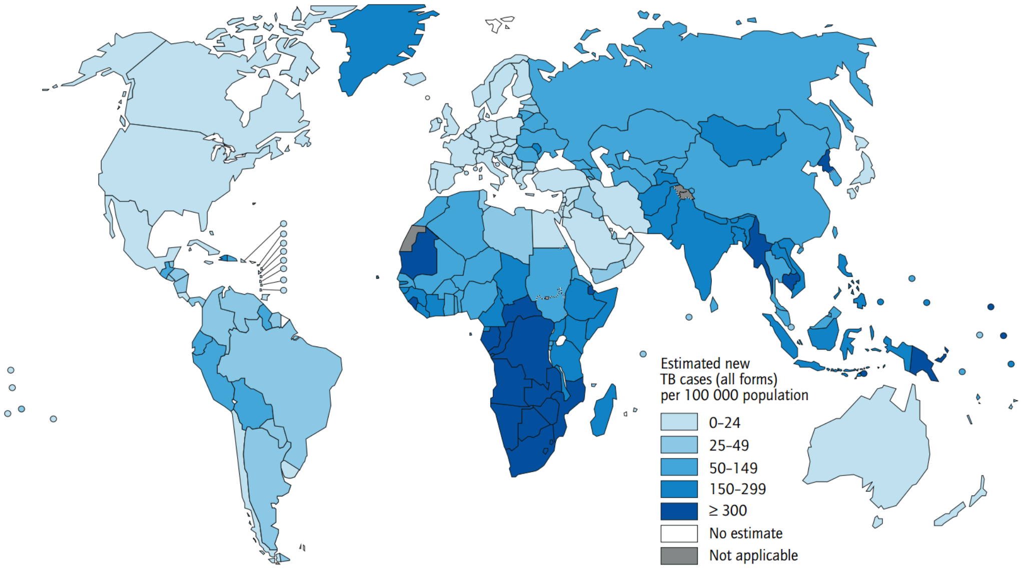


**Figure 2.2:** Estimated TB incidence rates for countries around the world in 2011 as given in the World Health Organization's global tuberculosis report 2012 [394].

which are both associated with poverty [213].

### 2.1.1 Mycobacteria Organisms

Mycobacteria organisms can be divided into a few main groups. Firstly, we have the mycobacteria that cause TB in humans and/or animals and these are grouped within the "Mycobacterium tuberculosis complex" (MTBC). Among the closely related species included in this group are *M. tuberculosis* (where 'M.' stands for 'Mycobacterium'), *M. Africanum*, *M. bovis*, *M. microti* and *M. canetti* [45]. Out of these, *M. tuberculosis* is by far the most commonly diagnosed species in humans, while *M. Africanum* and *M. canetti* are mainly seen in cases from Africa [79, 307]. Bovine TB is another alias for *M. bovis* and comes mainly from cattle but it can also occur in domestic animals such as sheep, goats, cats and dogs [19]. However, bovine TB is now much less common in developed countries since the introduction of pasteurised milk. Finally, *M. microti* is rare in humans and occurs mainly in small rodents [59].

Another group of mycobacteria are those which cause leprosy. This includes *M.*

*leprae* and *M. lepromatosis* [290]. Lastly, there are the nontuberculosis mycobacteria (NTM) group which are sometimes known as environmental mycobacteria or atypical mycobacteria. These mycobacteria can cause pulmonary disease similar to TB as well as other conditions such as skin disease [131]. This group includes a vast number of organisms. Some examples are *M. chelonae*, *M. abscessus*, *M. fortuitum*, *M. xenopi*, *M. gordonae* and *M. avium*.

### 2.1.2 Surveillance in Scotland

All mycobacterial isolates in Scotland are tested by the Scottish Mycobacteria Reference Laboratory (SMRL) which is the national reference laboratory there. This laboratory reports weekly results to HPS and also provides epidemiological information if requested. In 2000, the Enhanced Surveillance of Mycobacterial Infection (ESMI) scheme was established and coordinated by HPS [160]. This scheme was designed to incorporate the European minimum dataset for the collection of TB data (see Schwoebel et al. [327] for more details), which allows Scotland to contribute to the WHO and European TB data collection systems. Essentially, cases reporting to the ESMI scheme will have some clinical and demographic information available.

The trends from notifications data collected by HPS as illustrated in Figure 2.3 show that the numbers of TB cases have been increasing since 2005 [162]. In particular, there has been an increase in non-pulmonary cases. As part of the global effort to tackle TB, the Scottish Government wished to reverse these trends, and a TB action plan for Scotland was established by them in collaboration with HPS in 2011 [330]. The main goal of this action plan was to identify where TB-related services could be enhanced and within that goal, the main themes included looking at how surveillance could be improved; the work in this chapter contributes in this area.

### 2.1.3 Strain Typing

People suspected of having TB submit clinical samples such as a spatum sample. If *M. tuberculosis* grows in culture media from this specimen, then it becomes an

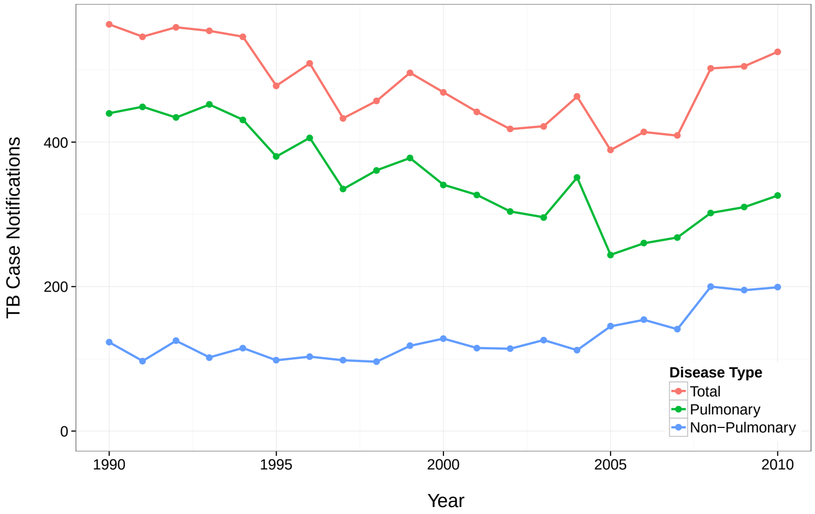


**Figure 2.3:** TB notifications in Scotland from 1990 to 2010. The plot was produced using data made available by HPS [162].

*isolate* and only these can have their genetic material analysed (genotyped). A key part of TB infection control comes from universal molecular strain typing of all MTBC isolates. This means that at least one isolate is submitted from every culture-positive patient with TB. This process is considered essential for epidemiological investigations such as analysing the spread of specific genotypes. This is because combining genotyping results with epidemiological information can help identify TB cases which are part of the same transmission chain. Moreover, it can help in identifying clusters of cases, unexpected outbreaks, false positives and can distinguish between activation of latent TB infection and recent or newly acquired infection with development of TB disease.

The first method for the typing of TB strains was IS*6110* restriction fragment length polymorphism analysis (RFLP) [363], but producing data using this method proved technically challenging. Another method named *spoligotyping* [195] was later developed which did not have the data production issues associated with RFLP. However this method suffered from having low discriminative power for distinguishing between strains. Hence, a better method for strain typing than RFLP or spoligotyping was needed.

### VNTR-MIRU

The most widely used method for TB strain typing at present is VNTR-MIRU (variable number tandem repeats of Mycobacterial interspersed repetitive units) which produces results in a standardised code and so can be analysed easily. Note that the method is also referred to as MIRU-VNTR in reports as well. In DNA, a *tandem repeat* occurs when there is a repeated pattern of two or more nucleotides which are directly adjacent to each other; an example of three tandem repeats would be ACA ACA ACA. If there are 10–60 nucleotides being repeated, it is termed as a *minisatellite*, and when the number of nucleotides being repeated is not known or variable, we have VNTR. In many bacterial genomes, there are regions of variable tandem base repeats (minisatellites) and these regions are prone to variations in size. If a sufficient number of regions are present, the variations in size can be used as the basis of a typing procedure – this is the fundamental idea behind distinguishing different *M. tuberculosis* strains using VNTR-MIRU.

This method only became possible as a method of TB strain typing after the publication of the *M. tuberculosis* H37Rv genome (total genetic content of *M. tuberculosis*) where 41 MIRU loci (specific positions of a DNA sequence) were reported. The *M. tuberculosis* strains are distinguished by examining the difference in the number of copies of tandem repeats in specific loci of the genome. Originally 12 of these 41 MIRU loci were considered enough to be useful for distinguishing between organisms. A fictitious example of a 12 loci MIRU profile would be:

$$123456654321$$

In this example, the first digit (1) refers to the first locus where there is one tandem repeat, the second digit refers to the second locus which has two repeats and the remaining digits can be interpreted similarly. Note that it is possible to have tandem repeats in excess of 9 and any double-digit repeats would have to be clearly separated in the data to be noticed. In our data, we distinguished double-digit repeats by using capital letters in the fashion of A = 10, B = 11, C = 12, etc.

Later it was found that levels of discrimination could be increased by using 15 loci, which Scotland had been using since the middle of 2005. This has since been changed to a standard panel of 24 loci in an attempt to optimise reproducibility and discrimination [365]. It is now the global standard for typing MTBC and has been used in Scotland since November 2010.

#### Incomplete MIRU Profiles

In some isolates, there can be a missing locus (or multiple missing loci) due to the quality/quantity of DNA samples as well as the general complexities of the multiplex polymerase chain reaction (PCR) procedure used to amplify the DNA. When a missing locus is present, it is represented by an underscore (_) in the data. For instance, in the following MIRU profile, the third locus is missing.

12_456654321

MIRU profiles with missing loci can only form potential rather than definitive clusters. In the majority of isolates with 24 loci profiles, there are between 0–3 missing loci. In our dataset, over 98% of 24 loci profiles had between 0–3 missing loci, and just under 50% were complete profiles (see Table 2.3 in Section 2.4 on finding potential strain clusters for more details). It is possible for isolates to have more than three missing loci but when there are more than three missing loci, a repeat DNA extraction is usually recommended where feasible.

### 2.1.4 Cluster Definitions

Definitions for different types of clusters of TB cases are given in the TB Strain Typing Cluster Investigation Handbook by Public Health England (PHE) [299]. We briefly summarise these definitions here. The definition of *clustered cases* is two or more persons with TB that are linked in time and share possible epidemiological links. This means that two cases would have spent time in the same locations at a similar time but the timing of the infectious period may not have been definite enough to be sure that transmission of infection between them would have taken place. The next step up from this is an *epidemiological cluster* where the epidemiological link is known. For this link, one case may have volunteered the name of another case while either of them was potentially infectious or the patients may have spent time in the same locations when either of them was infectious. Using strain typing data allows us to obtain *strain type clusters*. This is where two or more persons with TB have isolates with indistinguishable MIRU profiles and the dates of when they were diagnosed with TB are within two years of each other. An example of this is if we have two isolates with profiles:

123456654321
123456654321

The combination of epidemiological and strain typing evidence gives an *epidemiologically confirmed strain typing cluster*. This is where two or more persons with TB provide samples within two years of each other which produce indistinguishable MIRU profiles, and in addition to this, there are also known epidemiological links between them. For this type of cluster, we can confidently assert that the cases involved are genuinely linked.

### Potential Strain Clusters

In this investigation, we make use of either 15 loci MIRU profiles or 24 loci MIRU profiles as they were most widely available. For each MIRU length, we have separate rules for identifying potential clusters. For 15 loci profiles, the rule we use is that a case can have a maximum of one missing locus when forming a potential cluster with other cases that have complete profiles. However, all loci other than the missing locus must have exactly matching numbers of tandem repeats to the complete profile. Since around 99% of cases in our dataset with a 15 loci MIRU profile are either complete or have only one missing locus (Table 2.3), this rule will include almost all of the cases when looking for potential clusters. In this example:

123456787654321
1234_6787654321
123456_87654321

the second and third profiles can form a potential strain cluster with the complete profile at the top of the list. However, it must be noted that the second and third profiles cannot form a potential strain cluster on their own without the complete profile. In other words, all potential clusters must involve at least one complete profile. A further notable point is that incomplete profiles can be a member of more than one potential cluster. For instance, in the example below, the case with the incomplete MIRU profile shown on the second row would form two potential clusters.

| **Potential Cluster 1** | **Potential Cluster 2** |
|---|---|
| 123456787654321 | 123476787654321 |
| 1234_6787654321 | 1234_6787654321 |
| 123456_87654321 | |

For 24 loci profiles, we follow the rules set by PHE [296] and allow a maximum of two missing loci to form potential clusters. Thus, an example of a potential cluster involving 24 loci profiles could be:

123456781234567812345678
1234567_1234567812345678
123456781234567__2345678

Potential clusters must be investigated further to determine possible epidemiological links between cases, particularly under the scenario where incomplete profiles belong to multiple potential clusters. If there is deemed to be sufficient evidence for a true cluster to exist after the investigation, then we may have enough evidence to *impute* the missing digit by using the MIRU profiles of cases that are linked to the incomplete profile. Furthermore, in the scenario where an investigator requires further assurance as to what a missing digit may be, it may be possible to have the gaps filled in to confirm (or disconfirm) the existence of a cluster by repeating the genotyping procedure. However, even resampling does not guarantee that the missing digit will be known as the sample may again, give an incomplete reading. In any respect, investigating potential clusters by utilising incomplete profiles in this way means that some clusters which may have been previously missed can now be identified.

### Single Locus Differences

Although rare, it is nevertheless possible for linked TB cases to have a variation in one repeat unit between their MIRU profiles. The reasons why this happens are not fully known at present but it is believed that this may occur due to the evolution of a strain over a number of years or through transmission to different people. A study by Shorten et al. [332] looked at a large London outbreak to research the evolutionary relationship of Mycobacterium tuberculosis isolates from 13 patients that had strains that clustered together based on epidemiological investigations. Although these strains clustered, their MIRU profiles differed slightly. Based on their findings, they concluded that evolution had occurred during the course of transmission between cases, and a major reason for this may have been that the strains evolved through acquiring additional antimicrobial resistance. Moreover, another study by Walker et al. [370] used whole genome sequencing to detect microevolution within strains and their findings support the evidence that microevolutionary events can change a MIRU genotype within a host. Examples of this scenario have also been observed in Scotland. In one investigation, it was found that an index TB case transmitted the infection to their partner and they had identical 24 digit MIRU profiles with the exception of one digit (the 4th digit was a 6 in the index case but a 9 in the secondary case) [331].

## 2.2 Data

Data was available from all the clinical samples tested by the SMRL from January 2000 up to the beginning of August 2013. The vast majority of the data came from TB cases reporting to the ESMI scheme which offers extra information on TB cases over those cases not reporting to the ESMI scheme (we shall refer to these as “NOESMI” cases). Specifically, available data from cases reporting to the ESMI scheme includes details of healthboard, gender, age group (specific age or date of birth was not given), whether or not the form of TB is pulmonary or non-pulmonary, the date of notification, the culture result and MIRU profile. In contrast to this, we know only the date of notification, the culture result and MIRU profiles for the NOESMI cases. Note that the SMRL only considered obtaining MIRU profiles if the culture result was positive for a form of TB within the MTBC.

### 2.2.1 Analysis Dataset

In total, we had data on 5,701 TB case notifications – 5,479 of these reported to the ESMI scheme and 222 of these were NOESMI cases. For later analyses, we removed some of these cases as we could not be certain that they were actually TB cases. Firstly, we removed all cases who had a negative culture result. On top of this, we also removed cases where the culture result was contaminated, not known or failed to grow. Finally we removed cases that did not have their samples sent for culture at all; the numbers of cases removed are summarised in Table 2.1. All of the cases removed here were ESMI cases as all of the NOESMI cases tested positive for some form of TB. This left 4,106 cases which could be used in later analyses and we also note that 1,358 had 15 loci MIRU profiles and 1,141 had 24 loci profiles. These cases can be used for finding strain-type clusters.

| | $n$ |
|---|---:|
| *All TB Data* | |
| ESMI | 5,479 |
| NOESMI | 222 |
| Total | 5,701 |
| *Data Removals* | |
| Contaminated | 4 |
| Culture Negative | 974 |
| Culture Result Not Known | 117 |
| Failed to Grow | 23 |
| Not Sent for Culture | 477 |
| *Data for Analysis* | |
| Remaining | 4,106 |
| With 15 Loci MIRU Profile | 1,358 |
| With 24 Loci MIRU Profile | 1,141 |

**Table 2.1:** Numbers of TB notifications in our dataset along with numbers of cases removed before analysis for various reasons. Also shown are the numbers which did and did not report to the ESMI scheme and the number of cases with 15 and 24 loci MIRU profiles available.

### 2.2.2 A Note on Greater Glasgow and Clyde

Since our dataset spans 2000–2013 and contains healthboard information, we must note that the time period of the data coincides with the time when the Greater Glasgow and Clyde (GGC) healthboard was created (1st April 2006). It was created by amalgamating the Greater Glasgow healthboard with a large share of the Argyll and Clyde healthboard. Approximately 75% of the residents from Argyll and Clyde joined GGC while around 25% joined the Highland healthboard [180]. In order to make the output of our summary tables and graphs more concise and comparable over the whole time period of the data when looking at aggregate counts within healthboards, we assigned 25% of the Argyll and Clyde cases before 2006 to Highland and 75% of their cases to GGC.

## 2.3 Exploratory Analysis

To begin with, we explored the characteristics of the 3,884 TB cases that reported to the ESMI scheme between January 2000 and August 2013. To gain an idea as to whether or not the numbers of cases by sex, age group and within healthboards were proportional to the population at large, we compared case counts against the 2013 mid-year population estimates provided by the National Records of Scotland (NRS) [256]. Using these together allowed the calculation of an observed/expected ratio (O/E ratio) for each category along with an associated 95% confidence interval [90]; these are shown in Table 2.2.

Over the whole period, there were much more male cases (62%, $n$ = 2,415) than female cases (38%, $n$ = 1,467). Since the gender split in Scotland is almost equal between males and females, the results indicate that there are more male cases than we would expect based on the mid-year population counts within each gender (O/E ratio: 1.28, 95% CI: 1.22 to 1.33). A larger than expected proportion of cases come from those aged 15–34 (O/E ratio: 1.32, 95% CI: 1.25 to 1.39) and those aged 65+ (O/E ratio: 1.34, 95% CI: 1.26 to 1.43). However, there were very few cases in the 0–14 age range (only 55 cases or 1.4%).

| | **Cases** | **2013 Mid-Year Population** | **Expected Cases** | **Obs / Exp** | **95% CI** |
|---|---|---|---|---|---|
| *Sex* | | | | | |
| Female | 1,467 | 2,741,168 | 1,997 | 0.73 | 0.70 – 0.77 |
| Male | 2,415 | 2,586,532 | 1,885 | 1.28 | 1.23 – 1.33 |
| *Age Group* | | | | | |
| 0-14 | 55 | 852,005 | 621 | 0.09 | 0.07 – 0.12 |
| 15-34 | 1,324 | 1,376,449 | 1,003 | 1.32 | 1.25 – 1.39 |
| 35-64 | 1,576 | 2,152,384 | 1,568 | 1.01 | 0.96 – 1.06 |
| 65+ | 926 | 946,862 | 690 | 1.34 | 1.26 – 1.43 |
| *Healthboard* | | | | | |
| Ayrshire & Arran | 114 | 372,210 | 269 | 0.42 | 0.35 – 0.51 |
| Borders | 31 | 113,870 | 82 | 0.38 | 0.26 – 0.53 |
| Dumfries & Galloway | 47 | 150,270 | 109 | 0.43 | 0.32 – 0.58 |
| Fife | 123 | 366,910 | 265 | 0.46 | 0.39 – 0.55 |
| Forth Valley | 100 | 299,680 | 217 | 0.46 | 0.38 – 0.56 |
| Grampian | 303 | 579,220 | 418 | 0.72 | 0.64 – 0.81 |
| Greater Glasgow & Clyde | 1,791 | 1,137,930 | 822 | 2.18 | 2.08 – 2.28 |
| Highland | 128 | 321,000 | 232 | 0.55 | 0.46 – 0.66 |
| Lanarkshire | 338 | 652,580 | 471 | 0.72 | 0.64 – 0.80 |
| Lothian | 678 | 849,700 | 614 | 1.10 | 1.02 – 1.19 |
| Orkney | 3 | 21,570 | 16 | 0.19 | 0.04 – 0.56 |
| Shetland | 3 | 23,200 | 17 | 0.18 | 0.04 – 0.52 |
| Tayside | 177 | 412,160 | 298 | 0.59 | 0.51 – 0.69 |
| Western Isles | 13 | 27,400 | 20 | 0.66 | 0.35 – 1.12 |
| *Disease Type* | | | | | |
| Non-Pulmonary | 1,179 | | | | |
| Pulmonary | 2,705 | | | | |

**Table 2.2:** Observed and expected numbers of tuberculosis cases in Scotland reporting to the ESMI scheme over the period January 2000 to August 2013 by various characteristics. Expected numbers were estimated using 2013 mid-year population counts provided by the National Records of Scotland [256]. Confidence intervals were calculated assuming observed cases are Poisson variates [90]. Note that a very small number of cases were missing information on some of the characteristics – 2 with missing sex, 3 missing age and 8 missing healthboard.

By far the most cases came from the largest healthboard which is GGC (46% of cases, $n$ = 1,791), while many cases also came from the next largest healthboard, Lothian (17.5%, $n$ = 678). These were also the only two healthboards with a larger proportion of cases than would be expected relative to their respective populations (O/E ratio for GGC: 2.18, 95% CI: 2.08 to 2.28; O/E ratio for Lothian: 1.1, 95% CI: 1.02 to 1.19). The rest of the healthboards had less cases than expected based on their population size, but it must be noted that medium-sized healthboards such as Grampian and Lanarkshire also had a reasonable amount of cases with 303 and 338 cases respectively (around 8% of total cases each). With regards to the type of TB that cases were having, about 70% of cases had pulmonary TB ($n$ = 2,705) and 30% had non-pulmonary TB ($n$ = 1,179).

### 2.3.1 Trends Over Time

Since ESMI and NOESMI cases had date of notification, we could plot the combined total cases each year which is shown in Figure 2.4. In 2013, data only goes up to the beginning of August and hence, we made a simple estimation of counts for 2013 by multiplying the 2013 counts by $\frac{12}{7}$. It is important to note that while estimating counts in this way allows the trend over the whole time period to be more visible, we cannot rule out the possibility that the decreases in case counts seen in 2013 (Figures 2.4 and 2.5) could be due to a reporting delay – if a reporting delay does exist, we would only be able to obtain a more reliable count for 2013 with data extracted after some duration in 2014. A further note is that 122 cases did not have a date of notification and so they are not included in the case counts. It is clear that cases peaked firstly in 2004 and then peaked again between 2008 to 2010, but numbers of TB cases have been dropping since then. During the highest peaks between 2008 to 2010, there were just under 350 cases each year.

Looking at only the ESMI data allows us to see trends over time by gender, disease type, age group and healthboard, and these trends are shown in Figures 2.5 and 2.6. There we can see that female cases peaked in 2004 and have been decreasing since 2008 while male cases peaked in 2010 and have been decreasing since then. In Figure 2.5b, we can notice that the non-pulmonary cases have been

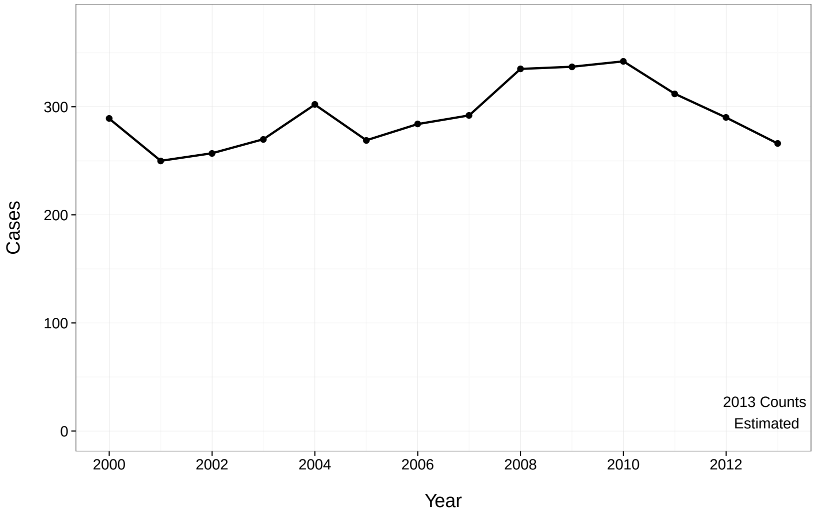
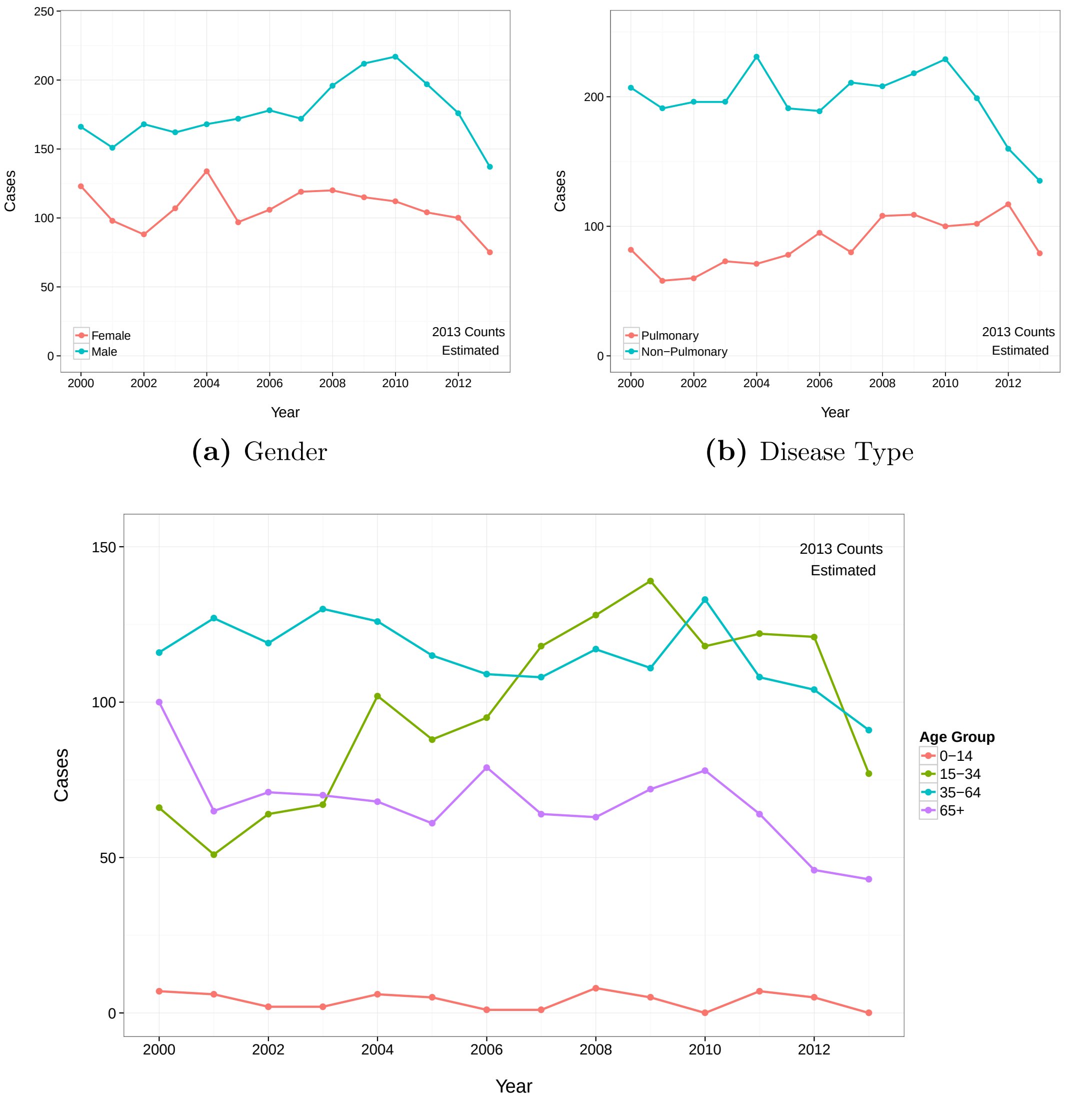


**Figure 2.4:** TB cases over time. Counts relate to the total of ESMI and NOESMI TB cases where year of notification was available. Note that in 2013, data only goes up to the beginning of August and an estimate for the total year counts has been made by multiplying by 12/7.

steadily increasing over time but numbers of pulmonary cases have been getting lower since 2010. Figure 2.5b illustrates that case numbers have been relatively stable over the years for those aged 0–14 and 35–64 and have been decreasing for those aged 65+ since 2010. However, the number of cases aged 15–34 increased from around 50 cases in 2001 to just under 150 cases in 2009 but since then, case counts have been decreasing slightly and seem to be stabilising in recent years.

The numbers of cases each year by healthboard are shown in Figure 2.6. Although there are some occasional bursts of cases in a year for a healthboard (for instance, Borders in 2009, Forth Valley in 2006, Highland in 2010, and Tayside in 2008), numbers of cases do not fluctuate dramatically over the period. An encouraging sign comes from the steady decrease in cases from Greater Glasgow and Clyde since 2009. However, numbers of cases in Lothian have remained at a similar level over the years.



**(a)** Gender

**(b)** Disease Type

**(c)** Age Group

**Figure 2.5:** TB cases over time by gender, disease type and age group where year of notification was available. Note that in 2013, data only goes up to the beginning of August and an estimate for the total year counts has been made by multiplying by 12/7.

### 2.3.2 Strain Clusters with Complete MIRU Profiles

Using only TB cases with complete 15 or 24 loci MIRU profiles, we found all of the strain clusters formed by having at least two TB cases with identical MIRU profiles. Out of the 1,193 cases with complete 15 loci MIRU profiles, 722 cases were in one of the 130 different strain clusters; from the 547 cases with complete 24

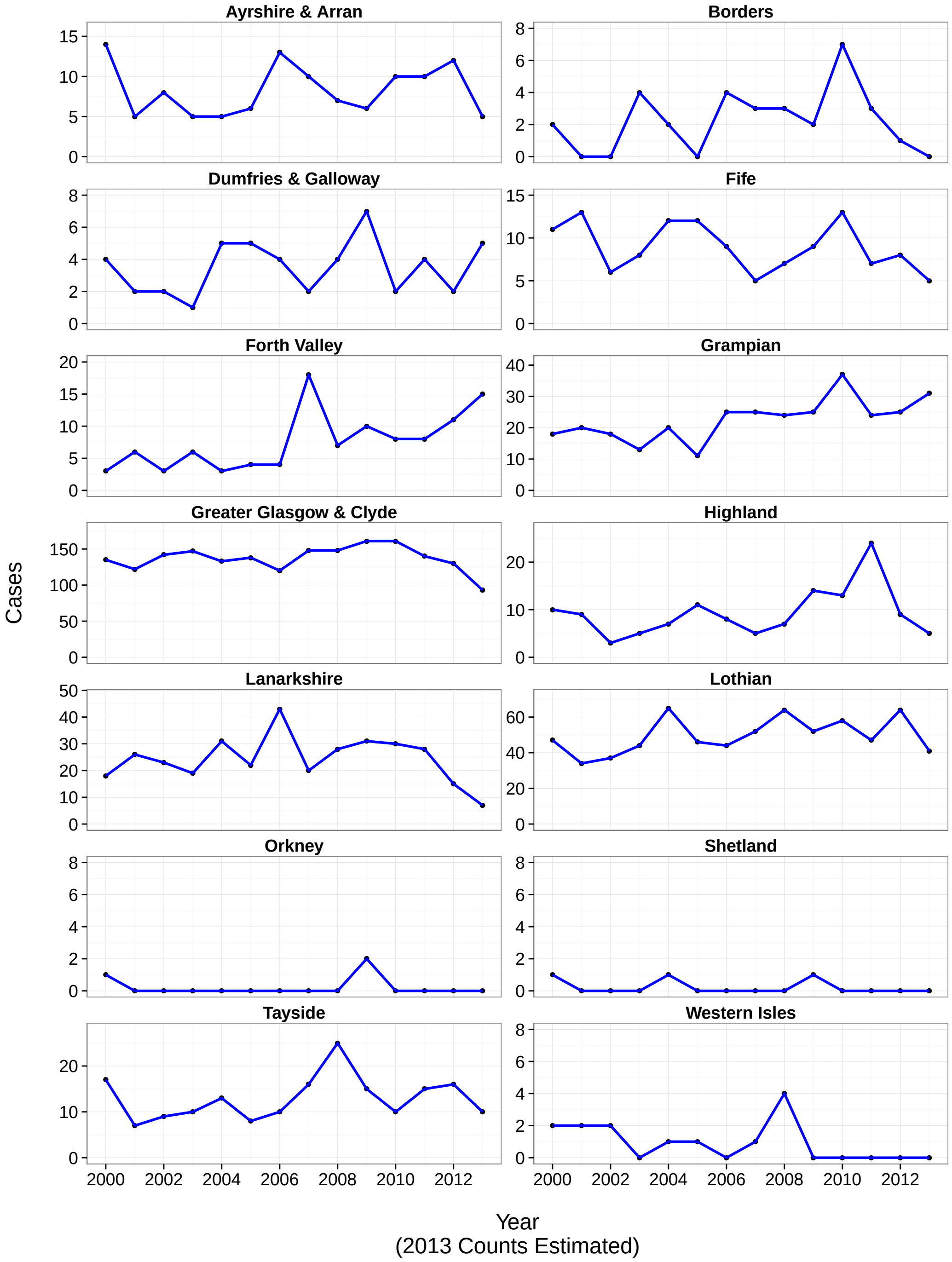


**Figure 2.6:** TB cases over time by healthboard where year of notification was available. Note that in 2013, data only goes up to the beginning of August and an estimate for the total year counts has been made by multiplying by 12/7. Also note that the $y$-axis varies by plot to account for differences in healthboard size.

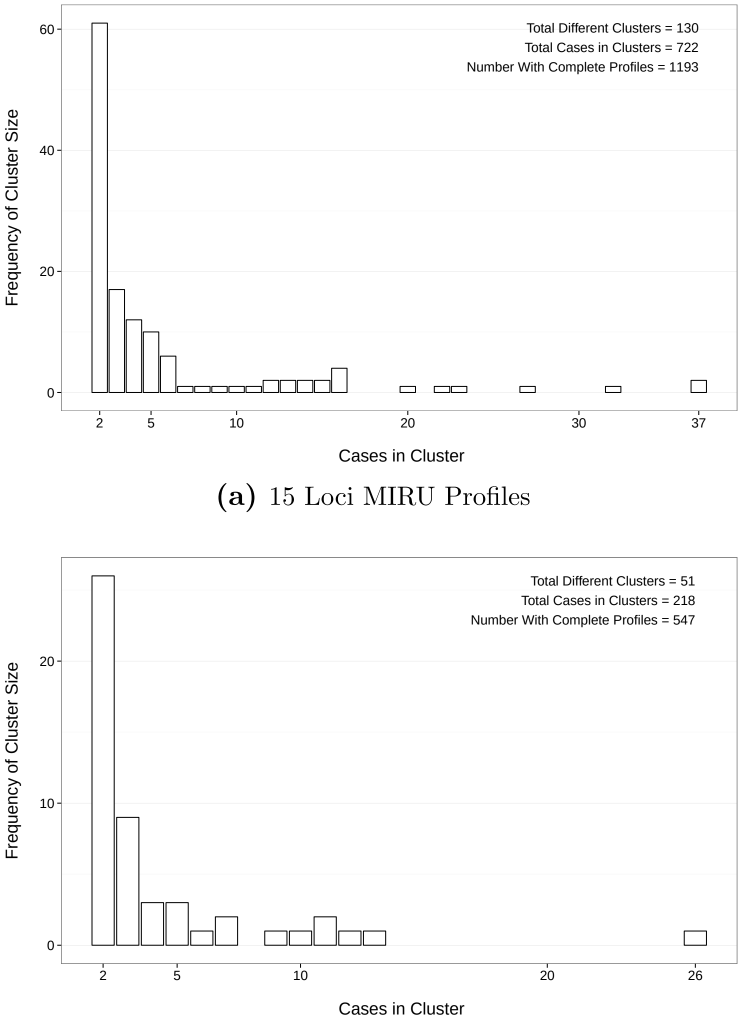
loci MIRU profiles, 218 were in one of the 51 different strain clusters. This means that 60.5% of those with complete 15 loci profiles share a profile with at least one other case and 39.9% of those with complete 24 loci profiles have a shared profile with another case.

Figure 2.7 shows the frequencies of clusters of different sizes. For 15 loci profiles, strain clusters varied in size from 2–37 cases, and there were 61 clusters of size 2 and two clusters with 37 cases in them. Meanwhile for 24 loci profiles, strain clusters had between 2 to 26 cases in them; 26 clusters with 2 cases in them and 1 cluster with the largest size of 26 cases.

## 2.4 Finding Potential Strain Clusters

To find potential strain clusters, we made use of 15 and 24 loci MIRU profiles which have missing loci. More specifically, for those with 15 loci profiles, we examined cases with one missing locus, and for those with 24 loci profiles, we looked at cases with up to two missing loci (more background on the topic of potential strain clusters can be found back in Section 2.1.4). Hence, for cases with 15 loci profiles, we wanted to investigate the 150 cases with one missing locus to see if they can form potential strain clusters (Table 2.3), while for cases with 24 loci profiles, we wanted to consider a much larger number of cases ($n = 496$).

Table 2.3 also shows that there was proportionately much more missing data for cases with 24 loci profiles compared with 15 loci profiles; almost 90% of the 15 loci profiles were complete while under 50% of the 24 loci profiles were complete. This is a clear sign that although the discriminatory power of the 24 loci VNTR-MIRU genotyping procedure is superior for differentiating samples compared with using using a smaller panel of VNTRs [65], this advantage is offset to an extent by the additional difficulty in obtaining a complete MIRU profile. Another implication of this finding is that it clearly illustrates a greater need to consider potential strain clusters when using 24 loci VNTR-MIRU genotyping as our results suggest there will be at least one missing locus around half of the time the procedure is carried



**(a)** 15 Loci MIRU Profiles

**(b)** 24 Loci MIRU Profiles

**Figure 2.7:** Histogram of strain cluster size using only cases with complete MIRU profiles. A strain cluster here is formed by having at least two TB cases having identical MIRU profiles and hence, the minimum strain cluster size is two.

out.

It is also important to consider where missing loci appear most frequently as this gives an idea of where imputation of missing digits would be highly beneficial. From Figure 2.8, we can observe that missing loci are most common in loci 9, 18, 22 and especially in loci 23 and 24 – 7.6% of MIRU profiles were missing data for locus 9, 6.5% for locus 18, 5.4% for locus 22, 30.8% for locus 23 and 13.6% for locus 24. Moreover, as most of the missing loci appeared in loci 23 and 24, this

| | 15 Loci Profiles | | 24 Loci Profiles | |
|---|---|---|---|---|
| **Missing Loci** | ***n*** | **%** | ***n*** | **%** |
| 0 | 1,193 | 87.85 | 547 | 47.94 |
| 1 | 150 | 11.05 | 358 | 31.38 |
| 2 | 12 | 0.88 | 138 | 12.09 |
| 3 | 1 | 0.07 | 81 | 7.10 |
| 4+ | 2 | 0.15 | 17 | 1.49 |

**Table 2.3:** Number of missing loci by MIRU profile length.

reinforces the point that there is greater utility for finding potential clusters when using 24 loci MIRU profiles.

### 2.4.1 Gower's Dissimilarity Coefficient

In order to find potential strain clusters, we must identify the MIRU profiles that are identical to other complete MIRU profiles except for the missing loci. To do this, we must calculate a measure of similarity between profiles with missing loci and all complete profiles. Gower introduced a general coefficient for measuring dissimilarity (or similarity) which is capable of dealing with a mixture of different types of data [136]; these include dichotomous (binary), qualitative and quantitative variables. For finding similarities, we can treat each locus as a qualitative variable and hence, Gower's dissimilarity coefficient is suitable for our task. For a pair of cases $i$ and $j$ with strains of TB genotyped using 15-loci, we would have $k$=15 variables, with each variable representing an individual locus. Hence the dissimilarity between the strains for a pair of cases on each locus can be denoted by $s_{ijk}$.

For qualitative variables, $s_{ijk} = 0$ (no dissimilarity) the digits are the same in both cases for locus $k$. Conversely $s_{ijk} = 1$ (dissimilar) if the digits differ between the two cases in locus $k$. Gower's method also has a quantity $\delta_{ijk}$ which represents the possibility of making comparisons. Here, $\delta_{ijk} = 1$ when a comparison can be made between two MIRU profiles (both profiles have a digit in locus position $k$),

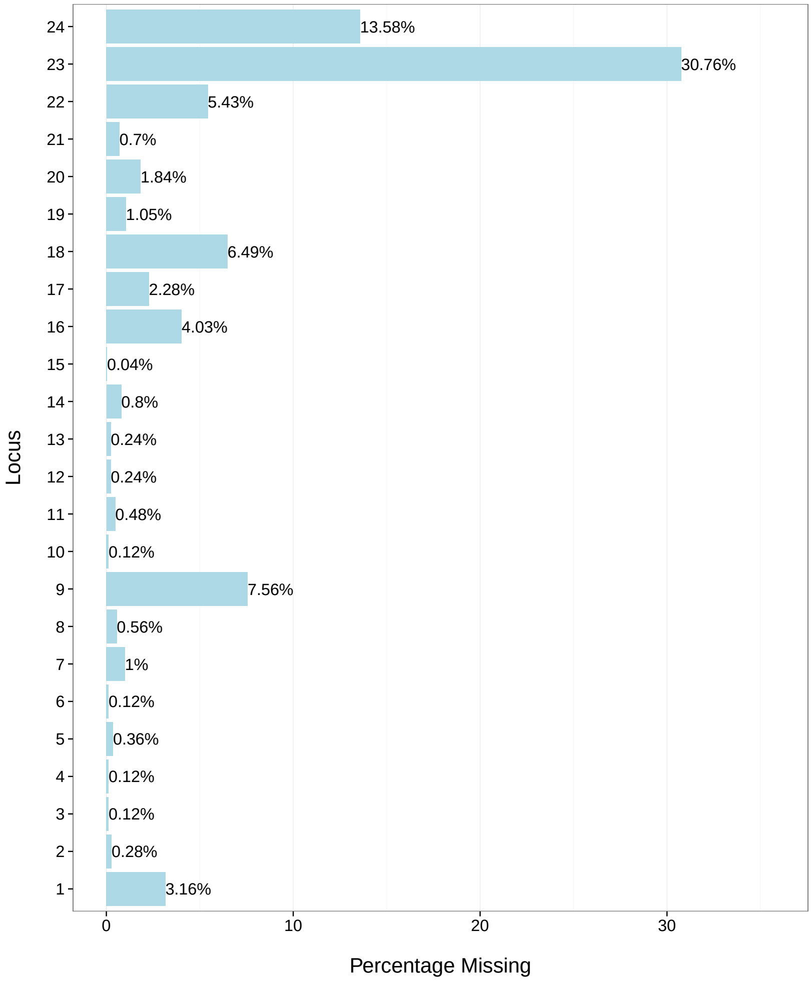


**Figure 2.8:** Percentages of missing numbers of tandem repeats in each locus.

and $\delta_{ijk} = 0$ when a comparison cannot be made (one or both profiles are missing a digit in locus position $k$). This effectively deals with profiles that have missing loci through the adjustment $s_{ijk} \times \delta_{ijk}$. As a simple example, consider a scenario where we have two profiles each genotyped by 15 loci:

$$\begin{array}{c} 323332532315323 \\ 3233325323\_5323 \end{array} \tag{2.1}$$

where _ denotes a missing locus, then $\delta_{ijk} = 0$ for $k = 11$ and $\delta_{ijk} = 1$ for the rest of the loci ($k \neq 11, k \leq 15$). Finally the dissimilarity between the pair of cases $i$

and $j$ is calculated as the average score taken over all possible comparisons.

$$s_{ij} = \frac{\sum_{k=1}^{K} s_{ijk} \times \delta_{ijk}}{\sum_{k=1}^{K} \delta_{ijk}} \tag{2.2}$$

So in fact, for the two profiles shown in (2.1), the dissimilarity score is $s_{ij} = \frac{0}{14} = 0$. We can clearly see that the purpose of the $\delta$ term is to fix the denominator as the number of non-missing loci, and we can also note that Equation (2.2) ensures that all scores are in the range 0–1, with a score of 1 meaning all digits are different between two profiles. Thus, the two profiles in the example are considered as not dissimilar at all and so potential clusters can be easily identified by looking at the incomplete MIRU profiles which produce dissimilarity scores of 0 with at least one complete MIRU profile. In our example, we would be particularly interested in discovering whether or not the missing digit matches the corresponding digit in the complete profile which is a 1. When incomplete profiles produce $s_{ij} > 0$ with all complete profiles, they do not form any potential clusters using this dissimilarity scoring system.

## 2.4.2 Probabilities for Potential Strain Clusters

One more important point to mention is that it is possible for incomplete profiles to belong to more than one potential cluster. In the following example, the incomplete profile shown on the second row could belong to two potential clusters:

$$\begin{array}{cc} 323332532315323 & 323332532325323 \\ 3233325323\_5323 & 3233325323\_5323 \end{array} \tag{2.3}$$

In this instance, we would be interested in finding out if the missing digit is a 1 or a 2. One way to discover what the missing digit is with, perhaps, a high degree of certainty is to repeat the genotyping procedure on the sample related to the incomplete profile. However, as an intermediate step, it would be beneficial to estimate a statistical probability that the incomplete profile belongs to either of these two clusters based on all available MIRU data (keeping in mind the possibility that it may not belong to any of these clusters). It is important to stress that these probabilities are only for exploratory purposes as certain assumptions need to be made in order to derive them. In essence, nothing absolutely definitive can come from these probabilities and epidemiological investigation is still necessary to determine *true* clusters. We propose two methods for estimating the probability.

### Method 1: Relative Frequency of Repeats in Each Locus

The first method we propose for calculating the probability that a missing locus is a particular digit is to use the relative frequency of a digit in a particular locus among all available MIRU profiles. If we denote $f_{dk}$ as the frequency of a specific number of tandem repeats $d$ in locus $k$, and $n_k$ as the total number of MIRU profiles with a non-missing digit in locus $k$, then the relative frequency is

$$r_{dk} = \frac{f_{dk}}{n_k} \tag{2.4}$$

The percentages of different numbers of tandem repeats in each locus ($r_{dk} \times 100\%$) are displayed in Figures 2.9 and 2.10. Note that much more data was available for the first 15 loci compared with loci 16–24 ($n$ = 2,499 for the first 15 loci and $n$ = 1,141 for loci 16–24). This is because we were able to use both, the 15 and 24 loci MIRU profiles when finding the percentages while for loci 16–24, we could obviously only use the 24 loci MIRU profiles.

From the figures, it is immediately apparent that certain loci more commonly have specific numbers of repeats. This finding can give us some degree of confidence in using relative frequencies to impute missing digits as the number of repeats in some loci seems to be reasonably predictable. For example, in loci 2, 6 and 9, there are very often two repeats, particularly for locus 6 where there were two repeats for approximately 95% of all the MIRU profiles. Furthermore, in loci 4, 13, 19 and 21, there are very often three repeats; this number of repeats occurred in over 80% of MIRU profiles for each of these loci. However for some loci such as locus

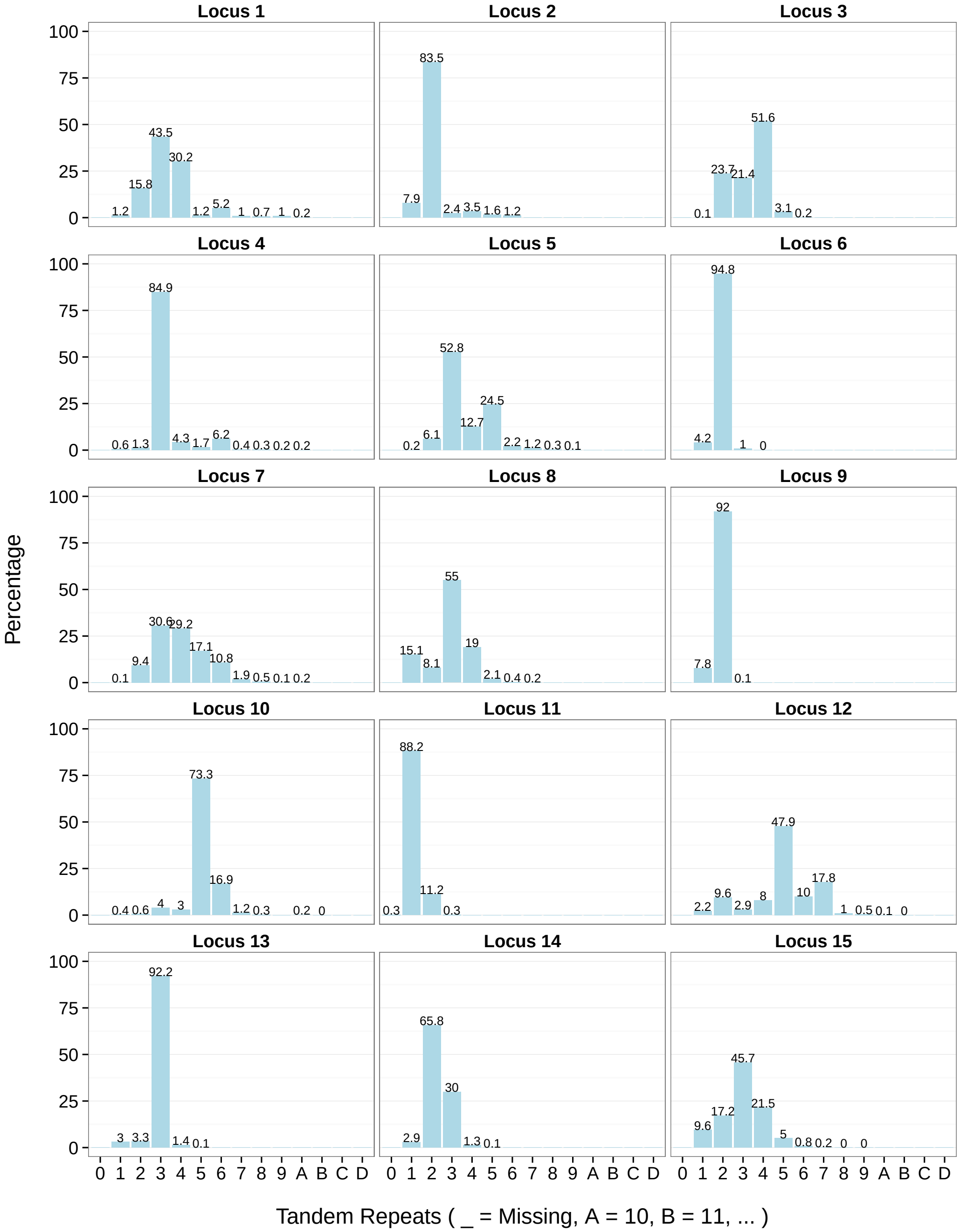


**Figure 2.9:** Percentages for numbers of tandem repeats in each locus from 1 to 15. The 15 and 24 loci MIRU profiles could be used here. Note that percentages have been rounded to one decimal place and are only displayed on the plot if the percentage is greater than zero.

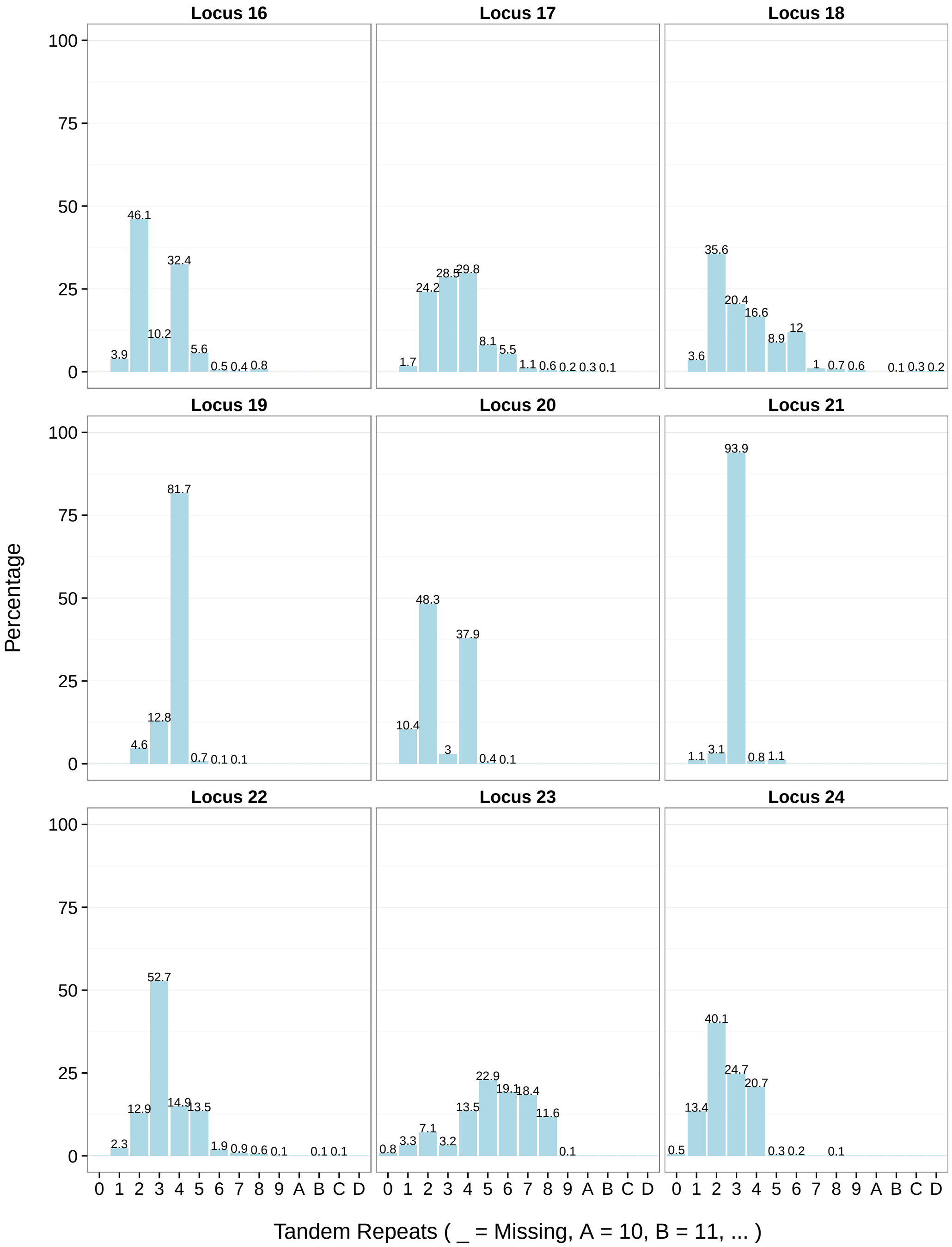


**Figure 2.10:** Percentages for numbers of tandem repeats in each locus from 16 to 24. Only 24 loci MIRU profiles could be used here. Note that percentages have been rounded to one decimal place and are only displayed on the plot if the percentage is greater than zero.

7, the number of repeats is less predictable (3 or 4 repeats is most likely but 2, 5 and 6 repeats also appear quite frequently).

Referring back to our example in Equation (2.3), we can obtain the relative frequency that the missing digit is a 1 or a 2, giving us probabilities using this method. Using the percentages from locus 11 in Figure 2.9 (and converting back to relative frequencies) gives $r_{1,11} = 0.88$ and $r_{2,11} = 0.11$ and hence, this method suggests that the missing digit in the example is far more likely to be a 1 rather than a 2.

When there are two missing loci (the maximum allowed for considering potential clusters in 24 loci profiles), we multiply together the relative frequencies for each missing locus to get a probability using this method. Multiplying the relative frequencies in this way assumes independence with regards to the *joint-missingness* between loci. We tested this assumption by performing a bootstrap test where random numbers of joint-missing loci were generated (while keeping the total number of joint-missing loci fixed to be the same as in the actual data), and then comparing the sum of the $\chi^2$ values for the actual data against the simulated data. The bootstrap test showed that the independence assumption is not valid here – largely due to the finding that most of the missing loci appear in loci 23 and 24 (see Figure 2.8) – but despite this, we will still use the method of multiplying the relative frequencies here when there are two missing loci as it provides a simple measure of probability for investigators to use, while bearing in mind that information derived from epidemiological investigations would take precedence anyway.

### Method 2: Incomplete Profile Must Belong to a Potential Cluster

Another method of calculating the probability that a missing locus is a specific digit is to assume that the incomplete profile must belong to an existing potential cluster. Essentially, this means that we assume that the missing digit can only be a digit from a potential cluster. If there are more MIRU profiles in potential clusters with digit $d$ in the missing locus position, then there is a higher probability of that missing digit being $d$. We can write this probability as

$$p_d = \frac{n_d}{n_{pc}} \tag{2.5}$$

where $n_d$ is the number of profiles which are in potential clusters with a digit $d$ and $n_{pc}$ is the total number of profiles in potential clusters. For the example from Equation (2.3), this method would assign a 50% chance of belonging to either of the potential clusters as there is only one case in each of the potential clusters.

### Comparison of Methods

As illustrated with the example given in Equation (2.3), the two probability methods can sometimes produce vastly different estimates. Since this can happen sometimes, we must consider the merits and shortcomings of using estimated probabilities from both methods for cluster investigations.

Method 1 is based more on the *big picture* regarding TB strains in Scotland as it looks at patterns in numbers of repeats for all of the MIRU data. Hence, Method 1 should work well when examining strains of TB that are common throughout Scotland or when the number of repeats in a locus is reasonably consistent for all strains. However, Method 1 is likely to be less useful when looking at missing numbers of repeats in loci which often have variable numbers of repeats; for instance, loci 7, 17 and 23 (see Figures 2.9 and 2.10), and this is likely to be an issue as locus 23 is missing in around 30% of profiles (Figure 2.8). Moreover, the probabilities generated using Method 1 when there are two missing loci may have questionable reliability as the method assumes independence between joint-missing loci while our tests showed this to be an invalid assumption.

In contrast to this, Method 2 focuses on much smaller subsets of MIRU profiles as it only compares incomplete profiles to other profiles that are almost exactly the same. Thus, it is likely to work better than Method 1 when looking at strains of TB that are rarer in Scotland and may also be more reliable when there are two missing loci. A disadvantage of Method 2 is that it disregards the possibility that cases belong to no existing cluster; with Method 1, this possibility is still considered and can be apparent when all the estimated probabilities are low.

Another issue to consider regarding the two methods is that the probability estimates could change over time as more TB case data becomes available. This problem is likely to affect Method 2 much more than Method 1. This is because Method 2 is based on smaller numbers and thus, will be less stable over time. However, the probability estimates for Method 1 will only be affected if there is a large influx of rare TB strains into Scotland.

### 2.4.3 Implementation in R

Since HPS receives updated data on TB cases in batches after some period of time (e.g. every quarter of a year), the functions written in R have been designed to take in batches of TB cases with MIRU profiles and find all incomplete profiles which form potential clusters. The R code for the functions is given in Appendix 2.A and the instructions for processing updates is also contained in the Appendix. Gower's dissimilarity coefficient was calculated using the `daisy()` function from the **cluster** package by Maechler et al. [232].

The first function `fun.initcompleteclus()` is to be run only once initially to find all the strain clusters using only cases with complete MIRU profiles; it does not need to be run when new batches of cases are received and is used just to create a starting cluster dataset. This function has been designed in this way as at the beginning we can find clusters very simply by just tabulating the MIRU profiles and seeing which profiles have frequency greater than one. Each cluster is given a cluster ID which is "SC" followed by a four digit number. So for instance, the first cluster is named as "SC0001". Note that these cluster IDs are fixed and static over time. Once the complete strain clusters have been dealt with, the process becomes more complicated.

All updating of clusters for new TB cases is taken care of by the `fun.clusupdate()` which calls upon two sub-functions named `fun.update.completeclus()` and `fun.`

assignmissclus(). The first of these examines any new cases with complete MIRU profiles for assignment to clusters while the latter function only looks at the incomplete MIRU profiles for the same purpose. If a new batch of cases all have complete MIRU profiles, then only the first sub-function is run by fun.clusupdate(), and conversely, if all cases have incomplete profiles, then only the second sub-function is run. For every new case, there are three possibilities:

1. Assignment to an existing cluster.
2. Assignment to a new cluster.
3. No assignment to any cluster.

The fun.update.completeclus() sub-function achieves assignment to clusters by using tabulations with appropriate use of subsetting operations which takes very little computational effort. However, the fun.assignmissclus() sub-function uses a 'for loop' where there is a cycle for each incomplete MIRU profile fed into fun.clusupdate(), which is more computationally demanding. This is not a problem, provided, there are not huge amounts of TB cases (for example, hundreds of thousands of cases).

In each cycle, one incomplete profile is considered and is compared against all complete MIRU profiles. This is done by creating a temporary dataset with the incomplete profile as the first item and then appending all complete profiles after that. In this temporary dataset, no other incomplete profiles are included. The comparisons are made by calculating the dissimilarity matrix using the daisy() function.

This gives a matrix $\mathbf{S}$, that is symmetric with $s_{ij} = s_{ji}$ and since all comparisons made against the same profile give a dissimilarity of 0, we also have $s_{ii} = 0$. As we only want comparisons between the incomplete profile against all other complete profiles, we only require the first row of the matrix (or equivalently, first column since the matrix is symmetric), excluding the first element $s_{1,1}$ which is the comparison with itself. From here, if we take the first row of the matrix, we can easily find potential strain clusters by examining the profiles with index $j$ in that first row which result in $s_{1,j} = 0$. Note that most of the complexity in the R code seen in Appendix 2.A is in collating which cases belong to which clusters, and to ensure that existing cluster IDs are retained after every update.

The last function, `fun.clusprob()` is used to calculate the probabilities that incomplete profiles belong to potential clusters using the two methods we described previously in Section 2.4.2. Note that this function also requires data relating to the relative frequencies of numbers of repeats for each loci. The final output tables are viewable as a spreadsheet in Microsoft Excel and this was achieved by using the **XLConnect** package in R [134]. Using this package allows the spreadsheet to be updated with an R script and thus, automates this part of the process.

### Using Results for Cluster Investigations

An example of the output produced after running the R cluster functions that could be useful for cluster investigations is shown in Table 2.4. There we can observe three incomplete MIRU profiles that form potential clusters. Looking more closely at the first incomplete profile shown on the table, we can see that it has formed two potential clusters which are clusters SC0047 and SC0069. Note that here, cluster SC0047 would have already existed and was made up of two cases with complete MIRU profiles which are identical, while SC0069 is a new potential cluster that is formed after joining with the incomplete profile. If it belonged to SC0047, we would impute the missing digits in locus 23 and locus 24 as 8 and 2 respectively and if it belonged to SC0069, we would impute them as 6 and 2.

Based on probability method 1, it is slightly more likely to belong to SC0069 than SC0047 but since both probabilities are low, this gives a reminder that we must keep in mind the possibility that it may not belong to any of these clusters. Probability method 2 gives the opposite recommendation – the incomplete profile is more likely to belong to SC0047 as it contains two cases compared with SC0047 which contains just one case. Another important consideration for this case is the epidemiological information, which perhaps also suggests that it may belong to none of the clusters as the case is from Grampian while the other cases are from

| MIRU | Clus ID | Clus Poss | Health-board | Gender | Age Group | Disease Type | Year Notified | Prob Method 1 | Prob Method 2 |
|---|---|---|---|---|---|---|---|---|---|
| 424352332515333456443 3__ | ASSIGN | 47, 69 | Grampian | F | 15-34 | P | 2007 | | |
| 424352332515333456443382 | SC0047 | | Lothian | F | 35-64 | NP | 2012 | 0.05 | 0.67 |
| 424352332515333456443382 | SC0047 | | GGC | M | 35-64 | P | 2012 | 0.05 | 0.67 |
| 424352332515333456443362 | SC0069 | | GGC | M | 35-64 | P | 2011 | 0.08 | 0.33 |
| 42435233_5153334564433_2 | ASSIGN | 47, 69 | GGC | M | 15-34 | P | 2008 | | |
| 424352332515333456443382 | SC0047 | | Lothian | F | 35-64 | NP | 2012 | 0.11 | 0.67 |
| 424352332515333456443382 | SC0047 | | GGC | M | 35-64 | P | 2012 | 0.11 | 0.67 |
| 424352332515333456443362 | SC0069 | | GGC | M | 35-64 | P | 2011 | 0.18 | 0.33 |
| 42235254251733354242 33_4 | ASSIGN | 43, 99 | GGC | M | 15-34 | NP | 2010 | | |
| 422352542517333542423354 | SC0043 | | Lothian | F | 65+ | P | 2010 | 0.23 | 0.67 |
| 422352542517333542423354 | SC0043 | | Lothian | M | 35-64 | NP | 2012 | 0.23 | 0.67 |
| 422352542517333542423384 | SC0099 | | Fife | M | 15-34 | P | 2013 | 0.12 | 0.33 |

**Table 2.4:** Example of TB cluster investigation output after running the cluster functions in R. The output shows three incomplete MIRU profiles that form potential clusters. The probability that they belong to certain potential clusters based on the two methods outlined in Section 2.4.2 are given in the last two columns of the table. Epidemiological investigation would also be needed to gain clarity on whether or not cases do actually belong to clusters.

healthboards which are relatively far away (Lothian and GGC). As well as this, the year of notification for the case was in 2007 compared with the other cases which have years of notification in 2011 and 2012.

For the third incomplete MIRU profile displayed in Table 2.4, there is seemingly less ambiguity for the investigation. Both probability methods suggest that it is more likely to belong to SC0043 than SC0099. Moreover, the case with the incomplete profile has a year of notification which also matches closely to one of the cases from SC0043. Therefore, the data brings a fairly strong argument for this case belonging to SC0043, but further investigation would still be required to confirm this conclusion. This example gives a clear demonstration of where the system can be used to guide a follow-up TB cluster investigation.

### Summary Results after Running R Functions

An overview of how many potential clusters were formed by each incomplete MIRU profile is shown in Table 2.5. Out of the 150 MIRU profiles with 15 loci and one missing locus, 59 formed one potential cluster, 25 formed two potential clusters and 66 incomplete profiles were not sufficiently similar to any complete MIRU profiles to form potential clusters. From the 496 MIRU profiles with 24 loci and up to two missing loci, 141 formed one potential cluster with one complete profile, 25 formed two potential clusters, 5 formed four potential clusters, but 325 did not form any potential clusters. These results imply that a large number of the cases could be isolated cases – 44% of cases with incomplete 15 loci profiles and 65.5% of cases with incomplete 24 loci profiles.

## 2.5 TB cases in 2012–13

The most recent TB cases are of considerable interest as we wish to discover and prevent any new or recent outbreaks as soon as possible. If outbreaks are not found quickly, then there is more potential for transmission of infection to others and the effort to stem the outbreak may be greatly increased. Therefore, using a subset of the most recent cases in the dataset, we further explored the patterns of trans-

| | **15 Loci** | | **24 Loci** | |
|---|---|---|---|---|
| | ***n*** | **%** | ***n*** | **%** |
| Incomplete Profiles Considered | 150 | | 496 | |
| *Potential Clusters Formed* | | | | |
| 0 | 66 | 44.0 | 325 | 65.5 |
| 1 | 59 | 39.3 | 141 | 28.4 |
| 2 | 25 | 16.7 | 25 | 5.0 |
| 3 | 0 | 0.0 | 0 | 0.0 |
| 4 | 0 | 0.0 | 5 | 1.0 |

**Table 2.5:** Numbers of potential clusters formed by cases with incomplete MIRU profiles.

mission. We selected those cases with 24 digit MIRU profiles with a maximum of two missing loci occurring within a year of the most recent case notification in the data. This is made up of 165 cases with date of notification between 30th July 2012 up to 30th July 2013. Using multivariate methods such as multidimensional scaling (MDS) can allow us to see an overview of how similar the strains of TB are between cases over this period. As mentioned previously in Section 2.1.4, it is possible for linked TB cases to have a single digit difference in their MIRU profile and we can also investigate this possibility here.

We can again use Gower's distance here but compared with previously, we will treat the number of repeats in each locus as quantitative variables. When treating all the variables as quantitative variables, Gower's distance essentially becomes a *range-normalised Manhattan distance* [136]. This looks at the absolute difference in the numbers of repeats between the MIRU profiles of cases and then normalises the distance using the range of variables so that distances will always be between 0–1. Calculating distances in this way allows us to explore the effect of giving some importance to the magnitude of the number of repeats. For instance, in the following example, the 2nd case is treated as being *more similar* to the 1st case than the 3rd case (in the first locus, '2' is closer to '1' than '3'; all other loci are identical between the 3 cases).

123456781234567812345678
223456781234567812345678
323456781234567812345678

When using Gower's distance, the missing loci are normally not considered in comparisons, and so if a case with one missing locus has 23 identical digits to a case with no missing loci, then they would be considered to be identical. To avoid this problem in this analysis, we have imputed the missing digits using probability method 2 where possible (using method 2 explicitly depends on the existence of cases with highly similar MIRU profiles to incomplete profiles), and where it was not possible to use method 2, we have used probability method 1 for imputation (method 1 can always be used for imputation).

### 2.5.1 Multidimensional Scaling

One powerful method of visually representing these dissimilarities in a spatial map is to use a dimension-reduction technique such as MDS [39] or more specifically here, classical MDS. This method essentially allows us to estimate coordinate values in, for example, 2D space, where the *interpoint distances* or distance between any two points (which here represent individual TB cases) is determined by the dissimilarity between them. In practice, the distances between all sets of points may not always be approximately proportional to their dissimilarities but the majority should be when the method performs well. The axes that these points lie on are artificially created and are sometimes called the *principal axes* [139]. These axes are artificial in the sense that they are not directly related to anything that is observed but are in fact, derived from the data with a goal of explaining as much of the variation in the data as possible.

Details on the optimisation schemes and goodness-of-fit indices used to estimate MDS coordinates can be found in Everitt et al. [106], but we will not discuss them here as we want to concentrate primarily on interpreting the results from the MDS solution. Firstly, we can informally test the adequacy of representing the data in two dimensions. Due to the presence of negative eigenvalues from the MDS solution and because the dissimilarity matrix is not Euclidean, we tested the adequacy by using a criteria suggested by Mardia et al. [235]. This makes use of squared eigenvalues rather than the raw eigenvalues.

$$\frac{\sum_{i=1}^{2} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2} \tag{2.6}$$

Here, $\lambda_i$ denotes the $i$th eigenvalue obtained from the MDS procedure and note that the numerator uses the two largest positive eigenvalues. Calculating this for our data gives a value of approximately 0.79. It has been suggested before that a value of around 0.8 is a reasonable fit [176] and so our result indicates that using two dimensions to represent the dissimilarities is probably justifiable. Figure 2.11 shows several plots which are 2D representations of the dissimilarities between the subset of recent TB cases after MDS. Effort has been made to distinguish those cases with missing loci but have now had data imputed as there is more uncertainty around them.

The plots illustrate that there are perhaps, three main clusters of cases – the largest cluster on the top-left; another sizeable cluster on the top-right; and a small cluster on the bottom-left. In the first plot of Figure 2.11 (top-left), colour is used to distinguish the organism of the case. We can observe that there is a lot of overlap between cases categorised as “M. Tuberculosis” and “M. Tuberculosis Complex” as we would expect, but three out of the four “M. Bovis” cases are reasonably distinctly separated from the main organisms. However, as there is a fair amount of distance between the bovine TB cases, there must be some noticeable differences in their MIRU profiles and it is unlikely that they form part of the same transmission chain. There also appears to be one case (located slightly below and left of centre) that is fairly isolated from all other cases. This is a case with MIRU profile 644652432102233327231346 that has number of differing loci ranging from 13 to 22 (median 18 differing loci) with the MIRU profiles of all other cases in the 12–13 period under examination. Therefore this case appears to be quite geneti-

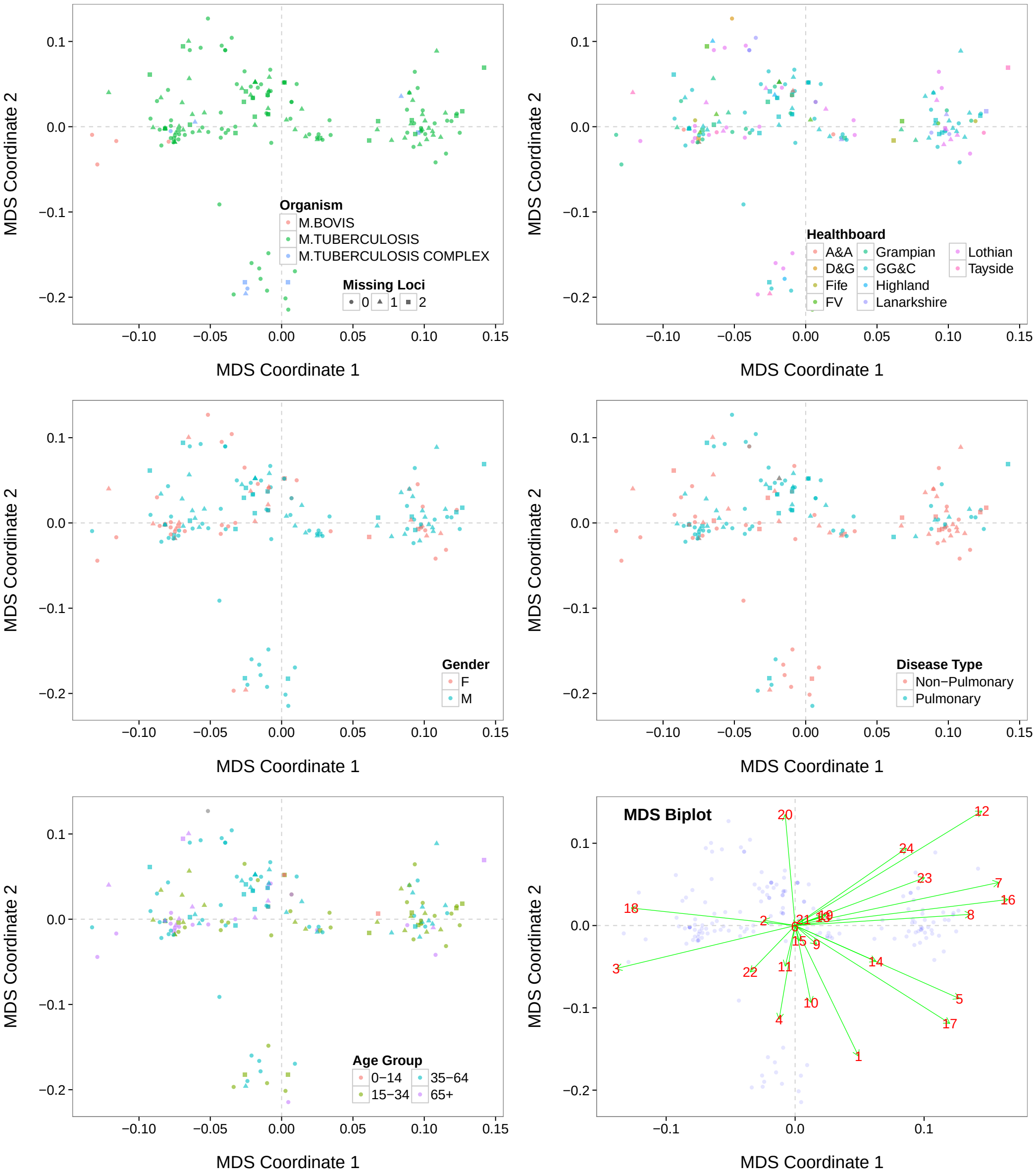


**Figure 2.11:** TB cases from 30th July 2012 to 30th July 2013 represented visually in two dimensions by MDS. Dissimilarities were estimated using Gower's distance. Plots show colours by organism, healthboard, gender, disease type and age group. The shape of the points distinguish cases with complete and incomplete MIRU profiles (legend only shown on top-left plot). Note that the `jitter` function in R was used to distinguish cases with identical MIRU profiles that would otherwise be laid directly on top of each other. The regression biplot (bottom-right) gives some idea about the loci that influence cases to be located where they are on the plots. Note that there is a slight change in scale on the $y$-axis on the biplot as compared with the other plots.

cally different from the other cases. The other plots on Figure 2.11 show no clear patterns by healthboard (top-right plot) or by age group (bottom-left plot). Most of the cases in the bottom cluster are males and many of them are non-pulmonary cases (centre two plots).

An MDS biplot can give an idea about how each of the loci relate to the MDS map. It was created by following a procedure as given in Greenacre et al. [139]. The direction and magnitude of the lines (biplot vectors) describe how influential each of the loci are in *pulling* points to a certain position on the plot. If the biplot vector for a particular locus has greater magnitude, then there will be more variance in that locus. The biplot vectors were found by a series of 24 linear regressions, one for each locus. In each regression, the response was the number of repeats in a locus and there were always the same two covariates which were the the $x$ and $y$ coordinates found from the MDS procedure. As the MDS dimensions are centred (the mean of each MDS dimension is zero), the constants from these regressions are the mean number of repeats for each of these loci which will be situated at the origin on the plot. Hence, all biplot vectors begin at the origin and the coefficients derived from the regressions for the MDS $x$ and $y$ coordinates give the endpoints for the vectors. The results from the regressions are given in Table 2.6 and these are subsequently superimposed onto the MDS plot in the bottom-right of Figure 2.11. Note that in the plot, the lines have been multiplied by an arbitrary scalar value to help separate the lines out more.

From the biplot, we can see that differences in loci 3 and 18 are important for cases being situated in the top-left. Similarly, differences in loci 7, 8, 12, 16 and 23 seem to have a bearing on cases being located towards the right side. Finally, differences in loci 1 and 4 appear to have most influence on the bottom cluster of cases. The various loci seen near the origin of the biplot have little influence as most of the cases likely have a similar number of repeats in these loci. As a *sense check*, we can calculate the variances in each of the loci and compare these against the length of the vectors; these are seen in the last two columns of Table 2.6. It should be easier to separate out cases when there is higher variance in some loci and this should be reflected in the length of the biplot vectors. In general, there is good agreement as loci 2, 6, 9, 13 and 21 all have low variance but there are some exceptions such as locus 15, which has moderate variance but is situated near the origin in the biplot. Some differences such as this are to be expected as the MDS solution will not perfectly represent the structure of the dissimilarity matrix in two dimensions. On top of this, the $R^2$ values show that some of the regressions used to derive the biplot vectors do not give a good fit to the data. Moderately high $R^2$ values of around 0.45 and above were achieved for most of the loci with higher variance but the loci with low $R^2$ values show the limitations in what can be interpreted from the biplot alone.

### Minimum Spanning Tree

One useful way of identifying possible misrepresentations of the dissimilarities in the MDS solution is to produce the *minimum spanning tree* (MST). In essence, this is a spanning tree which connects up all pairs of TB cases according to certain criteria. These are that every point (or case) must be visited at least once, the entire tree must be connected meaning that there must be some path between any pair of points, and finally, closed loops are not permitted. In total, there are $\binom{n}{2}$ straight line segments connecting up pairs of cases and the length of a spanning tree is the sum of the length of these segments; the MST is the tree which minimises the total length of the tree. Further details on algorithms which find the MST can be found in Gower and Ross [137].

Figure 2.12 displays the same graph as earlier from the MDS solution but with the line segments of the MST included. On the whole, the MDS procedure seems to have worked well as most of the links in the MST connect pairs of cases close to each other. Cases which are close together on the plot and connected in the MST may be interesting to investigate further as their MIRU profiles likely only differ slightly. In fact, a system for finding possible transmission pairs could be created by looking at those points which are connected by a segment in the MST as well as being within some threshold distance of each other on the MDS plot. As the likelihood of there being a transmission pair is small when the two cases have

| **Locus** | **Constant** | **Coeff MDS Coord 1** | **Coeff MDS Coord 2** | $R^2$ | **Vector Length** | **Var** |
|---:|---:|---:|---:|---:|---:|---:|
| 1 | 3.53 | 3.938 | -12.697 | 0.451 | 13.294 | 1.58 |
| 2 | 2.05 | -1.963 | 0.521 | 0.050 | 2.031 | 0.39 |
| 3 | 3.23 | -11.102 | -4.184 | 0.733 | 11.864 | 0.91 |
| 4 | 3.24 | -0.991 | -9.108 | 0.494 | 9.162 | 0.67 |
| 5 | 3.76 | 10.193 | -7.097 | 0.654 | 12.420 | 1.08 |
| 6 | 1.98 | -0.016 | -0.074 | 0.001 | 0.076 | 0.04 |
| 7 | 4.17 | 12.635 | 4.188 | 0.534 | 13.311 | 1.58 |
| 8 | 2.97 | 10.901 | 1.093 | 0.521 | 10.955 | 1.11 |
| 9 | 1.90 | 1.337 | -1.815 | 0.246 | 2.254 | 0.09 |
| 10 | 4.93 | 0.984 | -7.484 | 0.313 | 7.548 | 0.72 |
| 11 | 1.10 | -0.614 | -3.957 | 0.543 | 4.004 | 0.12 |
| 12 | 5.18 | 11.583 | 11.132 | 0.486 | 16.065 | 2.34 |
| 13 | 2.93 | 1.716 | 0.840 | 0.090 | 1.911 | 0.19 |
| 14 | 2.36 | 5.012 | -3.478 | 0.495 | 6.100 | 0.34 |
| 15 | 3.10 | 0.257 | -1.479 | 0.009 | 1.501 | 0.95 |
| 16 | 2.97 | 13.203 | 2.531 | 0.527 | 13.443 | 1.65 |
| 17 | 3.40 | 9.598 | -9.516 | 0.537 | 13.516 | 1.50 |
| 18 | 3.28 | -10.169 | 1.706 | 0.258 | 10.311 | 1.99 |
| 19 | 3.74 | 1.903 | 1.012 | 0.051 | 2.155 | 0.43 |
| 20 | 2.70 | -0.618 | 10.794 | 0.382 | 10.812 | 1.21 |
| 21 | 2.98 | 0.517 | 0.580 | 0.022 | 0.777 | 0.12 |
| 22 | 3.33 | -2.785 | -4.518 | 0.099 | 5.308 | 1.20 |
| 23 | 5.60 | 8.030 | 4.648 | 0.202 | 9.279 | 1.97 |
| 24 | 2.54 | 6.912 | 7.527 | 0.387 | 10.219 | 1.18 |

**Table 2.6:** Results from linear regressions used to find biplot vectors. The constant gives the mean number of repeats in each locus, and the two coefficients for the MDS coordinates give the direction and magnitude for the vectors which begin at the origin on the MDS plot. The variance column was calculated outwith the regression.

non-identical MIRU profiles, then it would probably be best to choose a restrictive threshold distance so that relatively few transmission pairs are considered. Note that these could be looked at by recreating the plot (and making it larger) with ID or row numbers for the cases so they can be easily identified rather than points, but we have produced the plot here with points to prevent it from being overly cluttered.

It is also possible to see that although some points are close on the graph, they

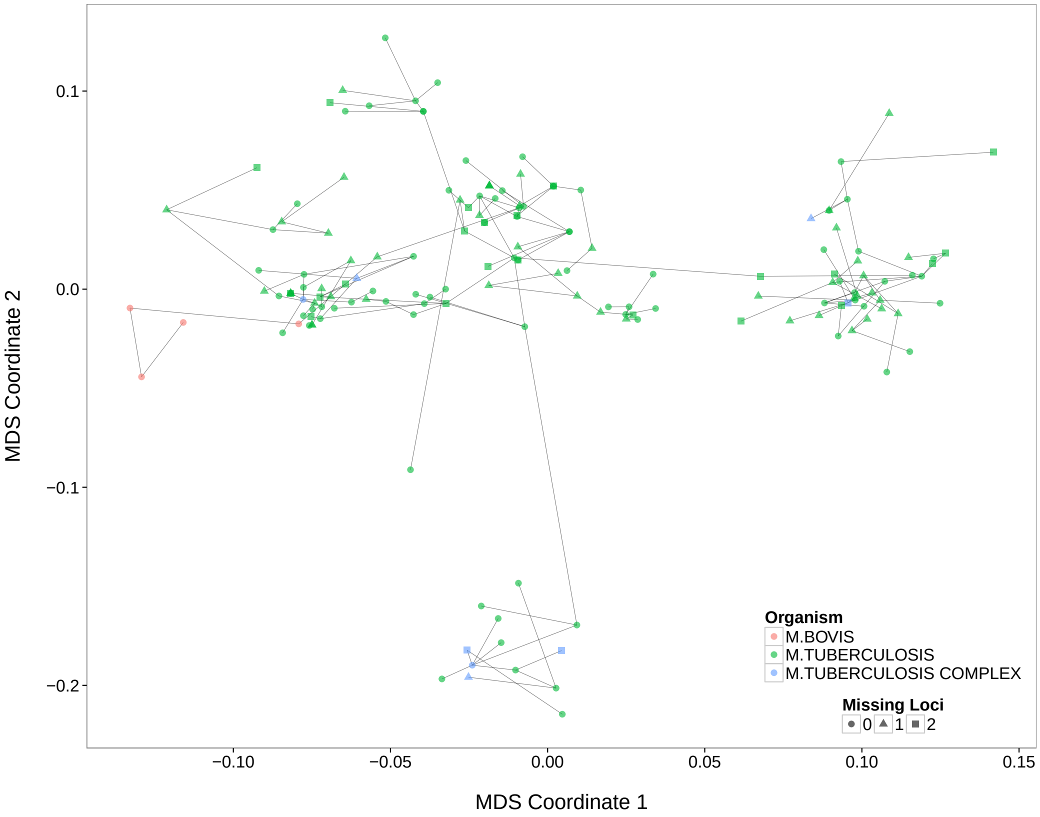


**Figure 2.12:** Minimum spanning tree imposed onto the MDS solution for TB cases from 30th July 2012 to 30th July 2013. Note that the `jitter` function in R was used to distinguish cases with identical MIRU profiles that would otherwise be laid directly on top of each other.

are not always connected with a line segment. When this happens, it is a sign of a possible distortion in the MDS solution whereby, the MDS solution does not accurately reflect the level of dissimilarity calculated earlier by Gower's distance. Some of the results from the MST make good practical sense. For example, the bovine TB cases are all linked by a line segment – even when there are *M. tuberculosis* cases situated nearer. Again, it should be noted that care should be taken when examining any links for those cases that have had data imputed due to the additional uncertainty associated with them.

Using the connected pairs of cases from the MST, we can further investigate those pairs which have MIRU profiles that differ in only one locus after data has been imputed. A summary of the 12 pairs of cases that fit this criteria can be seen in

Table 2.7, where the imputed digits in MIRU profiles have been underlined and the differing digit between pairs of MIRU profiles has been made bold for clarity. Since the connection between these pairs of cases is fairly speculative due to the differing locus, it is probably better to focus on pairs of cases that have a very short distance between them (relative to the scale of the distance measure used) and also those with no data imputations (or perhaps only one imputed digit). For example, the best candidates for investigation from Table 2.7 would be the 8th and 12th pairs on the list as no data imputations have been made in their MIRU profiles and the distance between these pairs on the MST is relatively short. Another strong candidate for investigation would be the 4th pair of cases where only one digit has been imputed in the MIRU profile of case B and the distance between them on the MST is the shortest out of all the pairs considered in the table.

| **MIRU Case A** | **MIRU Case B** | **Distance** | **Digits Imputed Case A** | **Digits Imputed Case B** |
|---|---|---|---|---|
| 22433231251532**4**224423552 | 22433231$\underline{2}$51532**5**22442355$\underline{2}$ | 0.0083 | 0 | 2 |
| 42443333$\underline{2}$5153222154233**7**$\underline{2}$ | 4244333325153222154233$\underline{\mathbf{5}}\underline{2}$ | 0.0119 | 2 | 2 |
| 4223525425173334424232**8**$\underline{2}$ | 4223525425173334424232**7**$\underline{2}$ | 0.0060 | 1 | 1 |
| **2**24332312515324224423552 | **3**2433231$\underline{2}$515324224423552 | 0.0052 | 0 | 1 |
| 422352542517333442423**2**7$\underline{2}$ | 42235254$\underline{2}$517333442423**3**7$\underline{2}$ | 0.0069 | 1 | 2 |
| 3233324325153254334433**8**3 | 3233324325153254334433$\underline{\mathbf{5}}$3 | 0.0179 | 0 | 1 |
| 42234264251732344244347$\underline{\mathbf{2}}$ | 42234264251732344244347**4** | 0.0104 | 1 | 0 |
| 4223426425173234424434**7**4 | 4223426425173234424434**6**4 | 0.0060 | 0 | 0 |
| 32433231$\underline{2}$51532422442355**2** | 32433231251532422442355**1** | 0.0052 | 1 | 0 |
| 4223526425173344424233$\underline{5}$**4** | 4223526425173344424233$\underline{5}$**1** | 0.0156 | 1 | 1 |
| 4243523325173334564433$\underline{\mathbf{5}}\underline{2}$ | 424352332517333456$\underline{4}$433**4**2 | 0.0060 | 2 | 1 |
| 61464243272233426**3**313441 | 61464243272233426**2**313441 | 0.0060 | 0 | 0 |

**Table 2.7:** Pairs of TB cases connected by the minimum spanning tree in 2012–13 that have MIRU profiles with differences in only one locus after data imputations. The imputed digits have been underlined and the differing digit between pairs of profiles has been made bold for clarity. Gower's distance, treating numbers of repeats in each locus as quantitative variables, was used as the distance measure.

### 2.5.2 Sensitivity Analysis

We can test the robustness of the classical MDS result by using *non-metric MDS* while using the same distance metric. The main difference between classical MDS and non-metric MDS is that the latter uses the rank order for dissimilarities rather than a direct numerical comparison. This is particularly applicable when we only know about a measure on an ordinal scale – for instance, survey results which were answered on a likert scale where we know a response of “strongly agree” is of higher order than “agree” but we cannot accurately quantify the extent of the increase in agreement.

For our TB data, we can postulate that when comparing the MIRU profiles of cases, a larger difference in the number of repeats in a locus may indicate a greater difference between their TB strains, but we do not know exactly how large the increase in dissimilarity is. Hence, there may be some merit in treating the dissimilarities in an ordered fashion. Details on the algorithms used in the non-metric MDS procedure to find the required coordinates in a plot can be found in Kruskal [211]. The basic idea is that a *stress* function is minimised after a number of iterations, at least to the point where some convergence criterion is reached.

An informal method for assessing a non-metric MDS solution is a *Shepard diagram*, which can be seen at the top of Figure 2.13. In a perfect scenario, the points would lie on the bisecting line but on the graph there are a number of points positioned further from the line towards the lower-right, and this suggests that the procedure has only worked moderately well. The other two plots displayed in Figure 2.13 are the biplot and MST. Overall, these plots look almost identical to the corresponding plots found from classical MDS. The biplot shows that the same loci as previously found are influential in the eventual location of cases on the plot and the MST exhibits similar patterns as found before. The results found using non-metric MDS have been encouraging in the sense that they largely agree with what was found before. This hints at some level of robustness in our results against the type of MDS procedure used as well as the distance measure used.

### Other Distance Measures

Keeping the distance measure the same between using classical MDS and non-metric MDS allowed a comparison to be made between the methods, but as part of the sensitivity analysis, we also tried using different distance measures for MDS

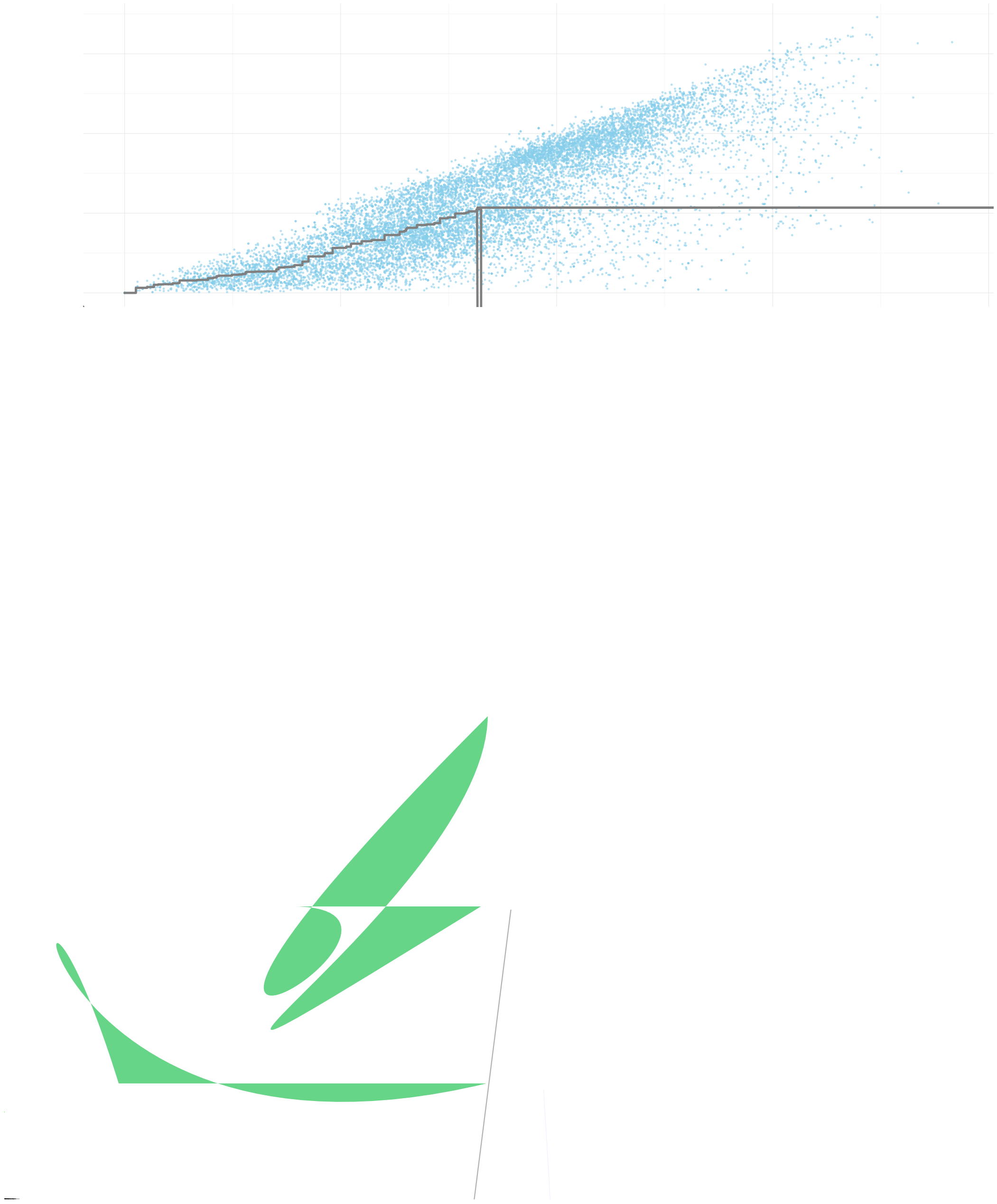


**Figure 2.13:** Non-metric MDS plots. Dissimilarities were estimated using Gower's distance. The shephard diagram (top) can be used to informally assess the quality of non-metric MDS solution. Note that the `jitter` function in R was used in the biplot and the minimum spanning tree to distinguish cases with identical MIRU profiles that would otherwise be laid directly on top of each other.

and non-metric MDS such as Euclidean distance and also simply using the number of differing loci between MIRU profiles. Both of these measures produced very similar results to what we already found; the main noticeable change was that when using the number of differing loci as the distance measure, all four bovine TB cases were more distinctly separated from the *M. Tuberculosis* cases.

## 2.6 Discussion

Disease outbreaks must be detected with haste so that effective interventions can be taken if necessary. Crucial to this process is the rapid detection of potential clusters of cases. If this can be achieved, then the effort and resources required to prevent further disease transmission can be greatly reduced. Here, we have attempted to design a number of bespoke systems which use statistical methods to assist in the detection of possible clusters of TB cases in Scotland. Although the annual incidence of TB in Scotland is relatively low as evidenced by our exploratory analysis of cases from 2000–2013 and HPS reports [162], there has been a recent rise in non-pulmonary cases and the Scottish Government wishes to act swiftly in order to reduce the burden on health caused by TB [330].

One major issue that HPS encountered in their investigations of TB cases was that they may have been failing to detect some strain clusters. This is due to the presence of missing loci in cases' MIRU profiles, which can happen as the genetic typing procedure does not always work perfectly. Our findings here show that this problem has increased since the introduction of using a 24-locus VNTR-MIRU panel compared with the previously utilised 15-locus panel. While the 24-locus panel offers greater discriminatory power than a smaller panel of VNTRs [65], we found that under half of 24-locus MIRU profiles were complete compared with almost 90% of 15-locus MIRU profiles being complete. Hence, a system was needed which could alert them to scenarios where the incomplete MIRU profiles are identical to other complete MIRU profiles, with the exception of the missing loci – these are prime suspects for potential clusters. Almost 99% of 15-locus MIRU profiles had one or zero missing loci and over 90% of 24-locus MIRU profiles had at most, two missing loci. For these reasons, it was decided that a maximum of one missing locus be allowed for 15-locus MIRU profiles and two missing missing loci be allowed for the now universally used 24-locus MIRU profiles. Furthermore, these rules for identification of potential clusters are in line with what has been used by Public Health England [296]. If we were to have allowed up to three missing loci in 24-locus profiles, then another 7% of the data could have been examined when finding potential clusters. However, it is likely that far too many potential clusters would be created and this would be counterproductive in investigations.

The system was developed in R [302] and can take in batches of TB cases with MIRU profiles, even when there are a mixture of complete and incomplete MIRU profiles. It has to be run separately for cases with 15 digit and 24 digit profiles, but going forward, this point is irrelevant as new cases will all have 24 digit profiles. HPS can run the program whenever they receive new data on TB cases (quarterly throughout the year) and this ensures that they minimise the chance of missing potential strain clusters. In addition to this, time and resources can be saved as the genotyping procedure can be repeated only when it is deemed that there is sufficient reason to do so.

For cases with incomplete MIRU profiles that were part of multiple potential clusters, we tried to estimate probabilities to describe the chance that they belonged to each of those possible clusters. It has to be stressed that the two methods used to assign probabilities are mainly for preliminary investigation purposes. Other qualitative information regarding the cases that is known to the investigator would be superior to the derived probabilities – for example, if cases were known to have come into contact with each other at times when one of them was infectious. In addition to this, other data that is routinely collected on cases such as date of notification, age group and healthboard should also be considered in investigations.

One of our methods for assigning probabilities used the relative frequency of the number of repeats in each of the 24 loci over all the available data. When there are cases with two missing loci, we derived a probability simply by multiplying relative frequencies together which assumes independence in joint-missing loci. We found that the independence assumption is probably not valid as most of the missing loci in our data appeared in loci 23 and 24. Hence a better probability estimate could be obtained by calculating probabilities conditional on exactly which two loci are missing.

In our exploratory analysis of the TB data, we found that the majority of the TB cases in Scotland were male and this finding is consistent with other studies [41, 249]. This adds evidence in favour for gender-focused interventions when tackling TB transmission. There are several possible explanations for why we are observing more males cases. Khan et al. [205] found that in Rawalpindi, women were less likely to test positive in smear tests for TB, and the reason for this was that women there were submitting poorer quality spatum samples compared with men. However this finding may only apply to low-income countries and may not be an explanation for gender disparities in developed countries. A possible explanation for the disparity which may be more generalisable worldwide could be to do with the impact that smoking has on TB. There has long been an association between smoking and pulmonary TB [40] and higher rates of smoking in men could be an important reason for observing more male cases [378].

Certain age groups also appear to be more at risk of TB. Compared with the age-structure of the population, we found a greater than expected proportion of TB cases were aged 15–34 and 65+. The finding that a large proportion of cases are in the elderly has also been seen in Germany [154]; in fact, that study found that the proportion of TB cases aged 60+ has been steadily increasing since 1976. For controlling TB, it is important to target interventions at this group as the study in Germany also found lower rates of treatment success in the elderly and higher rates of mortality. Moreover, the elderly could potentially seed outbreaks to other highly susceptible individuals in, for example, nursing homes.

There were far more cases in the GGC healthboard, relative to the size of population within GGC. There are many reasons why this finding may not be wholly unexpected. In Scotland, Glasgow has the highest population density and contains numerous areas with high deprivation. Hence, we are likely to see more cases in Glasgow as there is more crowding as well as higher rates of poverty there, and the link between poverty and TB has been found before, for instance in the US [32]. In addition to this, groups of people in Glasgow are known to have poorer health and shorter life expectancies compared to elsewhere in Scotland [138]. Therefore there are more likely to be higher proportions of people in Glasgow with weakened immune systems and perhaps, alcohol and drug misuse problems compared with other parts of Scotland. These factors are crucial as there is plenty of evidence which indicates that heavy alcohol use is a risk factor for incidence and re-infection of TB [305].

We also explored the use of MDS methods for finding clusters within the July 2012 to July 2013 period. The reason that we used only a subset of the cases in the MDS analysis was because it becomes difficult to interpret results and spot patterns when there are large numbers of cases. This is because the results from the MDS are displayed visually in two (or three) dimensions and the results are then interpreted with a degree of subjective impression by the reader. Using too many data points makes this task very challenging. Therefore, it was best to look at a subset of cases when using MDS, and we chose to use recent cases as this would probably provide most value to investigators.

Before using MDS, we first imputed the missing digits using our two probability methods. This was to ensure that all 24 loci would contribute to dissimilarity calculations. However, imputing digits like this does not allow the additional uncertainty associated with these cases to be clearly illustrated on the MDS plots. There could have been better ways to allow the missing digits to contribute to distance calculations. For instance, for MIRU profiles that clustered with other cases, with the exception of the missing loci, we could have adjusted the dissimilarity by taking a weighted average based on the size of the potential clusters associated with the incomplete MIRU profile. On the MDS plot, this would place the case with the incomplete profile closer to potential clusters with larger numbers of cases and further away from potential clusters with less cases. Another very simple way of ensuring that cases with imputed data would never have a dissimilarity of zero would be to add a fraction to imputed digits such as 0.5.

One of the main added benefits of using MDS was that it gave a visual overview of how similar the TB strains were between cases in Scotland over a period of one year. As part of the same analysis, we used the minimum spanning tree to connect pairs of cases together based on their similarity, and from there we could examine those pairs more closely. More specifically, for finding *speculative* clusters, we identified which pairs had a one locus difference between their MIRU profiles. This is notable as in rare circumstances, epidemiologically-linked cases can have slightly different MIRU profiles due to for example, evolution of a strain through transmission over a number of years or strains adapting to acquire antimicrobial resistance [331, 332]. Therefore it was necessary to provide a system for quickly identifying pairs of cases that have a difference in only one locus and may be linked.

The sensitivity of our results to the choice of using classical MDS was tested by rerunning the analysis using non-metric MDS which produced almost identical results. We then checked the sensitivity to the choice of distance measure being used by rerunning the analysis using Euclidean distance and the number of differing loci between MIRU profiles. Using Euclidean distance again produced very similar results, but when using the number of differing loci to measure dissimilarity, there was a greater amount of separation between *M. Bovis* cases and *M. Tuberculosis* cases. An explanation for this could be that Euclidean distance and Gower's distance both place too much importance on the magnitude of the difference in numbers of repeats in each locus, while using the number of differing loci will disregard the magnitude of the difference. Thus, perhaps an improved distance metric over any of the choices we tried would weight the dissimilarity based firstly on the number of differing loci between MIRU profiles and then secondly on the magnitude of the difference in numbers of repeats within each locus.

To find possible links between pairs of cases, we could also have explored the use of the eBURST algorithm [114, 122], but ultimately decided to create a bespoke system to have better control over the output produced from a tool. The eBURST software provides a tool that attempts to identify related groups of genotypes which can potentially be considered to be mutually exclusive. The algorithm was originally designed for use with multilocus sequence typing (MLST) data – a sequencing technique that has been found to have much lower discriminatory power for *M. tuberculosis* isolates than VNTR-MIRU [291]. This is because the genome of *M. tuberculosis* has little DNA sequence diversity and is thus, described as being *genetically monomorphic* [70]. However, eBURST could also be applied to the MIRU data as it essentially comes in the same form as MLST data; both types of data consist of strings of integers. More specifically, eBURST could have been used to find the pairs of cases that have single locus variations in the same manner as we done using the MST. In fact, the eBURST software can also find pairs that have double locus and triple locus variants and so it may be a useful way of quickly identifying these even more tenuously linked pairs.

The systems designed here have limitations as we have discussed, but they nevertheless serve as practical solutions for identifying potential clusters of TB cases which may otherwise have been missed by HPS. On that account, they could prove crucial in preventing clusters from growing to a size where they become difficult to manage. The systems should be applicable in Scotland for a number of years into the future as 24 loci VNTR-MIRU remains as the primary method for TB strain typing. However, looking further ahead, VNTR-MIRU will likely be superseded by more accurate methods for tracing TB outbreaks such as whole-genome sequencing (WGS) [310]. The cost of performing WGS continues to decrease [243], and the speed at which we can analyse this data will increase. On top of this, the increased discriminatory power allows for the detection of microevolution within TB strains [370] and the extra power could also alleviate the need to always perform such detailed epidemiological and contact investigations on TB cases. Therefore, it seems inevitable that WGS will become the universal standard when it becomes affordable for health services to perform WGS on all TB cases.

TB is a disease that is endemic in Scotland and here, we have looked at how genetic data can be used to detect cases that form potential strain clusters. In the next chapter, we use statistical methods to analyse data on a large and fairly recent epidemic in Scotland. This plays an important role in understanding how a disease may affect certain groups of a population differently.

# 2.A TB Appendix

## 2.A.1 R Functions

```
fun.initcompleteclus <- function(miru.df) {
    # Assigns clus names to the clusters of complete profiles
    # This function should be used only once to set up an initial clusters dataset
    #
    # Args:
    #   mirudf: df only needing 3 columns (but can have more)
    #            - ’miru’ which has the MIRU profiles
    #               All profiles MUST have the same number of loci (24 or 15)
    #            - ’nummiss’ which has the number of missing loci for each profile
    #            - ’id’ which is a unique id for each profile
    #
    # Returns:
    #   df with clus nums for TB cases with complete profiles


    mirulength <- unique(nchar(miru.df$miru))

    # Validity checks
    if (length(mirulength) != 1 | !mirulength %in% c(15, 24)) {
        stop("MIRU profiles should either all have 24 loci or all have 15 loci")
    }
    if (any(duplicated(miru.df$id))) {
        stop("There is more than one case with the same ID")
    }

    tab <- table(miru.df$miru[miru.df$nummiss == 0])
    clus_tab <- tab[tab > 1]  # clusters have at least 2 cases
    n <- length(clus_tab)

    clus.df <- data.frame(miru = names(clus_tab), clus = 1:n,
                          clus2 = paste0("SC", sprintf("%04d", 1:n)),
                          stringsAsFactors = FALSE)

    # Create df with all cases that have profiles which are part of clusters
    cases.df <- merge(miru.df, clus.df, by = "miru",
                      all.x = TRUE, stringsAsFactors = FALSE)
    cases.df$numclus <- ifelse(!is.na(cases.df$clus), 1, 0)
    cases.df$clusposs <- NA
    cases.df <- cases.df[order(cases.df$clus, cases.df$id), ]

    return(cases.df) # df contains all cases including those not part of a clus
}
```

./Chapters/TB/scripts/initcompleteclus.R

```r
fun.update.completeclus <- function(existcases.df, clus.df,
                                    newcases.df, mirulength = 24) {
    # Updates clusters for complete profiles
    # Function is designed to be used within 'fun.clusupdate'
    #
    # Args:
    #   existcases.df: df with the TB cases before the current update
    #   clus.df: df with current clusters of TB cases
    #   newcases.df: df with new TB cases with complete profiles
    #   mirulength: Number of loci in profiles. Default is 24 loci
    #
    # Returns:
    #   updated clus df after finding clus for new complete profiles


    # New cases have no clus info at present
    newcases.df$clus     <- NA
    newcases.df$clus2    <- NA
    newcases.df$numclus  <- 0
    newcases.df$clusposs <- NA


    # For new complete clus
    tab <- xtabs(~ miru, newcases.df)
    tab2 <- tab[tab > 1]
    tab.newclus <- subset(tab2, !names(tab2) %in% clus.df$miru)

    if (length(tab.newclus) > 0) {
        tab.newclus <- tab.newclus[sort(names(tab.newclus))]

        # Add new clus
        # These are profiles which form entirely new clus
        maxclus <- max(clus.df$clus)
        add.df  <- newcases.df[newcases.df$miru %in% names(tab2), ]
        add.df  <- add.df[order(add.df$miru), ]
        clus.df <- rbind(clus.df, add.df)
        clusno  <- rep(maxclus + 1:length(tab2), tab2)
        clus.df$clus[clus.df$miru %in% names(tab2)] <- clusno
        clus.df$numclus[clus.df$miru %in% names(tab2)] <- 1
    }

    # Add new complete profiles which can be added to existing clus
    tab3 <- tab[names(tab) %in% clus.df$miru]

    if (length(tab3) > 0) {
        add.df <- newcases.df[newcases.df$miru %in% names(tab3), ]
        add.df$clus <- clus.df$clus[match(add.df$miru, clus.df$miru)]
        add.df$numclus <- 1
        clus.df <- rbind(clus.df, add.df)
    }

    # Check if new profiles can be added to existing complete profiles
    # which are not currently part of a clus
    existcases.df <- subset(existcases.df, nummiss == 0)
    allcases.df <- rbind(existcases.df, newcases.df)
    tab <- xtabs(~ miru, allcases.df)
    tab <- tab[tab > 1]
    tab <- subset(tab, !names(tab) %in% clus.df$miru)

    if (length(tab) > 0) {
        maxclus <- max(clus.df$clus)
        tab <- tab[sort(names(tab))]
        add.df <- allcases.df[allcases.df$miru %in% names(tab), ]
        add.df <- add.df[order(add.df$miru), ]
        clus.df <- rbind(clus.df, add.df)
        clusno <- rep(maxclus + 1:length(tab), tab)
        clus.df$clus[clus.df$miru %in% names(tab)] <- clusno
        clus.df$numclus[clus.df$miru %in% names(tab)] <- 1
    }

    # Update longer clus names
    z.sel <- !is.na(clus.df$clus)
    clus.df$clus2[z.sel] <- paste0("SC", sprintf("%04d", clus.df$clus[z.sel]))
    clus.df <- clus.df[order(clus.df$clus, clus.df$nummiss), ]

    return(clus.df)
}
```

./Chapters/TB/scripts/updatecompleteclus.R

```
fun.assignmissclus <- function(existcases.df, clus.df,
                               newcases.df, mirulength = 24) {
    # Updates clus for incomplete profiles
    # Function is designed to be used within ’fun.clusupdate’
    #
    # Args:
    #   existcases.df: df with the TB cases before the current update
    #   clus.df: df with current clusters of TB cases
    #   newcases.df: df with new TB cases with incomplete profiles
    #   mirulength: Number of loci in profiles. Default is 24 loci
    #
    # Returns:
    #   updated clus df after finding clus for new incomplete profiles


    # Add clus vars to ’newcases.df’
    newcases.df$clus     <- NA
    newcases.df$clus2    <- NA
    newcases.df$numclus  <- 0
    newcases.df$clusposs <- NA

    allcases.df <- rbind(newcases.df, existcases.df)

    # Reset clus info in 'allcases.df' for all those who have missing loci
    # Some profiles with missing loci could now possibly belong to more clusters
    # Doing this will keep the clus names as they were even after reassignment
    allcases.df$clus[allcases.df$nummiss > 0]     <- NA
    allcases.df$clus2[allcases.df$nummiss > 0]    <- NA
    allcases.df$numclus[allcases.df$nummiss > 0]  <- 0
    allcases.df$clusposs[allcases.df$nummiss > 0] <- NA

    clus.df <- subset(clus.df, nummiss == 0)

    # On every cycle, check profiles with missing digits for assignment to clus
    miss.df <- subset(allcases.df, nummiss > 0)

    # Each case in the loop is compared with only complete profiles
    # and this list is contained in 'ref.df'
    # (Two cases with missing loci cannot form a clus)
    ref.df <- subset(allcases.df, nummiss == 0)

    # 'clus.df' updates after each loop,
    # 'ref.df' stays constant and is what each profile is compared to
    for (i in miss.df$id) {
        # i <- miss.df$id[1]
        dat <- rbind(miss.df[miss.df$id == i, ], ref.df)

        # Split MIRU profiles into columns
        z <- unlist(strsplit(dat$miru, ""))
        z <- matrix(z, ncol = mirulength, byrow = TRUE,
                    dimnames = list(dat$id, paste0("L", 1:mirulength)))
        z <- data.frame(z)
        z[z == "_"] <- NA  # Change _ to missing

        # Look at first row (profile with missing loci)
        z.miss <- as.matrix(daisy(z, metric = "gower"))[1, ]
        # Subset to have only matches (and not matches with itself)
        z.miss <- subset(z.miss, z.miss == 0 & names(z.miss) != i)

        # Run if there is at least one matching profile
        if (length(z.miss) > 0) {
            # Find matches to current existing clusters
            clusmatch <- clus.df$clus[match(names(z.miss), clus.df$id)]

            # If profile cannot be matched to existing clus,
            # it has to form new clus
            if (any(is.na(clusmatch))) {
                maxclus <- max(na.omit(clus.df$clus))

                # IDs for profiles which match but are not currently
                # assigned a clus name
                idmatch   <- names(z.miss)[is.na(clusmatch)]
                # Unique matching profiles which are not currently
                # assigned a clus name
                mirumatch <- sort(allcases.df$miru[allcases.df$id %in% idmatch])
                # Number of unique matching profiles
                nmatch    <- length(unique(mirumatch))
                z.tab     <- as.numeric(table(mirumatch))

                clusno  <- rep(maxclus + 1:nmatch, z.tab)
                clus.df <- rbind(clus.df,
                                 allcases.df[allcases.df$id %in% idmatch, ])

                # Update clus info for matching profiles which now make clus
                # but previously didn't
                clus.df$clus[clus.df$id %in% idmatch] <- clusno
                z.sel <- clus.df$id %in% idmatch & clus.df$numclus < 1
                clus.df$numclus[z.sel] <- 1
                clusmatch <- clus.df$clus[match(names(z.miss), clus.df$id)]
            }

            clusmatch <- unique(clusmatch)
            num.match <- length(clusmatch)

            if (num.match == 1) {
                allcases.df$clus[allcases.df$id == i] <- clusmatch
            }

            # ID Number is not in the clus df so add it now
            # One copy for each clus it matches to
            if (!i %in% clus.df$id) {
                repsdf  <- miss.df[rep(which(miss.df$id == i), num.match), ]
                clus.df <- rbind(clus.df, repsdf)
                clus.df$clus[clus.df$id == i] <- clusmatch
            }

            # Update clus info
            clus.df$numclus[clus.df$id == i] <- num.match
            allcases.df$numclus[allcases.df$id == i] <- num.match

            z.sel <- clus.df$id == i & clus.df$numclus > 1
            clus.df$clusposs[z.sel] <- paste(sort(clusmatch), collapse = ", ")

            z.sel <- allcases.df$id == i & allcases.df$numclus > 1
            allcases.df$clusposs[z.sel] <- paste(sort(clusmatch), collapse = ", ")
        }
```

```
    }

    clus.df <- clus.df[order(clus.df$clus, clus.df$nummiss), ]
    z.sel <- !is.na(clus.df$clus)
    clus.df$clus2[z.sel] <- paste0("SC", sprintf("%04d", clus.df$clus[z.sel]))
    rownames(clus.df) <- NULL

    out <- list(allcases = allcases.df, clus = clus.df)
    return(out)
}
```

./Chapters/TB/scripts/assignmissclus.R

```
fun.clusupdate <- function(existcases.df, clus.df, newcases.df) {
    # Updates clusters for new profiles
    # Profiles must all have 24 loci OR all profiles must have 15 loci
    # If using 24 loci, the max number of missing loci is 2
    # If using 15 loci, the max number of missing loci is 1
    #
    # Args:
    #   existcases.df: df with only the TB cases before the current update
    #   clus.df: df with current clusters of TB cases
    #   newcases.df: df with new TB cases
    #
    # Returns:
    #   updated clus df after finding clus for new profiles


    mirulength  <- unique(nchar(newcases.df$miru))
    init.numclus <- max(clus.df$clus) # Initial number of clusters

    # Validity checks
    if (length(mirulength) != 1 | !mirulength %in% c(15, 24)) {
        stop("MIRU profiles should either all have 24 loci or all have 15 loci")
    }

    if (mirulength == 24 & max(newcases.df$nummiss) > 2) {
        stop("There are 24 loci MIRU profiles with more than 2 missing digits")
    } else if (mirulength == 15 & max(newcases.df$nummiss > 1)) {
        stop("There are 15 loci MIRU profiles with more than 1 missing digit")
    }

    if (any(duplicated(c(existcases.df$id, newcases.df$id)))) {
        stop("There is more than one case with the same ID")
    }

    # Check ’existcases.df’ and ’newcases.df’ have the same vars
    if (any(grep("clus", names(existcases.df)))) {
        z.sel <- grep("clus", names(existcases.df))
        varnames.exist <- names(existcases.df)[-z.sel]
    } else {
        varnames.exist <- names(existcases.df)
    }
    if (any(grep("clus", names(newcases.df)))) {
        varnames.new <- names(newcases.df)[-grep("clus", names(newcases.df))]
    } else {
        varnames.new <- names(newcases.df)
    }

    if (!all(varnames.exist %in% varnames.new)) {
        stop("Var names should be identical in ’existcases.df’ and ’newcases.df’")
    }

    # Ensure var names are in the same order
    if (!identical(varnames.exist, varnames.new))  newcases.df[, varnames.exist]


    # Update clus for new cases with complete profiles first
    newcases_complete <- sum(newcases.df$nummiss == 0)
    newcases1 <- subset(newcases.df, nummiss == 0)

    if (newcases_complete == 0) {
        warning("No new cases with complete MIRU profiles")
    } else{
        cat(newcases_complete, "new cases with complete MIRU profiles\n")
        clus.df <- fun.update.completeclus(existcases.df, clus.df,
                                           newcases1, mirulength)
    }


    # Update clus for new cases with incomplete profiles
    newcases_incomplete <- sum(newcases.df$nummiss > 0)
    newcases2 <- subset(newcases.df, nummiss > 0)

    if (newcases_incomplete == 0) {
        warning("No new cases with incomplete MIRU profiles")
    } else{
        cat(newcases_incomplete, "new cases with incomplete MIRU profiles\n")
    }

    missclus.out <- fun.assignmissclus(existcases.df, clus.df,
                                       newcases.df, newcases.complete,
                                       mirulength)

    clus.df <- missclus.out$clus

    # Reporting and Output
    numclus <- max(clus.df$clus) # Number of clusters after update
    num.newclus <- numclus - init.numclus # Number of new clusters

    cat(num.newclus, "new clusters. Before update:", init.numclus,
        "clusters. After update:", numclus, "clusters.\n")

    out <- list(allcases =  missclus.out$allcases, clus = clus.df)
    return(out)
}
```

./Chapters/TB/scripts/clusupdate.R

```
fun.clusupdate <- function(existcases.df, clus.df, newcases.df) {
    # Updates clusters for new profiles
    # Profiles must all have 24 loci OR all profiles must have 15 loci
    # If using 24 loci, the max number of missing loci is 2
    # If using 15 loci, the max number of missing loci is 1
    #
    # Args:
    #   existcases.df: df with only the TB cases before the current update
    #   clus.df: df with current clusters of TB cases
    #   newcases.df: df with new TB cases
    #
    # Returns:
    #   updated clus df after finding clus for new profiles


    mirulength  <- unique(nchar(newcases.df$miru))
    init.numclus <- max(clus.df$clus) # Initial number of clusters

    # Validity checks
    if (length(mirulength) != 1 | !mirulength %in% c(15, 24)) {
        stop("MIRU profiles should either all have 24 loci or all have 15 loci")
    }

    if (mirulength == 24 & max(newcases.df$nummiss) > 2) {
        stop("There are 24 loci MIRU profiles with more than 2 missing digits")
    } else if (mirulength == 15 & max(newcases.df$nummiss > 1)) {
        stop("There are 15 loci MIRU profiles with more than 1 missing digit")
    }

    if (any(duplicated(c(existcases.df$id, newcases.df$id)))) {
        stop("There is more than one case with the same ID")
    }

    # Check 'existcases.df' and 'newcases.df' have the same vars
    if (any(grep("clus", names(existcases.df)))) {
        z.sel <- grep("clus", names(existcases.df))
        varnames.exist <- names(existcases.df)[-z.sel]
    } else {
        varnames.exist <- names(existcases.df)
    }
    if (any(grep("clus", names(newcases.df)))) {
        varnames.new <- names(newcases.df)[-grep("clus", names(newcases.df))]
    } else {
        varnames.new <- names(newcases.df)
    }

    if (!all(varnames.exist %in% varnames.new)) {
        stop("Var names should be identical in 'existcases.df' and 'newcases.df'")
    }

    # Ensure var names are in the same order
    if (!identical(varnames.exist, varnames.new))  newcases.df[, varnames.exist]


    # Update clus for new cases with complete profiles first
    newcases_complete <- sum(newcases.df$nummiss == 0)
    newcases1 <- subset(newcases.df, nummiss == 0)

    if (newcases_complete == 0) {
        warning("No new cases with complete MIRU profiles")
    } else{
        cat(newcases_complete, "new cases with complete MIRU profiles\n")
        clus.df <- fun.update.completeclus(existcases.df, clus.df,
                                           newcases1, mirulength)
    }


    # Update clus for new cases with incomplete profiles
    newcases_incomplete <- sum(newcases.df$nummiss > 0)
    newcases2 <- subset(newcases.df, nummiss > 0)

    if (newcases_incomplete == 0) {
        warning("No new cases with incomplete MIRU profiles")
    } else{
        cat(newcases_incomplete, "new cases with incomplete MIRU profiles\n")
    }

    missclus.out <- fun.assignmissclus(existcases.df, clus.df,
                                       newcases.df, mirulength)

    clus.df <- missclus.out$clus

    # Reporting and Output
    numclus <- max(clus.df$clus) # Number of clusters after update
    num.newclus <- numclus - init.numclus # Number of new clusters

    cat(num.newclus, "new clusters. Before update:", init.numclus,
        "clusters. After update:", numclus, "clusters.\n")

    out <- list(allcases =  missclus.out$allcases, clus = clus.df)
    return(out)
}
```

./Chapters/TB/scripts/clusprob.R

## 2.A.2 Instructions for Updating TB Clusters

### Excel Processing

Make a copy of the data file so that the original data file is always available as a backup – **Work with the new copy**.

**Enable VBA scripts** from options so that the "Developer" options are available in Excel.

**Import the "AmendBold" function ("AmendBold.bas")** into the Excel Workbook in the **"resources" folder**).

### ESMI Worksheet

Ensure everything in the "MIRUFINAL" and "UPDATED PROFILE" columns are **formatted as text** as the "AmendBold" function will not work unless data are formatted as text. If some are in numeric form, they will be easy to spot as they will be aligned right instead of aligned left. Delete the columns with headings "MIRU16", "MIRU25", "MIRU26" as they are not needed.

Sort Descending by "UPDATED PROFILE" column.

Copy any with updated profiles across to the "MIRUFINAL" column.

Delete the "UPDATED PROFILE" column as it is not needed anymore.

Create a new column with heading "MIRUFINAL2" in the first empty column

In the new column, use the **AmendBold function** to get the new version of the MIRU profile with #s to distinguish double-digit repeats (e.g. 10 will become 1#0#, etc). This is vital so R can distinguish double-digit repeats when the data is read into it.

If "MIRUFINAL" is in e.g. column "I" and "MIRUFINAL2" was created in column "K" then in Cell "K2" we would input the formula "= AmendBold(I2)".

### NO ESMI Sheet

Ensure everything in the "MIRU" and "MIRU UPDATED" columns are **formatted as text**. The reasons and process for this are the same as before with the ESMI worksheet.

Delete the columns with headings “MIRU16”, “MIRU25”, “MIRU26” as they are not needed.

Sort Descending by “MIRU UPDATED”.

Copy any with updated profiles across to the “MIRU” column.

Delete the “MIRU UPDATED” column as it is not needed anymore.

Create a new column with heading “MIRU2” in the first empty column

In the new column, use the AmendBold function to get the new version of the MIRU profile with #s to distinguish double-digit repeats.

### Combined Sheet

Create new worksheet called “Combined”.

Copy all contents from the ESMI worksheet into it first (use ctrl+a).

Then carefully add in the NO ESMI sheet as follows:
Add the “RefLabReportID” heading to the first empty column in the Combined worksheet.
Add the values for RefLabReportID to the combined worksheet, starting from the first available row, e.g. if the ESMI worksheet had 5480 rows in it, the first RefLabReportID would be in row 5481.

Add in the other column values as follows:
“LabReportYear” goes in “DATENOTYR”
“FirstReport” goes in “DATENOT”
“SpeciesTxt” goes in “Organism”
“MIRU” goes in “MIRUFINAL”
“MIRU2” goes in “MIRUFINAL2”

With reference to the example before, you should now have 5480 rows with ESMI data, and all the NO ESMI data should be contained in rows 5481 onwards - these will be missing ESMI data such as age and gender.

Save the combined worksheet as a csv file with filename “esmi_noesmi_combined.csv”. This is the file to be read into R.

### R Analysis

Make sure all the **required packages** are installed. These will be loaded automatically at startup when the workspace is loaded using the **.Rprofile file** located in the same directory as the workspace. The essential package to produce the final output is "XL-Connect".

Open the workspace file **"TB.RData"**. This is important as it will automatically set the working directory appropriately for the rest of the scripts to work in R. As explained above, it also loads the required packages, provided they have been installed.

Run the script **"data_setup.R"** to tidy the data and set up the dataframes. It will save a new workspace as **"TB2.RData"** at the end which is used from this point onwards. The main dataframe we work with is **"miru.df"** - this contains only those cases with MIRU profiles.

Load the "TB2.RData" workspace. Run the script **"freq_digits_by_locus.R"** to get **digits.mat** which is required to get probabilities for clusters, this is added to the "TB2.RData" workspace at the end. Note that the script also produces barplots which are saved in the "figs" folder.

Look in the **"output" folder**. If the file **"clus24.xlsx"** is already there, **delete** it as the function to produce the Excel sheet using the XLConnect package will not work if it the output file is already there.

Run the script **"run_functions_clus24.R"**. This produces the file **"clus24.xlsx"** which is the final output. It takes about a few minutes to run the cluster function and also a few minutes to run the XLConnect parts.

### 2.A.3 Visual Basic Script for AmendBold Excel Function

```
Function AmendBold(ByVal rngText As Range) As String
   Dim rngCell      As Range
   Dim strRetVal    As String
   Dim strChar      As String
   Dim lngI         As Long

   Set rngCell = rngText.Cells(1, 1)
   For lngI = 1 To Len(rngCell.Value)
      strChar = rngCell.Characters(lngI, 1).Text
      strRetVal = strRetVal & strChar
      If rngCell.Characters(lngI, 1).Font.FontStyle = "Bold" Then
         strChar = "#"
         strRetVal = strRetVal & strChar
         If strChar = " " Then strRetVal = strRetVal & strChar
      End If
   Next

   AmendBold = Application.Trim(strRetVal)

End Function
```

# Chapter 3

# Characteristics of 2009 Pandemic Influenza in Scotland

Chapter 2 on Tuberculosis was on identification of disease outbreaks for an endemic disease, and we developed methods which can be used to detect and alert epidemiologists to potential linked cases and outbreaks. One of the issues we came across there was missing data where MIRU profiles were frequently incomplete and the missing numbers of tandem repeats were addressed by data imputation. Missing data is an issue that comes up often in epidemiological research [345] and it will also be a feature of this chapter where we explore some of the statistical issues that can arise when analysing a large disease outbreak.

In March and early April 2009, a novel strain of Influenza A (H1N1) virus emerged originating from swine which gave rise to its colloquial name, *swine flu*. The 2009 pandemic H1N1 virus (referred to throughout this chapter as “pH1N1” where “p” refers to the fact that the outbreak was classified as a pandemic) was unique as it contains a combination of genes from both north American and Eurasian swine lineages that had never been previously identified in swine or human populations [343]. With any novel disease outbreak, it is important to gain an understanding of it’s characteristics as soon as possible. This includes looking at the symptoms that cases are reporting, and whether or not the disease affects certain groups of individuals more adversely. Information such as this is vital for mounting an efficient response to the outbreak.

In this chapter, we analyse a dataset provided by Health Protection Scotland

(HPS) which includes just over 1,000 early pH1N1 cases in Scotland. This equates to around 80% of all the cases that were eventually reported to HPS during the *containment phase* of the pandemic [161]. As the name suggests, this was the phase where the strategy was to contain the spread of the disease. An additional priority during this phase was to gather as much information as possible on the early cases in order to learn about the disease. Hence, these cases were interviewed and answered questionnaires where they gave details regarding their symptoms, what health services they used, their use of antivirals and also provided general demographic information. Note that in addition to this dataset, HPS also kept a larger database of early pH1N1 cases which included cases that did not provide the detailed data such as symptoms information but may have only provided information on when they acquired the infection; we utilise the larger database later in Chapter 5 where we look at estimating reproductive numbers for pH1N1 in Scotland during the containment phase.

A general aim in this chapter is to provide an overview of the information provided by these cases and to determine which groups of individuals may have been more severely or disproportionately affected by the disease. In addition to this, one of the main aims is to investigate the types of symptoms reported by cases in Scotland. More specifically, we wish to find out whether or not individuals with different characteristics are reporting different types and numbers of symptoms – this includes individuals belonging to different age groups, gender and socioeconomic groups. This allows for comparisons with what is being reported in other studies in the UK and other countries around the world. As well as this, we wish to explore if individuals with different characteristics were getting the same quality of treatment as measured by whether or not they received antivirals and also by looking at the length of time they had to wait to receive antivirals. In the containment phase, data was also collected on the contacts of cases. This gives us a rare opportunity to look at case-contact mixing patterns and relationships in the midst of a rapidly unfolding epidemic. For instance, we can explore how the numbers of contacts reported by cases varies depending on the characteristics of cases. As social contact patterns can have significant implications on transmission of communicable diseases, attempting to gain a better understanding around these patterns is an important goal.

## 3.1 Background

The first pH1N1 cases recorded by the Centers for Disease Control and Prevention (CDC) occurred in Mexico [60, 227] and the United States [132] but there have been suggestions that the initial transmission to humans occurred several months before the outbreak was formally recognised [343]. The virus became a real cause for concern when increased rates of hospitalisation and death due to severe pneumonia in young adults were observed in Mexico [289].

Efforts were made by the Mexican government to keep the virus under control by closing public and private facilities as well as general measures like social distancing, timely medical care, and personal hygiene. However, the virus was able to spread rapidly through human-to-human contact, particularly in enclosed environments. The efforts from the Mexican government to contain the virus proved to be insufficient and after only a few weeks of proper surveillance, the virus had spread to 30 countries.

The rapid spread around the globe was inevitable due to the high volumes of air travel throughout the world and in particular, the United States. As a consequence, on the 11th of June, the World Health Organization (WHO) raised the pandemic alert to the highest level of 6 signalling the official start of the 2009 influenza pandemic [63]. By this time, there were almost 30,000 confirmed cases reported from 74 countries. Transmissibility appeared to be substantially higher than that of seasonal influenza and there was the belief that it may have been more comparable with that of previous influenza pandemics [123] such as the 1918 Spanish influenza pandemic. Hence, the 2009 strain quickly became a matter of serious interest and importance for public health organisations around the world.

### 3.1.1 Pandemic Influenza in the UK

The first two confirmed cases in the United Kingdom were a couple of travellers returning to Scotland from Mexico with respiratory symptoms on the 26th of April 2009 [155, 286]. These cases were officially recorded on the 27th of April. HPS subsequently implemented an enhanced surveillance scheme to track cases as well as their close contacts. Initially most cases resulted from travellers returning to the UK from Mexico and the United States (imported cases), but on the 1st of May 2009, the first indigenously acquired case was reported and since then there was a rapidly rising increase in the number of indigenous cases reported [18]. By the 2nd of July 2009, 7,447 cases were reported in the UK.

### 3.1.2 Known Symptoms

The vast majority of people suffering from 2009 pH1N1 experienced mild symptoms and made a full recovery within about a week without the need of any kind of medical intervention. The symptoms of this strain of influenza were typically the same as the symptoms for seasonal influenza; people with the virus usually had a fever or a high temperature (in excess of 38 °C or 100.4 °F). Other symptoms included unusual levels of tiredness, headache, runny nose, sore throat, shortness of breath, coughing, loss of appetite, muscle ache, and possibly, diarrhoea and vomiting [240]. On rare occasions, the disease caused severe illness which may require hospitalisation and these have been seen to mostly affect adults between 30–50 years [63]. This is in stark contrast to the case of the common seasonal influenza where the majority of deaths and hospitalisations occur in frail elderly people.

### 3.1.3 High-Risk Groups

People with certain underlying conditions were considered to belong to *high-risk groups* for pandemic influenza as these conditions may render them to be more susceptible to adverse outcomes such as hospitalisation and death. These conditions include chronic respiratory disease, coronary heart disease, kidney disease, liver disease, neurological disease, immunosuppression and diabetes [287]. Moreover, patients who had drug treatment for asthma in the past three years were also considered to be at increased risk of adverse reactions following influenza infection.

Pregnant women were of particular concern because during pregnancy, a woman's immune system becomes weakened [52, 186]. A study conducted by Oluyomi-Obi et al. [271] in Manitoba in Canada looked at thirty pregnant women, of which six were admitted to the intensive care unit. For these six women, the time from onset of symptoms to life-threatening deterioration was on average, just five days and two of the patients died. Although the study was fairly small, it gives cause for concern and highlights why it is vitally important for pregnant women to take extra precautions during periods when pandemic influenza is circulating.

## 3.2 Data

### 3.2.1 Pandemic Influenza Cases

The data used in this analysis consists of the pandemic influenza cases who reported their illness to HPS between April to July 2009 and were also given questionnaires regarding their symptoms (see Figure 3.2 for a visual representation of which cases have been included). Specifically, the earliest date reported by a case was the 25th of April and the last reported date was the 9th of July; the corresponding self-reported dates of symptoms onset for these cases were 23rd April and 30th June, respectively. The collection of detailed data from cases ceased from around that time as after the 2nd of July, the containment phase of the pandemic in Scotland was declared as over due to the rapidly growing number of cases [161]. From here the focus shifted to treating those considered to be most at risk of serious outcomes following infection rather than trying to reduce the risk of infection from cases to their close contacts. Furthermore, from this point, it was clear that it would take far too much resource from the health service to swab and collect detailed clinical and epidemiological information from so many people presenting with respiratory symptoms and influenza-like-illness.

All cases fall under three categories which are *possible*, *probable* and *confirmed*, with the level of assurance that a case is a true case rising through these three categories in that order. Possible cases were people who had a history of acute respiratory illness (ARI) and also recent travel to a known affected area (such as the US or Mexico) or had been in contact with a confirmed or probable case. Probable cases were defined as a person who had the characteristics of a possible case but has also tested positive for influenza A but was not subtypeable. Finally, a confirmed case referred to an individual who tested positive for the new influenza A(H1N1) virus by the use of specific reverse transcription polymerase chain reaction (RT-PCR) confirmed by sequence analysis [286]. During the containment phase of the pandemic, anybody suspected of having been infected with pandemic influenza was asked to provide a swab for laboratory-testing.

A large variety of clinical and epidemiological information was collected from case questionnaires and interviews. This includes demographic details such as gender, date of birth and their residential postcode; clinical illness history (for example, dates for onset of symptoms and types of symptoms reported); whether or not cases had underlying risk factors such as associated chronic diseases; whether or not females were pregnant when diagnosed; and details of their use of different health services available from the National Health Service (NHS). Finally, follow-up information on if cases received antiviral drugs was available as well.

The postcodes of cases were linked to datazones (DZs) which are small areas of statistical geography in Scotland covering between around 500 to 1,000 residents. These DZs could subsequently be matched to a deprivation quintile measured by the 2009 Scottish Index of Multiple Deprivation (SIMD) [328] to give an approximation around the socioeconomic status of cases. For clarity, we mention here that deprivation quintile one refers to people living in DZs belonging to the 20% most deprived DZs in Scotland.

### 3.2.2 Modifications and Assumptions

The dataset originally consisted of 1,157 cases, but for this analysis we wanted to concentrate on only laboratory-confirmed pandemic influenza cases. This reduced the size of the dataset down to 1,138 cases; nineteen cases not classified as "confirmed" and with no positive influenza laboratory test result recorded were dropped (two cases classified as "possible", fourteen as "probable" and three not classified with any status).

Data was available on 21 different symptoms which are listed in Table 3.1. However, in order to make the analysis more concise, some similar symptoms were grouped together – dry and productive coughs were combined together as *cough*, muscle ache and joint ache were combined as *muscle or joint ache*, and nausea and vomiting were combined as *nausea or vomiting*. This leaves data on 18 different symptoms which were used in all analyses regarding symptoms instead of the original 21 symptoms.

### Missing Value Mechanism for Symptoms Data

Since many cases did not provide responses for all eighteen symptoms (see Table 3.1, noting that the total number of cases for analysis was 1,138), data imputations were made in order to improve the completeness of the data. For the imputations, we made the assumption that patients would often leave a question blank rather than answer it as "No" when they did not have a symptom. Therefore, if a case had answered at least one question regarding what symptoms they had whilst other symptoms for that case were not filled in, then those blank entries were assumed to be answered as "No". However, if all entries regarding symptoms were missing for a case, then the blank entries were treated as "Missing". The process for imputing missing responses is shown schematically on Figure 3.1.

The assumptions regarding the mechanism behind how missing data arose were considered to be reasonable as if a person had at least filled in one question regarding their symptoms, then we know that they have looked at the form and in this scenario, they may have decided to only fill in information for symptoms they had to save time. However if no questions regarding symptoms were answered, then we cannot easily make any prediction about that person's symptoms and they may not have looked at the symptoms questionnaire at all. After imputations, data on

| Symptom | Responses Before Imputation | Imputed No Responses |
|---:|---:|---:|
| Dry Cough, Productive Cough | 1,020 | 31 |
| Muscle ache, Joint ache | 968 | 83 |
| Nausea, Vomiting | 991 | 60 |
| Fever | 1,031 | 20 |
| Runny Nose | 986 | 65 |
| Sneezing | 972 | 79 |
| Sore Throat | 995 | 56 |
| Shortness of Breath | 949 | 102 |
| Diarrhoea | 966 | 85 |
| Fatigue | 978 | 73 |
| Loss of Appetite | 948 | 103 |
| Chills | 944 | 107 |
| Conjunctivitis | 936 | 115 |
| Headache | 954 | 97 |
| Seizures | 946 | 105 |
| Altered Consciousness | 944 | 107 |
| Nose Bleed | 947 | 104 |
| Rash | 950 | 101 |
| Number missing all symptoms | 87 | |
| Total responses after imputation | 1051 | |

**Table 3.1:** Numbers of symptom responses before and after data imputation. Note that for analysis, dry and productive coughs were combined, muscle ache and joint ache were combined, and nausea and vomiting were grouped together, leaving 18 different symptoms instead of the original 21 symptoms. The number of responses refers to cases who gave either ‘Yes’ or ‘No’ responses. Imputed ‘No’ responses were made according to the process shown in Figure 3.1.

all 18 symptoms was available for 1,051 cases which is 92.4% of the 1,138 cases under consideration (Table 3.1).

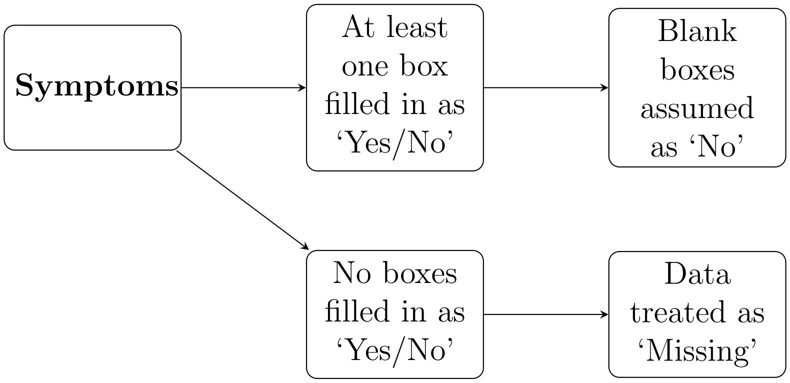

Symptoms
At least one box filled in as 'Yes/No'
Blank boxes assumed as 'No'
No boxes filled in as 'Yes/No'
Data treated as 'Missing'

**Figure 3.1:** Flowchart showing the assumptions made when dealing with missing data on symptoms that patients reported.

### 3.2.3 Contacts of Cases

Most cases provided some details regarding their close contacts, who were also interviewed to discover demographic information on contacts among other important details such as if they were symptomatic, and the type of relation they had to the case. For the type of relation to the case, assumptions were made to categorise contacts as household and non-household contacts – close family members (brothers, sisters, mothers, fathers) were assumed to be household contacts, while friends, co-workers, grandparents and other social contacts were assumed to be non-household contacts.

There were a number of issues with the contacts dataset which imposed limitations on what could be achieved through analysis using it. Although there was data on 3,809 individual contacts of cases, only 3,514 (92.3%) of them could be linked to a case belonging to the cases dataset, and only these contacts were included (Figure 3.2). This means that around 8% of the contacts data could not be used when examining case-contact mixing. Moreover, the data on contacts was on the whole, very incomplete and difficult to extract meaningful information from due to all or much of the data in some variables being entered in free text form (see Table 3.2 for more details on issues with variables in the contacts dataset).

Despite these issues, it was still possible to undertake some meaningful investigations like analysing how numbers of reported contacts varied by the characteristics of cases. The data also allowed us to look into case-contact mixing patterns such as if cases in particular age groups often had more contacts in particular age groups. With regards to these mixing patterns, we explored three main patterns: ages of case against ages of their contacts; gender of cases against gender of their contacts; and deprivation quintile of cases against deprivation quintile of their contacts.

| **Variable** | **% Missing** | **Notes** |
|---:|---:|:---|
| Link Number to Case | 7.7 | 7.3% of contacts linked to a case not in the cases dataset; 0.4% had no link to a case |
| Gender | 3.8 | |
| Date of Birth | 17.1 | Date of birth was missing for all the contacts of around 6% of cases |
| Deprivation | 55.3 | Postcodes missing or partially missing |
| Town | 40.5 | Lack of consistency for analysis as most contacts were from different towns |
| Contact Type | 21.9 | Modified to Household or Non-household contact based on assumptions |
| Date Last Contact | 49.5 | Many data entry errors; not all records were specific dates |
| Other | 71.7 | Free text for other information on contacts, e.g if they travelled with the case or if contact was a work colleague of a case |
| Symptoms | 70.0 | Modified to if contact reported having symptoms or not having symptoms |
| Outcome | 74.4 | Free text for actions taken or advised for contact, e.g. prescribing of antivirals, speaking to a GP, if contact has been swabbed |

**Table 3.2:** Percentages of missing data for variables in the contacts dataset. The percentages were taken from the total number of contacts, which includes contacts that could not be linked to a case from the cases dataset ($n$ = 3,809). Note that for some variables, the amount of data that can be utilised in analysis may be even lower than reported here as even when data had been filled in, it was not always entered correctly.

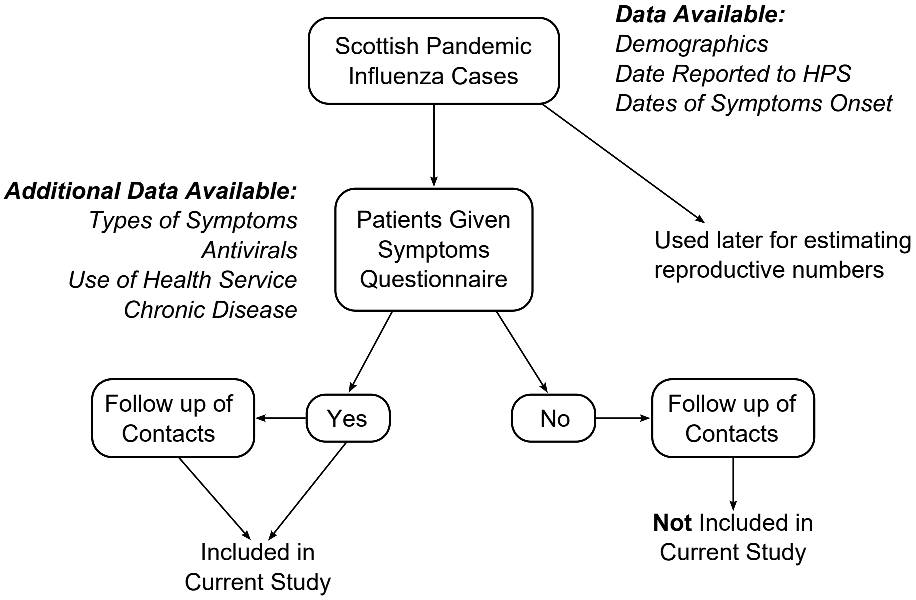

Data Available:
Demographics
Date Reported to HPS
Dates of Symptoms Onset
Scottish Pandemic Influenza Cases
Additional Data Available:
Types of Symptoms
Antivirals
Use of Health Service
Chronic Disease
Patients Given Symptoms Questionnaire
Used later for estimating reproductive numbers
Follow up of Contacts
Yes
No
Follow up of Contacts
Included in Current Study
Not Included in Current Study

**Figure 3.2:** Visual representation showing which Scottish pandemic influenza patients reported to HPS were included in the current study – only cases that were given the symptoms questionnaire have been included. In addition to this, contacts of these cases were also used for this study. Note that all cases (those given and not given the questionnaire) are used in Chapter 5 for estimating reproductive numbers.

## 3.3 Statistical Methods

General descriptive analyses were performed on the data to provide an overview of the characteristics of cases with the aim of finding out if certain groups of individuals were disproportionately affected by pandemic influenza in Scotland, and then also comparing our findings with what other studies have reported. This included looking at the demographics of cases; investigating the numbers and types of clinical symptoms reported by cases; finding how many reported underlying medical conditions and how many had contact with different health services. Furthermore, we also examined treatment by considering how many cases received antivirals and if they received them, how long they had to wait to receive antivirals. Note that throughout this chapter, we will often refer to the length of time that cases had to wait for antivirals simply as the *treatment delay.*

Standard statistical tests were employed to find if there were differences by age and gender related to the numbers of symptoms reported, acquisition of antivirals and the delay before receiving antivirals. These included $\chi^2$ tests, and non-parametric tests such as Mann-Whitney and Kruskal-Wallis tests; non-parametric tests were preferred here as numbers of symptoms and treatment delays were not normally distributed. For some comparisons, we calculated proportions and to derive confidence intervals around these proportions, we used the Wilson score interval [388]. Compared with the normal approximation method for deriving an interval, Wilson's method has the advantage of not allowing a negative lower limit for proportions and the procedure also works better for small sample sizes [3] which may occur for rarer symptoms.

### 3.3.1 Linear-by-Linear Associations

We also wanted to find if numbers of symptoms reported, acquisition of antivirals and treatment delay varied by deprivation quintile. We hypothesised that for instance, numbers of symptoms reported may increase linearly with increasing deprivation while we did not believe that there would be a linear trend for age. Hence, we used linear-by-linear tests of association to test for this trend in deprivation. The linear-by-linear test statistic is denoted as $M^2$ throughout this chapter and is calculated as in Equation (3.1) (see Agresti [2, chap. 2] for more details).

$$M^2 = (N-1)\,r^2 \tag{3.1}$$

$N$ refers to the total number of cases, and when $N$ is large, $M^2$ approximately follows a $\chi^2$ distribution on 1 degree of freedom. Furthermore, $r$ denotes Pearson's correlation coefficient which can be calculated after categories within ordered variables are assigned scores. For instance, deprivation quintiles one to five can simply be given scores of one to five, and other ordered variables can be assigned scores similarly.

### 3.3.2 Logistic Regression and Multiple Testing

The types of symptoms reported by cases have been found to differ by age group and gender in other studies and surveillance reports (for example, see Jhung et al. [190], McLean et al. [240], Pebody et al. [286]). Thus, we also wished to compare the symptoms reported by Scottish cases to ascertain if our findings would be consistent with those studies. Differences in the types of symptoms reported by age groups of cases, by gender and by deprivation status of cases were found by using logistic regression [174]; odds ratios (ORs) were subsequently found by exponentiating the regression equation of the form shown in Equation (3.2), where $p_{\text{symp}}$ refers to the probability of reporting one of the eighteen symptoms.

$$\log\left(\frac{p_{\text{symp}}}{1-p_{\text{symp}}}\right) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_n x_{n,i} \tag{3.2}$$

As we wanted to look at each of the eighteen symptoms one-by-one, we had to perform a regression for each symptom and for this, we had to take into consideration the issue of multiple-testing [319]. To account for this issue, we used the Bonferroni correction [35] which imposes a more stringent significance threshold – for eighteen symptoms, an original 5% significance level is reduced to $\alpha = \frac{0.05}{18} \approx 0.0028$. The Bonferroni correction produces conservative results but we decided to use it here as we desired to have a lower chance of attaining false positive results.

### 3.3.3 Principal Component Analysis

Principal component analysis (PCA) [175, 193] was used to try and determine how different groups of symptoms clustered among the cases. For our application, PCA attempts to capture most of the variation from the symptoms data in a smaller set of linear combinations of the original variables which are named principal components (PCs). To do this, the procedure transforms the variables into an equal number of linearly uncorrelated variables which are the PCs. Hence for our data, the eighteen variables are transformed into eighteen PCs where the first PC captures the most variation from the data and each PC thereafter, captures a monotonically decreasing amount of the variance from the data. Similarly to multidimensional scaling (MDS) which was used for cluster analysis in Chapter 2, a motivation for using PCA is that it allows for a visualisation of high-dimensional data in two or perhaps, three dimensions (provided that the variation can be sufficiently captured in a small set of PCs), and this makes it much easier to explore the data for the purpose of identifying possible clusters.

## 3.4 Results

### 3.4.1 Demographics

From the 1,138 laboratory-confirmed early cases of 2009 pH1N1 in Scotland, Table 3.3 shows that the gender split was even with 562 (49.4%) male cases and 576 (50.6%) female cases. The ages of cases ranged from 0–82 and around 10.6% of cases were aged 0–4, but the majority of cases were aged 15–64 (54.9%) or 5–14 (32.4%), which indicates that school age children and those in the working ages were predominantly affected. This is supported with the median age of cases being 17 and an IQR of 9–28. Notably, only 15 cases were aged 65+ (1.3%), and also note that age was not known for eight cases as their dates of birth were missing. Figure 3.3 illustrates that the gender split for cases remained almost equal even when stratified into the four age groups ($\chi^2 = 2.42$, $p = 0.49$).

Over three quarters (79.9%) of cases were from the Greater Glasgow and Clyde (GGC) healthboard, with numerous cases also coming from Highland and Lanarkshire (see Table 3.3); all of the other healthboards had less than 20 cases each. There was a slight tendency for cases to be living in more deprived areas of Scotland with just under half of all cases coming from areas belonging to the two most deprived quintiles (48.2%). It should also be noted that 156 cases were missing deprivation information – 68 cases did not give their postcode, 52 cases only gave partial details of their postcode, and the remaining 42 either wrongly recorded their postcode or their postcode could not be matched to a DZ in Scotland.

| **Variable** | **Group** | **Cases (%)** |
|---|---|---|
| Gender | Female | 576 (50.62) |
| | Male | 562 (49.38) |
| Age Group | 0-4 | 121 (10.63) |
| | 5-14 | 369 (32.43) |
| | 15-64 | 625 (54.92) |
| | 65+ | 15 (1.32) |
| | Unknown | 8 (0.70) |
| Healthboard | Ayrshire and Arran | 8 (0.70) |
| | Dumfries and Galloway | 4 (0.35) |
| | Fife | 6 (0.53) |
| | Forth Valley | 19 (1.67) |
| | Greater Glasgow and Clyde | 909 (79.88) |
| | Grampian | 2 (0.18) |
| | Highland | 120 (10.54) |
| | Lanarkshire | 46 (4.04) |
| | Lothian | 7 (0.62) |
| | Tayside | 17 (1.49) |
| Deprivation Quintile | Most Deprived | 334 (29.35) |
| | 2 | 214 (18.80) |
| | 3 | 161 (14.15) |
| | 4 | 126 (11.07) |
| | 5 | 147 (12.92) |
| | Unknown | 156 (13.71) |

**Table 3.3:** Counts and percentages for early 2009 pH1N1 cases in Scotland by gender, age group, healthboard and deprivation quintile.

### 3.4.2 Symptoms

The numbers and percentages of cases reporting each of the 18 symptoms are presented in Figure 3.4. The most common symptoms reported by cases were consistent with what was mentioned previously in Section 3.1.2; these include 80–90% of cases reporting a dry or productive cough, fever and fatigue; 70–75% of cases having sore throat and headache; and 60–70% reporting muscle or joint ache, loss of appetite, runny nose and chills. In addition to this, 48.3% of cases reported

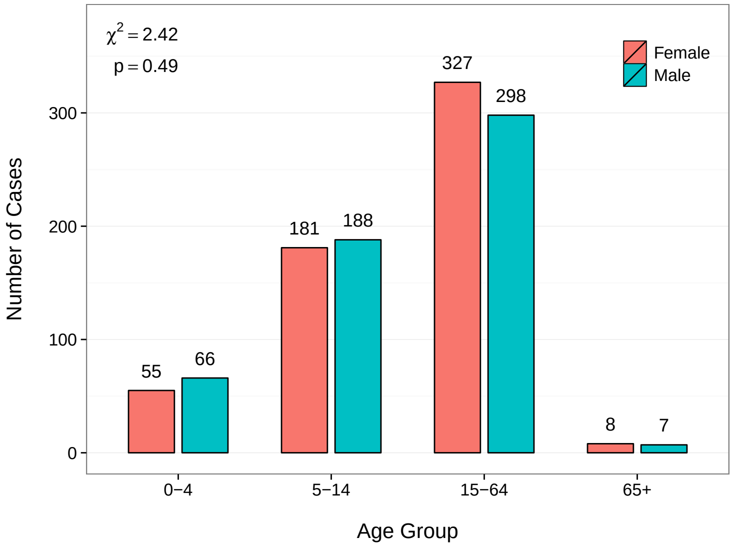


**Figure 3.3:** Gender distribution by age group for early cases of 2009 pH1N1 in Scotland.

nausea or vomiting, while about a quarter of cases reported diarrhoea (24.6%). A minority of cases (around 10% of cases or less) reported symptoms such as seizures, rash, nose bleed, altered consciousness and conjunctivitis.

### Differences in Symptoms by Age Groups

The ORs in Table 3.4 give the odds of reporting each of the 18 symptoms for individuals in different age groups compared with individuals aged 15–64, which was used as the reference category as they make up the largest group of cases. Here, five cases that reported on symptoms were not used in the model as their age was unknown and so 1,046 cases were included in the model instead of the 1,051 that did report on symptoms. Note that an additional three cases also had unknown age but they also had unknown responses for all 18 symptoms and therefore would not be used anyway. An overall $p$-value for each symptom was derived which comes from a general test of association between reporting a symptom and age. Note that when considering which results to be interesting for discussion here, we looked for an overall $p \approx 0.0028$ or lower due to using Bonferonni corrections for multiple testing (described earlier in Section 3.3.2). As well as the ORs, the number of cases within age groups that reported at least one symptom ($N$) and percentages of cases within age groups reporting each symptom are presented in

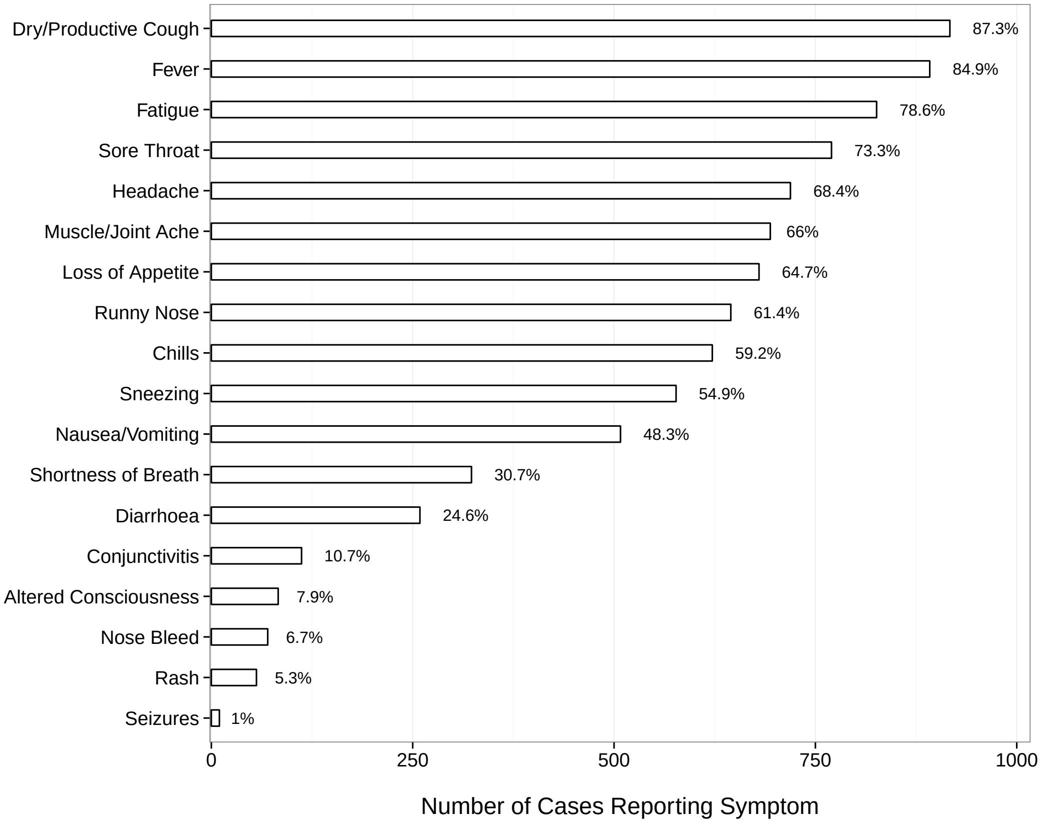


**Figure 3.4:** Symptoms reported by early 2009 pH1N1 cases in Scotland. The percentages were calculated using the numbers of cases answering either 'Yes' or 'No' after imputation (see Table 3.1).

the table along with the associated 95% confidence interval around the percentage.

Significant differences were observed between age groups for seven symptoms which were fever, runny nose, sore throat, shortness of breath, muscle/joint ache, chills and headache. For all of these symptoms except fever, significant differences were found when comparing the 0–4 age group with those aged 15–64. Those aged 0–4 were around twice as likely to report runny nose (OR 2.16, 95% CI 1.35–3.45) compared with those aged 15–64. However, those in the youngest age group were less likely to report the other symptoms; they were around 60% less likely to report sore throat and chills; about 72% less likely to report shortness of breath; and approximately 85% less likely to report muscle/joint ache and headache (see Table 3.4 for the exact OR estimates).

The significant difference for fever was observed in the 5–14 age group, who were around twice as likely to report that symptom compared with cases aged 15–64 (OR 2.10, 95% CI 1.39–3.19). In comparison to those aged 15–64, those aged 5–14 were also found to be just over 60% less likely to report muscle or joint ache and just over 70% less likely to report shortness of breath. No significant differences were observed when comparing the 65+ age group to the 15–64 age group, and for some symptoms (seizures, nose bleed and rash), comparisons could not be made as no cases aged 65+ reported these symptoms.

| **Symptom** | **Age** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| Fever | Overall | 1,046 (84.9) | 82.6 – 86.9 | – | < 0.001[a] |
| | 0–4 | 114 (88.6) | 81.5 – 93.2 | 1.76 (0.95 – 3.25) | 0.072 |
| | 5–14 | 340 (90.3) | 86.7 – 93.0 | 2.10 (1.39 – 3.19) | < 0.001 |
| | 15–64 | 580 (81.6) | 78.2 – 84.5 | 1.00 | – |
| | 65+ | 12 (58.3) | 32.0 – 80.7 | 0.32 (0.10 – 1.02) | 0.053 |
| Dry/Prod Cough | Overall | 1,046 (87.4) | 85.2 – 89.3 | – | 0.027[a] |
| | 0–4 | 114 (86.0) | 78.4 – 91.2 | 0.68 (0.38 – 1.23) | 0.204 |
| | 5–14 | 340 (83.8) | 79.5 – 87.4 | 0.58 (0.39 – 0.86) | 0.006 |
| | 15–64 | 580 (90.0) | 87.3 – 92.2 | 1.00 | – |
| | 65+ | 12 (75.0) | 46.8 – 91.1 | 0.33 (0.09 – 1.27) | 0.107 |
| Runny Nose | Overall | 1,046 (61.5) | 58.5 – 64.4 | – | < 0.001[a] |
| | 0–4 | 114 (77.2) | 68.7 – 83.9 | 2.16 (1.35 – 3.45) | 0.001 |
| | 5–14 | 340 (57.9) | 52.6 – 63.1 | 0.88 (0.67 – 1.15) | 0.356 |
| | 15–64 | 580 (61.0) | 57.0 – 64.9 | 1.00 | – |
| | 65+ | 12 (33.3) | 13.8 – 60.9 | 0.32 (0.10 – 1.07) | 0.065 |
| Sneezing | Overall | 1,046 (55.1) | 52.0 – 58.1 | – | 0.264[a] |
| | 0–4 | 114 (55.3) | 46.1 – 64.1 | 0.98 (0.65 – 1.46) | 0.906 |
| | 5–14 | 340 (54.7) | 49.4 – 59.9 | 0.95 (0.73 – 1.25) | 0.733 |
| | 15–64 | 580 (55.9) | 51.8 – 59.9 | 1.00 | – |
| | 65+ | 12 (25.0) | 8.9 – 53.2 | 0.26 (0.07 – 0.98) | 0.047 |
| Sore Throat | Overall | 1,046 (73.1) | 70.4 – 75.7 | – | < 0.001[a] |
| | 0–4 | 114 (55.3) | 46.1 – 64.1 | 0.42 (0.28 – 0.63) | < 0.001 |
| | 5–14 | 340 (76.5) | 71.7 – 80.7 | 1.10 (0.81 – 1.51) | 0.538 |
| | 15–64 | 580 (74.7) | 71.0 – 78.0 | 1.00 | – |
| | 65+ | 12 (75.0) | 46.8 – 91.1 | 1.02 (0.27 – 3.81) | 0.978 |
| Short Breath | Overall | 1,046 (30.8) | 28.1 – 33.6 | – | < 0.001[a] |

***Continued on next page***

| **Symptom** | **Age** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| | 0–4 | 114 (16.7) | 10.9 – 24.6 | 0.28 (0.16 – 0.47) | < 0.001 |
| | 5–14 | 340 (17.1) | 13.4 – 21.4 | 0.29 (0.21 – 0.40) | < 0.001 |
| | 15–64 | 580 (41.9) | 37.9 – 46.0 | 1.00 | – |
| | 65+ | 12 (16.7) | 4.7 – 44.8 | 0.28 (0.06 – 1.28) | 0.100 |
| Fatigue | Overall | 1,046 (78.7) | 76.1 – 81.1 | – | 0.256[a] |
| | 0–4 | 114 (73.7) | 64.9 – 80.9 | 0.80 (0.51 – 1.27) | 0.345 |
| | 5–14 | 340 (81.8) | 77.3 – 85.5 | 1.28 (0.91 – 1.80) | 0.149 |
| | 15–64 | 580 (77.8) | 74.2 – 81.0 | 1.00 | – |
| | 65+ | 12 (83.3) | 55.2 – 95.3 | 1.43 (0.31 – 6.61) | 0.647 |
| Musc/Joint Ache | Overall | 1,046 (66.1) | 63.1 – 68.9 | – | < 0.001[a] |
| | 0–4 | 114 (33.3) | 25.3 – 42.4 | 0.15 (0.09 – 0.23) | < 0.001 |
| | 5–14 | 340 (57.1) | 51.7 – 62.2 | 0.39 (0.29 – 0.52) | < 0.001 |
| | 15–64 | 580 (77.4) | 73.8 – 80.6 | 1.00 | – |
| | 65+ | 12 (83.3) | 55.2 – 95.3 | 1.46 (0.32 – 6.74) | 0.629 |
| Loss Appetite | Overall | 1,046 (64.9) | 62.0 – 67.7 | – | 0.262[a] |
| | 0–4 | 114 (62.3) | 53.1 – 70.6 | 0.97 (0.64 – 1.47) | 0.896 |
| | 5–14 | 340 (69.1) | 64.0 – 73.8 | 1.32 (0.99 – 1.75) | 0.058 |
| | 15–64 | 580 (62.9) | 58.9 – 66.8 | 1.00 | – |
| | 65+ | 12 (66.7) | 39.1 – 86.2 | 1.18 (0.35 – 3.96) | 0.791 |
| Nausea/Vomit | Overall | 1,046 (48.6) | 45.5 – 51.6 | – | 0.057[a] |
| | 0–4 | 114 (45.6) | 36.8 – 54.8 | 0.96 (0.64 – 1.44) | 0.854 |
| | 5–14 | 340 (53.8) | 48.5 – 59.1 | 1.34 (1.02 – 1.75) | 0.033 |
| | 15–64 | 580 (46.6) | 42.5 – 50.6 | 1.00 | – |
| | 65+ | 12 (25.0) | 8.9 – 53.2 | 0.38 (0.10 – 1.43) | 0.153 |
| Chills | Overall | 1,046 (59.4) | 56.4 – 62.3 | – | < 0.001[a] |
| | 0–4 | 114 (42.1) | 33.4 – 51.3 | 0.43 (0.29 – 0.65) | < 0.001 |
| | 5–14 | 340 (60.3) | 55.0 – 65.4 | 0.90 (0.68 – 1.19) | 0.458 |
| | 15–64 | 580 (62.8) | 58.8 – 66.6 | 1.00 | – |
| | 65+ | 12 (33.3) | 13.8 – 60.9 | 0.30 (0.09 – 1.00) | 0.049 |
| Headache | Overall | 1,046 (68.5) | 65.6 – 71.2 | – | < 0.001[a] |
| | 0–4 | 114 (32.5) | 24.6 – 41.5 | 0.15 (0.10 – 0.24) | < 0.001 |
| | 5–14 | 340 (68.8) | 63.7 – 73.5 | 0.71 (0.53 – 0.95) | 0.024 |
| | 15–64 | 580 (75.7) | 72.0 – 79.0 | 1.00 | – |
| | 65+ | 12 (50.0) | 25.4 – 74.6 | 0.32 (0.10 – 1.01) | 0.052 |
| Diarrhoea | Overall | 1,046 (24.7) | 22.1 – 27.4 | – | 0.017[a] |
| | 0–4 | 114 (31.6) | 23.8 – 40.6 | 1.27 (0.82 – 1.96) | 0.289 |
| | 5–14 | 340 (19.1) | 15.3 – 23.6 | 0.65 (0.47 – 0.90) | 0.009 |
| | 15–64 | 580 (26.7) | 23.3 – 30.5 | 1.00 | – |

***Continued on next page***

| **Symptom** | **Age** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| | 65+ | 12 (16.7) | 4.7 – 44.8 | 0.55 (0.12 – 2.53) | 0.441 |
| Conjunctivitis | Overall | 1,046 (10.7) | 9.0 – 12.7 | – | 0.029[a] |
| | 0–4 | 114 (18.4) | 12.4 – 26.5 | 2.29 (1.32 – 3.98) | 0.003 |
| | 5–14 | 340 (10.9) | 8.0 – 14.6 | 1.24 (0.79 – 1.93) | 0.343 |
| | 15–64 | 580 (9.0) | 6.9 – 11.6 | 1.00 | – |
| | 65+ | 12 (16.7) | 4.7 – 44.8 | 2.03 (0.43 – 9.52) | 0.369 |
| Nose Bleed | Overall | 1,046 (6.7) | 5.3 – 8.4 | – | 0.753[a] |
| | 0–4 | 114 (4.4) | 1.9 – 9.9 | 0.62 (0.24 – 1.60) | 0.324 |
| | 5–14 | 340 (7.4) | 5.0 – 10.6 | 1.07 (0.64 – 1.80) | 0.794 |
| | 15–64 | 580 (6.9) | 5.1 – 9.3 | 1.00 | – |
| | 65+ | 12 (0.0) | – | – | – |
| Rash | Overall | 1,046 (5.4) | 4.1 – 6.9 | – | 0.012[a] |
| | 0–4 | 114 (12.3) | 7.5 – 19.6 | 3.11 (1.56 – 6.18) | 0.001 |
| | 5–14 | 340 (5.0) | 3.1 – 7.9 | 1.17 (0.62 – 2.20) | 0.629 |
| | 15–64 | 580 (4.3) | 2.9 – 6.3 | 1.00 | – |
| | 65+ | 12 (0.0) | – | – | – |
| Alter Conscious | Overall | 1,046 (7.9) | 6.4 – 9.7 | – | 0.141[a] |
| | 0–4 | 114 (3.5) | 1.4 – 8.7 | 0.36 (0.13 – 1.02) | 0.054 |
| | 5–14 | 340 (7.1) | 4.8 – 10.3 | 0.76 (0.46 – 1.25) | 0.273 |
| | 15–64 | 580 (9.1) | 7.1 – 11.8 | 1.00 | – |
| | 65+ | 12 (16.7) | 4.7 – 44.8 | 1.99 (0.42 – 9.32) | 0.383 |
| Seizures | Overall | 1,046 (1.0) | 0.5 – 1.8 | – | 0.343[a] |
| | 0–4 | 114 (2.6) | 0.9 – 7.5 | 3.89 (0.86 – 17.63) | 0.078 |
| | 5–14 | 340 (0.9) | 0.3 – 2.6 | 1.28 (0.29 – 5.76) | 0.746 |
| | 15–64 | 580 (0.7) | 0.3 – 1.8 | 1.00 | – |
| | 65+ | 12 (0.0) | – | – | – |

**Table 3.4:** Odds ratios and 95% confidence intervals for differences in symptoms reported by age groups. The reference age category was 15–64. $N$ refers to the total number of non-missing symptom responses in each age group and the percentages relate to how many within age groups have a symptom.
**a**: $p$-value for Wald test of the overall variable.

### Differences in Symptoms by Gender

Table 3.5 presents the ORs of reporting each symptom for males compared with females. Note that a separate overall $p$-value was not required here as only one comparison is being made for each symptom. After accounting for Bonferroni corrections, one statistically significant difference in reporting symptoms between males and females was found, which was for nausea/vomiting. Around 44% of men reported nausea/vomiting while around 53% of women reported that symptom (OR 0.69, 95% CI 0.54–0.88). Furthermore, there was also a tendency for males to be less likely to report most of the symptoms as seen by many of the ORs being less the one in Table 3.5, with the main exceptions being cough, diarrhoea and rash.

### Differences in Symptoms by Deprivation

Similarly to when we looked for differences in reported symptoms by age, we derived an overall $p$-value for each symptom using a general test of association when comparing cases that lived in areas belonging to different quintiles of deprivation. However compared with the model for age where we dropped cases with unknown age, those cases that had unknown deprivation were included in the model in their own category as they made up a sizeable amount of all cases that reported symptoms ($n$ = 125, 12%).

After accounting for multiple testing using Bonferonni corrections, there were no statistically significant differences in the symptoms reported by cases from different deprivation quintiles. However, we can note that there was a fairly large difference in the the amount of cases reporting nausea/vomiting from the most deprived quintile compared with the those from deprivation quintile two and also those with unknown deprivation quintile. Approximately 58% of cases from quintile one reported nausea/vomiting while around 43% of cases from quintile two and 38% of cases with unknown deprivation status reported that symptom. Another point to highlight is that for a number of symptoms, the ORs were all less than one when comparing deprivation quintiles 2–5 with the most deprived quintile. Specifically, these symptoms were cough, shortness of breath, nausea/vomiting, diarrhoea, conjunctivitis, rash and altered consciousness. Hence, this gives the general impression that more symptoms were reported by cases living in areas belonging to the 20% most deprived DZs in Scotland.

| **Symptom** | **Gender** | ***N*** **(%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| Fever | F | 531 (86.3) | 83.1 – 88.9 | 1.00 | – |
| | M | 520 (83.5) | 80.0 – 86.4 | 0.80 (0.57 – 1.13) | 0.207 |
| Dry/Prod Cough | F | 531 (86.6) | 83.5 – 89.3 | 1.00 | – |
| | M | 520 (87.9) | 84.8 – 90.4 | 1.12 (0.78 – 1.61) | 0.542 |
| Runny Nose | F | 531 (62.5) | 58.3 – 66.5 | 1.00 | – |
| | M | 520 (60.2) | 55.9 – 64.3 | 0.91 (0.71 – 1.16) | 0.438 |
| Sneezing | F | 531 (58.4) | 54.1 – 62.5 | 1.00 | – |
| | M | 520 (51.3) | 47.1 – 55.6 | 0.75 (0.59 – 0.96) | 0.022 |
| Sore Throat | F | 531 (75.9) | 72.1 – 79.3 | 1.00 | – |
| | M | 520 (70.6) | 66.5 – 74.3 | 0.76 (0.58 – 1.00) | 0.052 |
| Short Breath | F | 531 (34.7) | 30.7 – 38.8 | 1.00 | – |
| | M | 520 (26.7) | 23.1 – 30.7 | 0.69 (0.53 – 0.90) | 0.005 |
| Fatigue | F | 531 (81.0) | 77.4 – 84.1 | 1.00 | – |
| | M | 520 (76.2) | 72.3 – 79.6 | 0.75 (0.56 – 1.01) | 0.057 |
| Musc/Joint Ache | F | 531 (67.8) | 63.7 – 71.6 | 1.00 | – |
| | M | 520 (64.2) | 60.0 – 68.2 | 0.85 (0.66 – 1.10) | 0.222 |
| Loss Appetite | F | 531 (66.3) | 62.2 – 70.2 | 1.00 | – |
| | M | 520 (63.1) | 58.8 – 67.1 | 0.87 (0.67 – 1.12) | 0.276 |
| Nausea/Vomit | F | 531 (52.9) | 48.7 – 57.1 | 1.00 | – |
| | M | 520 (43.7) | 39.5 – 47.9 | 0.69 (0.54 – 0.88) | 0.003 |
| Chills | F | 531 (59.1) | 54.9 – 63.2 | 1.00 | – |
| | M | 520 (59.2) | 55.0 – 63.4 | 1.00 (0.79 – 1.28) | 0.974 |
| Headache | F | 531 (71.8) | 67.8 – 75.4 | 1.00 | – |
| | M | 520 (65.0) | 60.8 – 69.0 | 0.73 (0.56 – 0.95) | 0.019 |
| Diarrhoea | F | 531 (22.0) | 18.7 – 25.8 | 1.00 | – |
| | M | 520 (27.3) | 23.7 – 31.3 | 1.33 (1.00 – 1.76) | 0.048 |
| Conjunctivitis | F | 531 (10.4) | 8.0 – 13.2 | 1.00 | – |
| | M | 520 (11.0) | 8.6 – 13.9 | 1.07 (0.72 – 1.58) | 0.751 |
| Nose Bleed | F | 531 (7.3) | 5.4 – 9.9 | 1.00 | – |
| | M | 520 (6.0) | 4.2 – 8.3 | 0.80 (0.49 – 1.30) | 0.369 |
| Rash | F | 531 (4.9) | 3.4 – 7.1 | 1.00 | – |
| | M | 520 (5.8) | 4.1 – 8.1 | 1.19 (0.69 – 2.04) | 0.529 |
| Alter Conscious | F | 531 (8.1) | 6.1 – 10.7 | 1.00 | – |
| | M | 520 (7.7) | 5.7 – 10.3 | 0.95 (0.60 – 1.48) | 0.807 |
| Seizures | F | 531 (0.9) | 0.4 – 2.2 | 1.00 | – |
| | M | 520 (1.0) | 0.4 – 2.2 | 1.02 (0.29 – 3.55) | 0.973 |

**Table 3.5:** Odds ratios and 95% confidence intervals for differences in symptoms reported by males and females. The reference gender category was females. *N* refers to the total number of non-missing symptom responses for each gender and the percentages relate to how many within a gender have a symptom.

| **Symptom** | **Depriv** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| Fever | Overall | 1,051 (84.9) | 82.6 – 86.9 | – | 0.786[a] |
| | Most Dep | 317 (85.8) | 81.5 – 89.2 | 1.00 | – |
| | 2 | 198 (83.3) | 77.5 – 87.9 | 0.83 (0.51 – 1.35) | 0.447 |
| | 3 | 153 (83.7) | 77.0 – 88.7 | 0.85 (0.50 – 1.44) | 0.541 |
| | 4 | 121 (87.6) | 80.6 – 92.3 | 1.17 (0.63 – 2.19) | 0.625 |
| | 5 | 137 (83.2) | 76.1 – 88.5 | 0.82 (0.47 – 1.42) | 0.478 |
| | Unknown | 125 (85.6) | 78.4 – 90.7 | 0.98 (0.54 – 1.78) | 0.956 |
| Dry/Prod Cough | Overall | 1,051 (87.3) | 85.1 – 89.1 | – | 0.509[a] |
| | Most Dep | 317 (89.0) | 85.0 – 92.0 | 1.00 | – |
| | 2 | 198 (83.8) | 78.1 – 88.3 | 0.64 (0.38 – 1.08) | 0.095 |
| | 3 | 153 (86.3) | 79.9 – 90.8 | 0.78 (0.44 – 1.39) | 0.401 |
| | 4 | 121 (88.4) | 81.5 – 93.0 | 0.95 (0.49 – 1.83) | 0.875 |
| | 5 | 137 (88.3) | 81.9 – 92.7 | 0.94 (0.50 – 1.76) | 0.843 |
| | Unknown | 125 (87.2) | 80.2 – 92.0 | 0.85 (0.45 – 1.59) | 0.602 |
| Runny Nose | Overall | 1,051 (61.4) | 58.4 – 64.3 | – | 0.482[a] |
| | Most Dep | 317 (62.1) | 56.7 – 67.3 | 1.00 | – |
| | 2 | 198 (61.1) | 54.2 – 67.6 | 0.96 (0.66 – 1.38) | 0.814 |
| | 3 | 153 (63.4) | 55.5 – 70.6 | 1.06 (0.71 – 1.57) | 0.792 |
| | 4 | 121 (66.1) | 57.3 – 73.9 | 1.19 (0.77 – 1.84) | 0.441 |
| | 5 | 137 (55.5) | 47.1 – 63.5 | 0.76 (0.51 – 1.14) | 0.183 |
| | Unknown | 125 (59.2) | 50.4 – 67.4 | 0.88 (0.58 – 1.35) | 0.567 |
| Sneezing | Overall | 1,051 (54.9) | 51.9 – 57.9 | – | 0.500[a] |
| | Most Dep | 317 (55.8) | 50.3 – 61.2 | 1.00 | – |
| | 2 | 198 (58.6) | 51.6 – 65.2 | 1.12 (0.78 – 1.60) | 0.540 |
| | 3 | 153 (52.9) | 45.1 – 60.7 | 0.89 (0.60 – 1.31) | 0.555 |
| | 4 | 121 (60.3) | 51.4 – 68.6 | 1.20 (0.79 – 1.84) | 0.396 |
| | 5 | 137 (51.1) | 42.8 – 59.3 | 0.83 (0.55 – 1.23) | 0.352 |
| | Unknown | 125 (48.0) | 39.4 – 56.7 | 0.73 (0.48 – 1.11) | 0.137 |
| Sore Throat | Overall | 1,051 (73.3) | 70.5 – 75.9 | – | 0.581[a] |
| | Most Dep | 317 (73.2) | 68.1 – 77.8 | 1.00 | – |
| | 2 | 198 (70.2) | 63.5 – 76.1 | 0.86 (0.58 – 1.28) | 0.463 |
| | 3 | 153 (70.6) | 62.9 – 77.2 | 0.88 (0.57 – 1.35) | 0.555 |
| | 4 | 121 (75.2) | 66.8 – 82.0 | 1.11 (0.69 – 1.80) | 0.667 |
| | 5 | 137 (77.4) | 69.7 – 83.6 | 1.25 (0.78 – 2.01) | 0.348 |
| | Unknown | 125 (75.2) | 67.0 – 81.9 | 1.11 (0.69 – 1.79) | 0.665 |
| Short Breath | Overall | 1,051 (30.7) | 28.0 – 33.6 | – | 0.058[a] |
| | Most Dep | 317 (36.6) | 31.5 – 42.0 | 1.00 | – |
| | 2 | 198 (29.8) | 23.9 – 36.5 | 0.74 (0.50 – 1.08) | 0.114 |

***Continued on next page***

| **Symptom** | **Depriv** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| | 3 | 153 (26.8) | 20.4 – 34.3 | 0.63 (0.41 – 0.97) | 0.036 |
| | 4 | 121 (29.8) | 22.3 – 38.4 | 0.73 (0.47 – 1.15) | 0.180 |
| | 5 | 137 (24.1) | 17.7 – 31.9 | 0.55 (0.35 – 0.87) | 0.010 |
| | Unknown | 125 (30.4) | 23.0 – 38.9 | 0.76 (0.49 – 1.18) | 0.219 |
| Fatigue | Overall | 1,051 (78.6) | 76.0 – 81.0 | – | 0.042[a] |
| | Most Dep | 317 (80.4) | 75.7 – 84.4 | 1.00 | – |
| | 2 | 198 (70.7) | 64.0 – 76.6 | 0.59 (0.39 – 0.89) | 0.011 |
| | 3 | 153 (79.7) | 72.7 – 85.3 | 0.96 (0.59 – 1.55) | 0.858 |
| | 4 | 121 (81.8) | 74.0 – 87.7 | 1.09 (0.64 – 1.88) | 0.744 |
| | 5 | 137 (82.5) | 75.3 – 87.9 | 1.14 (0.68 – 1.93) | 0.611 |
| | Unknown | 125 (77.6) | 69.5 – 84.0 | 0.84 (0.51 – 1.39) | 0.504 |
| Musc/Joint Ache | Overall | 1,051 (66.0) | 63.1 – 68.8 | – | 0.096[a] |
| | Most Dep | 317 (70.0) | 64.8 – 74.8 | 1.00 | – |
| | 2 | 198 (60.1) | 53.2 – 66.7 | 0.64 (0.44 – 0.94) | 0.021 |
| | 3 | 153 (69.9) | 62.3 – 76.6 | 1.00 (0.65 – 1.52) | 0.983 |
| | 4 | 121 (69.4) | 60.7 – 76.9 | 0.97 (0.62 – 1.53) | 0.901 |
| | 5 | 137 (62.0) | 53.7 – 69.7 | 0.70 (0.46 – 1.07) | 0.096 |
| | Unknown | 125 (61.6) | 52.8 – 69.7 | 0.69 (0.45 – 1.06) | 0.089 |
| Loss Appetite | Overall | 1,051 (64.7) | 61.8 – 67.5 | – | 0.082[a] |
| | Most Dep | 317 (67.5) | 62.2 – 72.4 | 1.00 | – |
| | 2 | 198 (59.6) | 52.6 – 66.2 | 0.71 (0.49 – 1.03) | 0.068 |
| | 3 | 153 (62.7) | 54.9 – 70.0 | 0.81 (0.54 – 1.21) | 0.308 |
| | 4 | 121 (74.4) | 65.9 – 81.3 | 1.40 (0.87 – 2.24) | 0.164 |
| | 5 | 137 (65.7) | 57.4 – 73.1 | 0.92 (0.60 – 1.41) | 0.706 |
| | Unknown | 125 (57.6) | 48.8 – 65.9 | 0.65 (0.43 – 1.00) | 0.050 |
| Nausea/Vomit | Overall | 1,051 (48.3) | 45.3 – 51.4 | – | 0.004[a] |
| | Most Dep | 317 (58.0) | 52.5 – 63.3 | 1.00 | – |
| | 2 | 198 (43.4) | 36.7 – 50.4 | 0.56 (0.39 – 0.79) | 0.001 |
| | 3 | 153 (50.3) | 42.5 – 58.1 | 0.73 (0.50 – 1.08) | 0.115 |
| | 4 | 121 (43.0) | 34.5 – 51.9 | 0.54 (0.36 – 0.83) | 0.005 |
| | 5 | 137 (44.5) | 36.5 – 52.9 | 0.58 (0.39 – 0.87) | 0.008 |
| | Unknown | 125 (38.4) | 30.3 – 47.2 | 0.45 (0.29 – 0.69) | < 0.001 |
| Chills | Overall | 1,051 (59.2) | 56.2 – 62.1 | – | 0.707[a] |
| | Most Dep | 317 (61.8) | 56.4 – 67.0 | 1.00 | – |
| | 2 | 198 (58.6) | 51.6 – 65.2 | 0.87 (0.61 – 1.25) | 0.464 |
| | 3 | 153 (55.6) | 47.6 – 63.2 | 0.77 (0.52 – 1.14) | 0.194 |
| | 4 | 121 (58.7) | 49.8 – 67.1 | 0.88 (0.57 – 1.34) | 0.546 |
| | 5 | 137 (62.0) | 53.7 – 69.7 | 1.01 (0.67 – 1.52) | 0.966 |
| | Unknown | 125 (55.2) | 46.5 – 63.6 | 0.76 (0.50 – 1.16) | 0.201 |

***Continued on next page***

| **Symptom** | **Depriv** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| Headache | Overall | 1,051 (68.4) | 65.5 – 71.2 | – | 0.786[a] |
| | Most Dep | 317 (69.7) | 64.4 – 74.5 | 1.00 | – |
| | 2 | 198 (71.7) | 65.1 – 77.5 | 1.10 (0.74 – 1.63) | 0.628 |
| | 3 | 153 (67.3) | 59.5 – 74.2 | 0.89 (0.59 – 1.35) | 0.599 |
| | 4 | 121 (66.1) | 57.3 – 73.9 | 0.85 (0.54 – 1.32) | 0.468 |
| | 5 | 137 (71.5) | 63.5 – 78.4 | 1.09 (0.70 – 1.70) | 0.698 |
| | Unknown | 125 (60.0) | 51.2 – 68.2 | 0.65 (0.42 – 1.00) | 0.051 |
| Diarrhoea | Overall | 1,051 (24.6) | 22.1 – 27.3 | – | 0.010[a] |
| | Most Dep | 317 (31.5) | 26.7 – 36.9 | 1.00 | – |
| | 2 | 198 (26.8) | 21.1 – 33.3 | 0.79 (0.53 – 1.18) | 0.249 |
| | 3 | 153 (20.9) | 15.2 – 28.0 | 0.57 (0.36 – 0.91) | 0.017 |
| | 4 | 121 (17.4) | 11.6 – 25.1 | 0.46 (0.27 – 0.77) | 0.003 |
| | 5 | 137 (21.2) | 15.2 – 28.7 | 0.58 (0.36 – 0.94) | 0.025 |
| | Unknown | 125 (19.2) | 13.3 – 27.0 | 0.52 (0.31 – 0.85) | 0.010 |
| Conjunctivitis | Overall | 1,051 (10.7) | 8.9 – 12.7 | – | 0.820[a] |
| | Most Dep | 317 (12.6) | 9.4 – 16.7 | 1.00 | – |
| | 2 | 198 (10.1) | 6.6 – 15.1 | 0.78 (0.44 – 1.37) | 0.387 |
| | 3 | 153 (11.1) | 7.1 – 17.1 | 0.87 (0.47 – 1.58) | 0.639 |
| | 4 | 121 (9.1) | 5.2 – 15.5 | 0.69 (0.34 – 1.40) | 0.306 |
| | 5 | 137 (10.2) | 6.2 – 16.4 | 0.79 (0.41 – 1.50) | 0.469 |
| | Unknown | 125 (8.0) | 4.4 – 14.1 | 0.60 (0.29 – 1.24) | 0.171 |
| Nose Bleed | Overall | 1,051 (6.7) | 5.3 – 8.3 | – | 0.083[a] |
| | Most Dep | 317 (5.4) | 3.4 – 8.4 | 1.00 | – |
| | 2 | 198 (6.6) | 3.9 – 10.9 | 1.24 (0.59 – 2.61) | 0.571 |
| | 3 | 153 (11.1) | 7.1 – 17.1 | 2.21 (1.09 – 4.45) | 0.027 |
| | 4 | 121 (3.3) | 1.3 – 8.2 | 0.60 (0.20 – 1.83) | 0.372 |
| | 5 | 137 (8.8) | 5.1 – 14.7 | 1.69 (0.79 – 3.65) | 0.178 |
| | Unknown | 125 (5.6) | 2.7 – 11.1 | 1.05 (0.42 – 2.59) | 0.921 |
| Rash | Overall | 1,051 (5.3) | 4.1 – 6.9 | – | 0.789[a] |
| | Most Dep | 317 (6.3) | 4.1 – 9.5 | 1.00 | – |
| | 2 | 198 (4.0) | 2.1 – 7.8 | 0.63 (0.27 – 1.45) | 0.273 |
| | 3 | 153 (4.6) | 2.2 – 9.1 | 0.71 (0.29 – 1.72) | 0.451 |
| | 4 | 121 (5.8) | 2.8 – 11.5 | 0.91 (0.38 – 2.21) | 0.838 |
| | 5 | 137 (4.4) | 2.0 – 9.2 | 0.68 (0.27 – 1.73) | 0.419 |
| | Unknown | 125 (6.4) | 3.3 – 12.1 | 1.02 (0.44 – 2.37) | 0.972 |
| Alter Conscious | Overall | 1,051 (7.9) | 6.4 – 9.7 | – | 0.284[a] |
| | Most Dep | 317 (10.7) | 7.8 – 14.6 | 1.00 | – |
| | 2 | 198 (5.6) | 3.1 – 9.7 | 0.49 (0.24 – 0.99) | 0.047 |
| | 3 | 153 (7.8) | 4.5 – 13.2 | 0.71 (0.36 – 1.41) | 0.326 |

***Continued on next page***

| **Symptom** | **Depriv** | ***N* (%)** | **95% CI** | **OR (95% CI)** | ***p*** |
|---|---|---|---|---|---|
| | 4 | 121 (6.6) | 3.4 – 12.5 | 0.59 (0.26 – 1.31) | 0.195 |
| | 5 | 137 (7.3) | 4.0 – 12.9 | 0.66 (0.31 – 1.37) | 0.260 |
| | Unknown | 125 (6.4) | 3.3 – 12.1 | 0.57 (0.26 – 1.27) | 0.167 |
| Seizures | Overall | 1,051 (1.0) | 0.5 – 1.7 | – | 0.913[a] |
| | Most Dep | 317 (0.9) | 0.3 – 2.7 | 1.00 | – |
| | 2 | 198 (1.0) | 0.3 – 3.6 | 1.07 (0.18 – 6.45) | 0.943 |
| | 3 | 153 (2.0) | 0.7 – 5.6 | 2.09 (0.42 – 10.49) | 0.369 |
| | 4 | 121 (0.0) | – | – | – |
| | 5 | 137 (1.5) | 0.4 – 5.2 | 1.55 (0.26 – 9.39) | 0.633 |
| | Unknown | 125 (0.0) | – | – | – |

**Table 3.6:** Odds ratios and 95% confidence intervals for differences in symptoms reported by deprivation quintiles. the reference category was the most deprived quintile. $N$ refers to the total number of non-missing symptom responses in each deprivation quintile and the percentages relate to how many within deprivation quintiles have a symptom.
**a**: $p$-value for Wald test of the overall variable.

### Multiple Logistic Regression

We fitted a multivariable logistic regression model with age, gender and deprivation together to investigate how results from our univariable analyses would be affected when also adjusting for the additional covariates – for example, with age, we would examine the effect on results when also adjusting for gender and deprivation. The percentage changes in the estimated ORs from the multivariable model relative the univariable ORs are illustrated in Figure 3.5, and for symptoms that produced statistically significant differences between groups in univariable analyses, the ORs and 95% CIs are reported in Table 3.7.

From Figure 3.5, we can observe that when comparing deprivation quintile two against deprivation quintile one for fever, the OR changed by around 10%, and similarly for loss of appetite, the OR changed by around the same percentage when comparing those aged 65+ against cases aged 15–64. Larger changes in ORs were found for seizures, but this was perhaps expected as only 10 cases reported seizures and even minor adjustments to a model could have a larger impact on the ORs for that symptom. For all of the other symptoms, the ORs from the

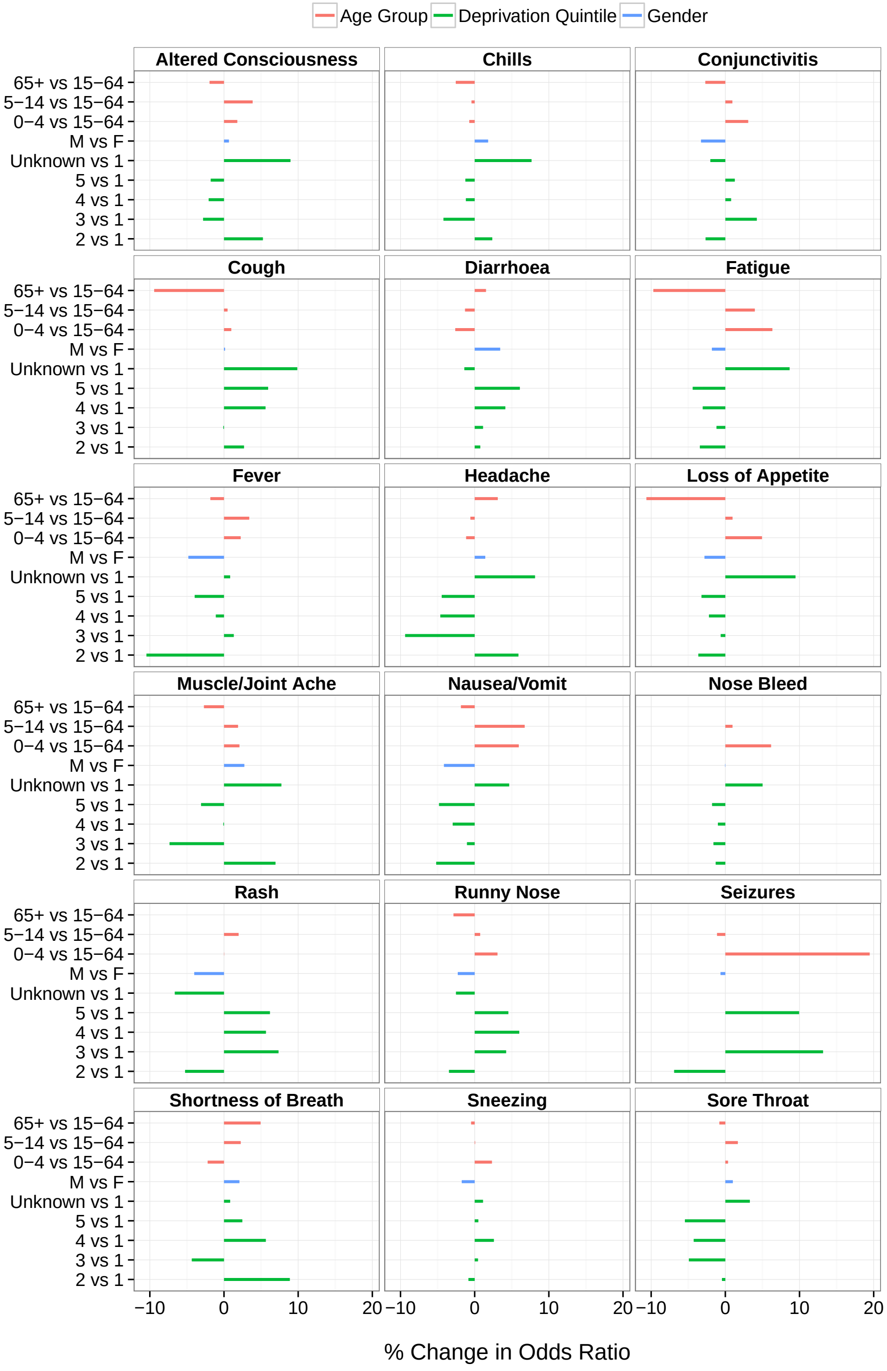


**Figure 3.5:** Percentage change in odds ratios for multiple logistic regression on symptoms against age, gender and deprivation compared with univariable regression. The change was calculated relative to the univariable odds ratio.

| | OR (95% CI) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Fever** | **Runny Nose** | **Sore Throat** | **Short Breath** | **Musc/Joint Ache** | **Nausea/Vomit** | **Chills** | **Headache** |
| *Age* | | | | | | | | |
| 0-4 | 1.80 | 2.23 | 0.42 | 0.27 | 0.15 | 1.02 | 0.43 | 0.15 |
| | (0.97 – 3.34) | (1.39 – 3.57) | (0.28 – 0.64) | (0.16 – 0.46) | (0.10 – 0.23) | (0.68 – 1.54) | (0.28 – 0.65) | (0.10 – 0.24) |
| 5-14 | 2.18 | 0.89 | 1.12 | 0.29 | 0.40 | 1.43 | 0.90 | 0.70 |
| | (1.43 – 3.31) | (0.67 – 1.17) | (0.82 – 1.54) | (0.21 – 0.41) | (0.29 – 0.53) | (1.09 – 1.88) | (0.68 – 1.18) | (0.52 – 0.95) |
| 15-64 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 65+ | 0.31 | 0.31 | 1.01 | 0.29 | 1.42 | 0.38 | 0.29 | 0.33 |
| | (0.10 – 1.01) | (0.09 – 1.05) | (0.27 – 3.82) | (0.06 – 1.36) | (0.30 – 6.61) | (0.10 – 1.43) | (0.09 – 0.98) | (0.10 – 1.06) |
| *Gender* | | | | | | | | |
| Female | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Male | 0.77 | 0.89 | 0.77 | 0.70 | 0.88 | 0.66 | 1.02 | 0.74 |
| | (0.54 – 1.08) | (0.69 – 1.14) | (0.58 – 1.02) | (0.53 – 0.93) | (0.67 – 1.15) | (0.52 – 0.85) | (0.80 – 1.31) | (0.56 – 0.98) |
| *Deprivation* | | | | | | | | |
| Most Deprived | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2 | 0.74 | 0.92 | 0.86 | 0.80 | 0.69 | 0.53 | 0.89 | 1.17 |
| | (0.45 – 1.22) | (0.64 – 1.34) | (0.58 – 1.28) | (0.54 – 1.19) | (0.46 – 1.02) | (0.37 – 0.76) | (0.62 – 1.29) | (0.77 – 1.76) |
| 3 | 0.86 | 1.10 | 0.84 | 0.61 | 0.92 | 0.72 | 0.74 | 0.81 |
| | (0.50 – 1.47) | (0.73 – 1.65) | (0.54 – 1.29) | (0.39 – 0.94) | (0.59 – 1.43) | (0.49 – 1.07) | (0.50 – 1.10) | (0.53 – 1.25) |
| 4 | 1.16 | 1.26 | 1.06 | 0.78 | 0.97 | 0.53 | 0.87 | 0.81 |
| | (0.61 – 2.18) | (0.81 – 1.97) | (0.65 – 1.73) | (0.48 – 1.24) | (0.60 – 1.57) | (0.34 – 0.81) | (0.56 – 1.33) | (0.51 – 1.29) |
| 5 | 0.79 | 0.79 | 1.18 | 0.56 | 0.68 | 0.55 | 1.00 | 1.04 |
| | (0.45 – 1.38) | (0.53 – 1.20) | (0.74 – 1.91) | (0.35 – 0.90) | (0.44 – 1.05) | (0.37 – 0.83) | (0.66 – 1.51) | (0.66 – 1.65) |
| Unknown | 0.99 | 0.86 | 1.15 | 0.76 | 0.74 | 0.47 | 0.82 | 0.70 |
| | (0.54 – 1.82) | (0.56 – 1.33) | (0.71 – 1.87) | (0.48 – 1.22) | (0.46 – 1.18) | (0.31 – 0.73) | (0.53 – 1.26) | (0.44 – 1.12) |

**Table 3.7:** Odds ratios and 95% confidence intervals for differences in symptoms reported by age groups, gender and deprivation. Only symptoms which produced statistically significant differences in univariable analyses are shown. The reference categories were age group 15–64, female gender and the most deprived deprivation quintile.

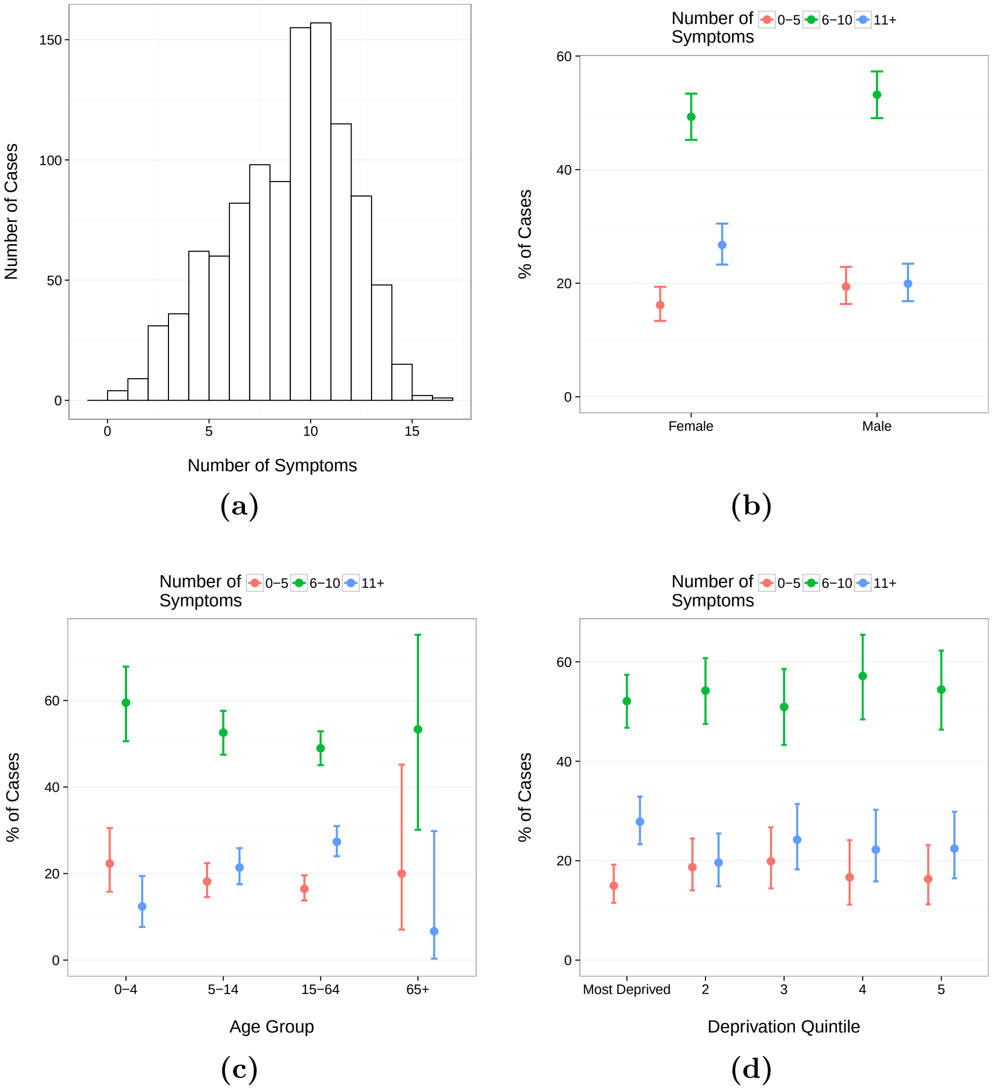
multivariable model changed by less than 10% for comparisons between age, gender and deprivation. Therefore on the whole, adjusting for multiple factors as in the multivariable model only had a minimal impact on the ORs for reporting each of the symptoms.

### Number of Symptoms

For our analysis on the numbers of symptoms reported by cases – out of a possible total of 18 symptoms for each case – we looked at only the 702 cases that returned complete data on symptoms along with the 349 cases that returned partial data and subsequently had "No" responses imputed. Out of these cases, the number of symptoms reported ranged from 0–16 but the histogram (Figure 3.6a) shows that the majority of cases reported between 6–12 symptoms; the median number of symptoms reported by patients was 9 (IQR 6–11 symptoms). Only four cases (0.4%) reported no symptoms and these can be considered as reported asymptomatic cases.

We wanted to look at whether or not there were differences in numbers of reported symptoms between gender, age group and deprivation quintile, with a hypothesis that more deprived groups would report more symptoms. To visually explore this, numbers of symptoms were grouped into 0–5, 6–10 and 11+ and error-bar plots were produced using these categorisations. The clearest difference between males and females that can be seen on Figure 3.6b is that a higher proportion of females reported 11+ symptoms than males (29% vs 21.5%). The difference in numbers of symptoms reported between males and females was found to be statistically significant using a Mann-Whitney test ($p = 0.011$, Table 3.10).

From Figure 3.6c, the clearest differences between age groups lie with the lower proportions of cases reporting 11+ symptoms in age groups 0–4 (13.2%) and 65+ (8.3%), compared with 23.2% and 29.5% for age groups 5–14 and 15–64, respectively. Note that for those aged 65+, the error bars around estimated proportions were wide due to the small number of cases in that age category. Moreover, a Kruskal-Wallis test found that there was a significant difference in the numbers of



(a)

(b)

(c)

(d)

**Figure 3.6:** Numbers of symptoms reported by early pH1N1 cases overall and by gender, age group and deprivation quintile. Percentages were taken from only cases that reported either complete or partial data on symptoms ($n$ = 1,051). Error bars around the estimated proportions were calculated using Wilson's method [388]. Note that the $y$-axis scale varies between plots.

symptoms reported between age groups ($p < 0.001$). Post-hoc pairwise comparisons using Dunn's test [94] with Bonferonni adjustments for significance found that the main difference was between age groups 0–4 and 15–64 (as expected from the percentages illustrated in Figure 3.6c). No clear trend in the numbers of symptoms by deprivation quintile can be seen from Figure 3.6d, and a linear-by-linear test did not indicate that there was any evidence of a linear association.

### Clustering of Symptoms

PCA was used on the data for the 18 symptoms in order to find if different groups of these symptoms clustered among the cases. From the scatterplot of the first two PCs (which account for around a third of the total variance in the data) shown at the bottom of Figure 3.7, we can observe that the procedure split the cases into three main clusters. Moreover, the biplot (top of Figure 3.7), illustrates that the symptoms that most discriminate the groups on the second PC are sneezing and runnynose. This point is also clear from Table 3.8, which clearly shows that 100% of cases in PCA group 1 (bottom group on the plot) reported those two symptoms while 0% of cases in PCA group 3 (top group on the plot) reported those symptoms.

The cases were split into groups that were roughly indicative of their severity of illness, with highest severity in group 1 and lowest severity in group 3. This statement becomes clearer by deriving the average numbers of reported symptoms for cases in each group – a mean of 10 reported symptoms was found for group 1, followed by a mean of 8 for group 2 and finally, a mean of 6 for group 3. Furthermore, the percentages of cases in each group reporting each of the symptoms generally decreased from group 1 to group 2 and then group 3 (Table 3.8). The only exceptions to this were rarely reported symptoms such as altered consciousness and seizures which were highest in group 1 but lower in group 2 than in group 3.

In the most severe illness group, the 10 symptoms that most cases reported were chills, cough, fatigue, fever, headache, loss of appetite, muscle/joint ache, runny nose, sneezing and sore throat. Over 65% of cases in group 1 reported each of these symptoms. For cases in group 2 that reported a mean of 8 symptoms, the most common symptoms were the same as for group 1 but perhaps excluding chills and sneezing, which under 40% of cases in group 2 reported. Finally, for cases in group 3, the most commonly reported symptoms were the same as for group 2 but with the exceptions of loss of appetite and runny nose.

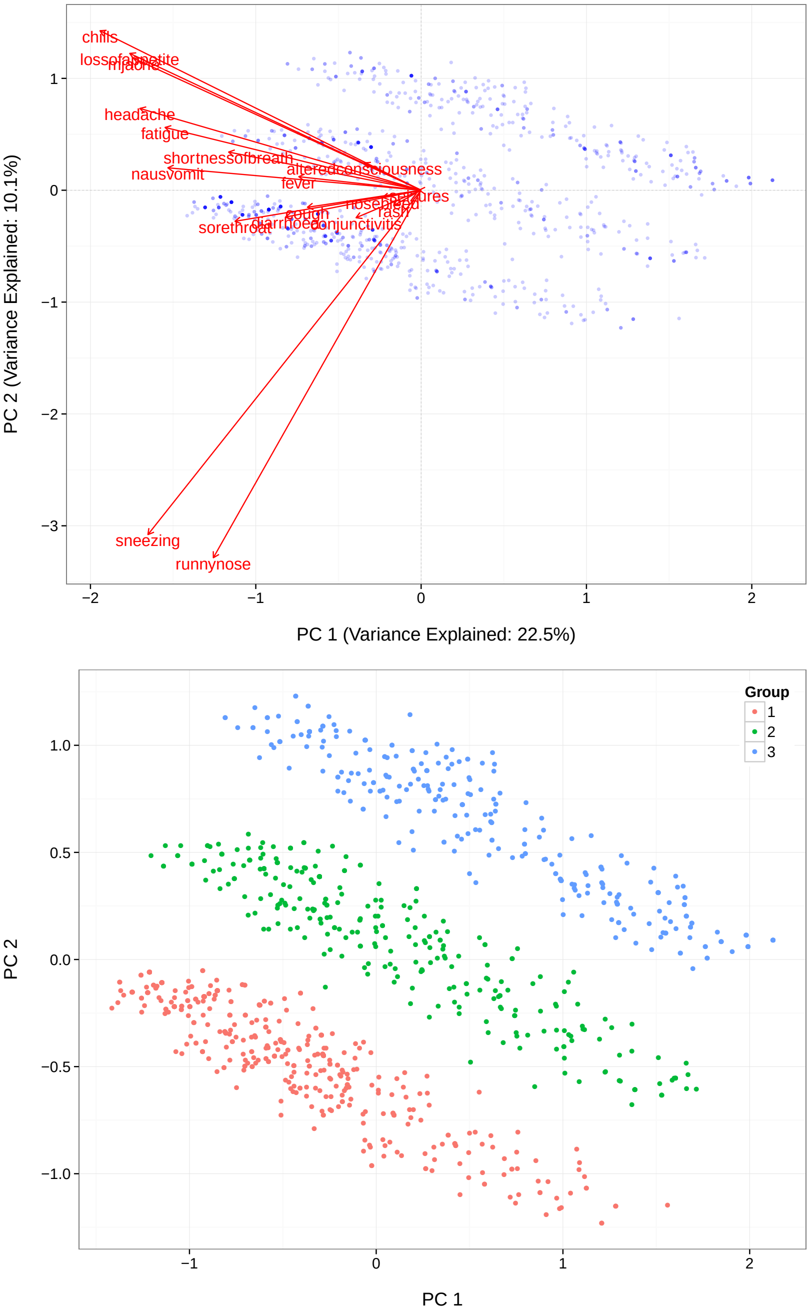


**Figure 3.7:** Principal components analysis plots for the pH1N1 symptoms data. At the top, the biplot is shown and at the bottom, a scatterplot for the first two principal components is shown with colours illustrating how groups were split up. Runny nose and sneezing were the symptoms that most discriminate the groups; everybody in group one reported those symptoms and nobody in group three reported those symptoms.

| | Cases (%) | | |
|---|---|---|---|
| **Symptom** | **PCA Group 1** | **PCA Group 2** | **PCA Group 3** |
| Altered Consciousness | 41 (8.9) | 20 (6.6) | 22 (7.6) |
| Chills | 307 (66.7) | 174 (57.8) | 141 (48.6) |
| Conjunctivitis | 70 (15.2) | 29 (9.6) | 13 (4.5) |
| Cough | 424 (92.2) | 269 (89.4) | 224 (77.2) |
| Diarrhoea | 139 (30.2) | 73 (24.3) | 47 (16.2) |
| Fatigue | 404 (87.8) | 228 (75.7) | 194 (66.9) |
| Fever | 411 (89.3) | 259 (86.0) | 222 (76.6) |
| Headache | 356 (77.4) | 198 (65.8) | 165 (56.9) |
| Loss of Appetite | 320 (69.6) | 208 (69.1) | 152 (52.4) |
| Muscle/Joint Ache | 340 (73.9) | 182 (60.5) | 172 (59.3) |
| Nausea/Vomiting | 261 (56.7) | 139 (46.2) | 108 (37.2) |
| Nose Bleed | 43 (9.3) | 15 (5.0) | 12 (4.1) |
| Rash | 36 (7.8) | 13 (4.3) | 7 (2.4) |
| Runny Nose | 460 (100.0) | 185 (61.5) | 0 (0.0) |
| Seizures | 5 (1.1) | 2 (0.7) | 3 (1.0) |
| Shortness of Breath | 174 (37.8) | 80 (26.6) | 69 (23.8) |
| Sneezing | 460 (100.0) | 117 (38.9) | 0 (0.0) |
| Sore Throat | 385 (83.7) | 202 (67.1) | 183 (63.1) |

**Table 3.8:** Numbers and percentages of pH1N1 cases reporting each of the 18 symptoms within three groups produced from principal components analysis (PCA). See Figure 3.7 for an illustration of the groups chosen following PCA.

### 3.4.3 Treatment

For exploring treatment, we looked at all 1,138 laboratory-confirmed cases and assumed unknown responses to be “No” when calculating proportions. This applies to the analyses on health services, antivirals but not treatment delay as for that analysis, we could only use cases that received antivirals and also reported both, a date for receiving antivirals and a date of infection (date of onset of symptoms or date of report or date of confirmation).

Cases most often reported that they telephoned or visited their GP (58.2%) while NHS Direct and other health services were the next most commonly used services with 25.4% and 26.4% of cases using these services respectively (Figure 3.8). This was followed by visiting walk-in centres which were used by only 8.6% of cases. Less than 5.5% of cases used the remaining services which were contacting out of hours services, attending accident and emergency departments and being admitted to hospitals. Note that cases can attend multiple health services and therefore, the percentages on Figure 3.8 do not add up to 100%.

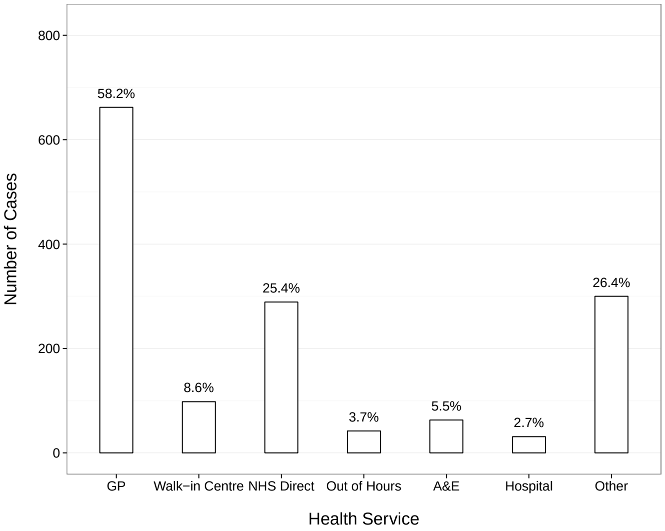


**Figure 3.8:** Health services used by pH1N1 cases with percentages calculated from the 1,138 laboratory-confirmed early pH1N1 cases in Scotland. Note that cases can attend multiple health services and therefore, the percentages do not add up to 100%.

Out of the cases that reported the use of multiple health services, 461 (40.5%) used two health services, 88 (7.7%) used three health services and only 10 cases (under 1%) used four health services; nobody reported the use of more than four different health services. For those that used at least two different health services, Table 3.9 shows how many cases used different pairs of health services (noting that the numbers along the diagonal of the table give how many cases reported the use of each individual health service). The most commonly reported pairs of health services used were GP with other health services ($n = 178$), NHS Direct with other health services ($n = 125$), GP with NHS Direct ($n = 118$), and GP with walk-in centres ($n = 57$).

| | **GP** | **Walk-in Centre** | **A&E** | **Hospital** | **NHS Direct** | **Other** | **Out of Hours** |
|---|---|---|---|---|---|---|---|
| **GP** | 662 | | | | | | |
| **Walk-in Centre** | 57 | 98 | | | | | |
| **A&E** | 23 | 4 | 63 | | | | |
| **Hospital** | 12 | 0 | 7 | 31 | | | |
| **NHS Direct** | 118 | 41 | 19 | 11 | 289 | | |
| **Other** | 178 | 6 | 9 | 4 | 125 | 300 | |
| **Out of Hours** | 13 | 5 | 3 | 5 | 22 | 5 | 42 |

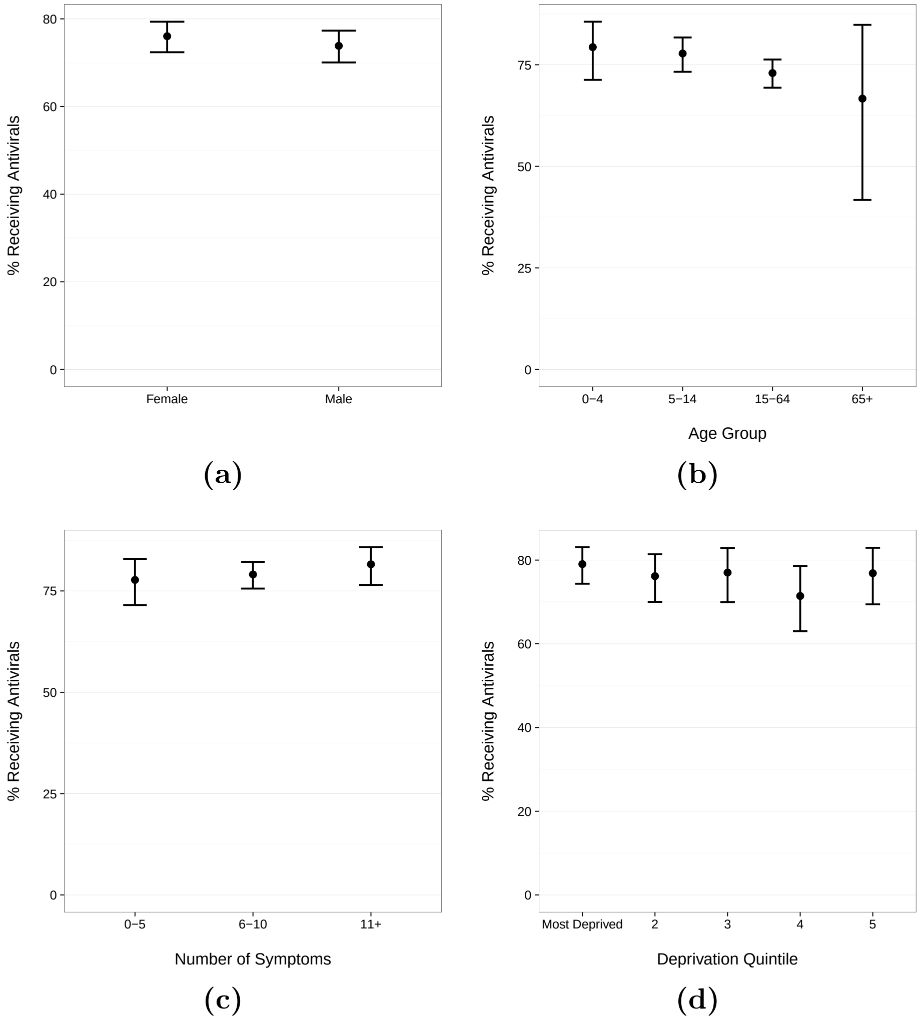
**Table 3.9:** Numbers of pH1N1 cases using at least two different health services. Note that the numbers on the diagonal give the use of a single health service.

From those that used at least three health services, 40 cases used GP with NHS Direct and other health services, and 13 used GP with NHS Direct and walk-in centres; less than 10 cases used any of the remaining three-way combinations of health services. It was rare for cases to use four different health services but when this happened, it was usually due to cases needing to be admitted to hospital or an accident and emergency department. For instance, three cases reported the use of GP with NHS Direct, walk-in centres and visiting the accident and emergency department.

### Antivirals

Almost three quarters of cases (74.3%) received oseltamivir (trade name: Tamiflu) as their antiviral whilst only 0.6% of cases received Zanamivir (Relenza). No cases received Amantadine or Rimantadine, and 0.3% of cases received other antivirals. This gives an overall figure of 75.2% of cases that received antiviral drugs.

For antiviral use, we hypothesised that antiviral use would be higher for those reporting more symptoms and would be lower for those from more deprived areas. We also wished to find if there were differences by gender and age but did not expect any linear trend in antiviral use with age. Figure 3.9 shows the percentages of cases receiving antivirals by gender, age group, number of symptoms reported and deprivation quintile. Figure 3.9b shows decreasing percentages of cases receiving antivirals by age group and from, Figure 3.9c, there appears to be a slight trend of higher antiviral use when cases have more symptoms.



**(a)** **(b)**

**(c)** **(d)**

**Figure 3.9:** Percentages of early pH1N1 cases receiving antivirals by gender, age group, numbers of symptoms and deprivation quintile. Error bars around the estimated proportions were calculated using Wilson's method [388]. Note that the $y$-axis scale varies between plots.

To more formally test for differences, a $\chi^2$ test was used to test for a difference in the proportions of cases receiving antivirals by gender, and the same test was also used to test for an association between receiving antivirals with age group. Linear-by-linear tests were used to test if there was a linear association between receiving antivirals with numbers of symptoms and also with deprivation quintile.

None of these tests indicated that there were significant differences or associations (see Table 3.10).

A previous review by Ward et al. [377] on the use of oseltamivir against pandemic influenza found that the most frequently reported adverse effects of using the drug were nausea, vomiting and diarrhoea. To investigate if cases that received oseltamivir were more likely to report these symptoms compared with cases that did not receive oseltamivir, we used only the 992 cases that reported at least one symptom and also reported on whether or not they received oseltamivir. We found that cases were no more likely to report diarrhoea (OR 1.13, 95% CI 0.77 – 1.7) or nausea/vomiting (OR 1.23, 95% CI 0.88 – 1.73) based on if they received oseltamivir.

### Treatment Delay

The distribution of treatment delay times is given in Figure 3.10. From those cases for which a treatment delay length could be derived ($n = 778$), 37.7% received antivirals within the recommended 48 hours, and around half received treatment within 3–6 days of having symptoms (50.4%). Therefore, around 88% cases of received antivirals within a week of being symptomatic, but 12% of cases had to wait a week or more to receive antivirals.

For treatment delay, we wanted to test the same hypotheses as for antiviral use. To visually explore differences in treatment delay by gender, age group, number of symptoms and deprivation quintile, treatment delay was grouped into 0–2, 3–6 and 7+ days and the error-bar plots shown in Figure 3.11 were created. The plots do not show any obvious differences in treatment delay by these variables. Mann-Whitney and Kruskal-Wallis tests were used to test if there were significant differences between treatment delay and gender, and between treatment delay and age group; linear-by-linear tests were used to test for linear associations between treatment delay with numbers of symptoms and with deprivation quintile. These tests did not produce any significant results (summary results are shown in Table 3.10).

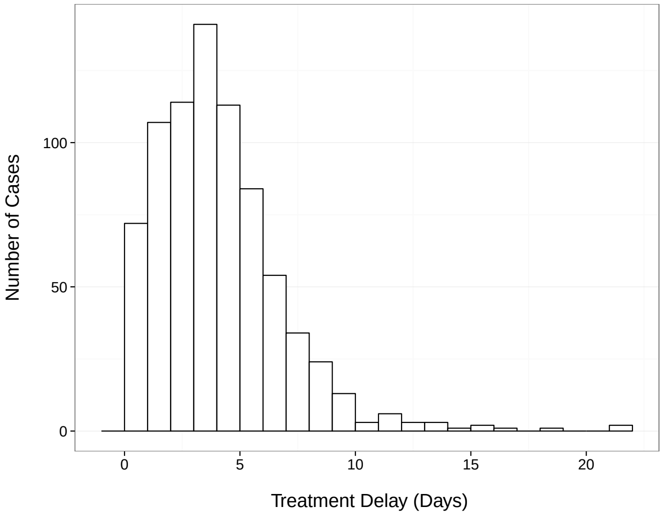


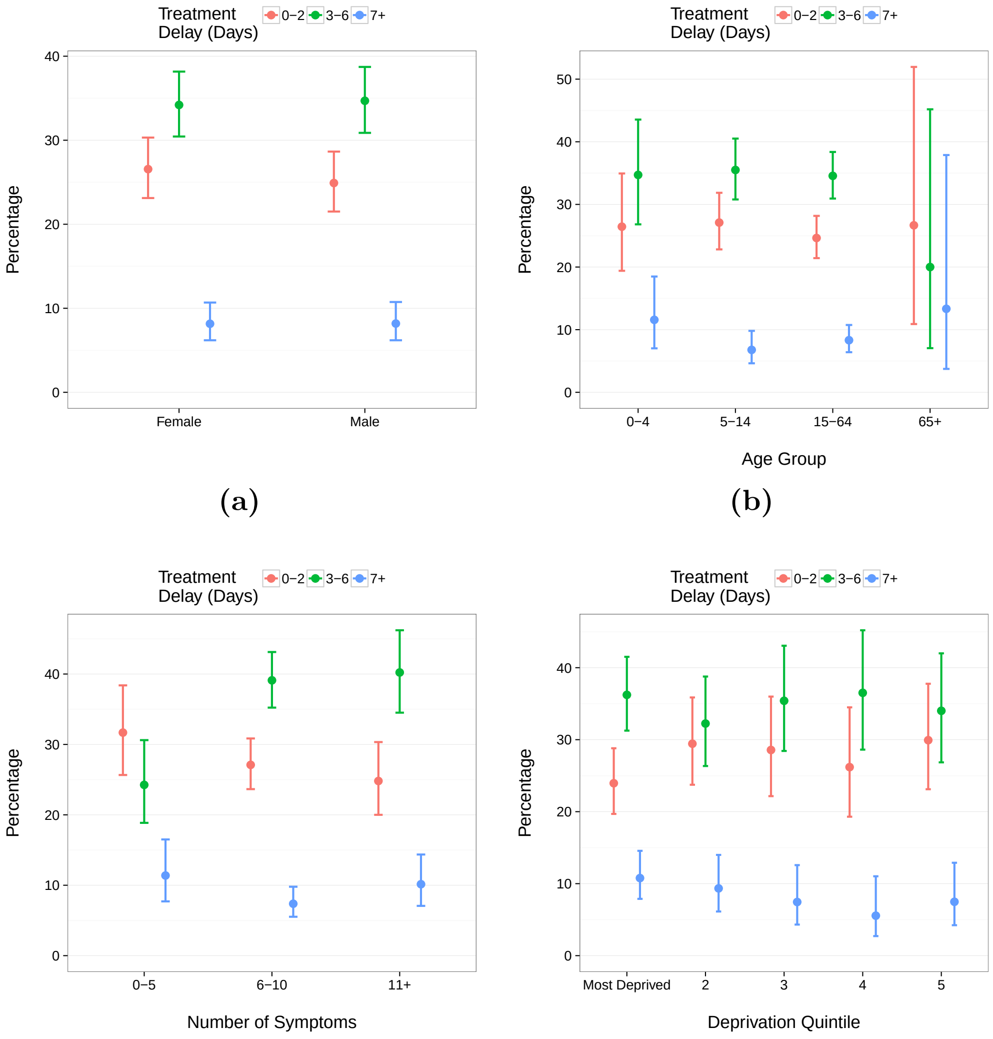
**Figure 3.10:** Time between onset of symptoms and receiving antivirals for early pH1N1 cases in Scotland. Note that this only includes cases for which the dates of receiving antivirals and symptoms onset were known ($n = 778$).

| **Dependent Variable** | **Independent Variable** | **Test** | ***p*** |
|---|---|---|---|
| Number of Symptoms | Gender | Mann-Whitney | 0.011 |
| | Age Group | Kruskal-Wallis | $< 0.001$ |
| | Deprivation Quintile | Linear-by-Linear | 0.283 |
| Received Antivirals | Gender | Chi-Square | 0.431 |
| | Age Group | Chi-Square | 0.197 |
| | Number of Symptoms | Linear-by-Linear | 0.293 |
| | Deprivation Quintile | Linear-by-Linear | 0.289 |
| Treatment Delay | Gender | Mann-Whitney | 0.962 |
| | Age Group | Kruskal-Wallis | 0.703 |
| | Number of Symptoms | Linear-by-Linear | 0.136 |
| | Deprivation Quintile | Linear-by-Linear | 0.121 |

**Table 3.10:** Summary statistical test results for analyses on numbers of symptoms, receiving antivirals and treatment delay. Note that for linear-by-linear tests, number of symptoms was grouped into 0–5, 6–10 and 11+, and treatment delay was grouped into 0–2, 3–6 and 7+ days.

### 3.4.4 Chronic Diseases

Out of the 1,138 pH1N1 cases in our sample, 98 cases (8.6%) had at least one associated chronic disease, with 17 cases having more than one associated chronic



(a)

(b)

(c)

(d)

**Figure 3.11:** Percentages of early pH1N1 cases waiting 0–2 days, 3–6 days and over a week for treatment by gender, age group, numbers of symptoms and deprivation quintile. Error bars around the estimated proportions were calculated using Wilson's method [388]. Note that the $y$-axis scale varies between plots.

disease (Figure 3.12b). From those that had at least one chronic disease, by far the most prevalent disease was lung disease (37 cases out of 98 with at least one chronic disease, 37.8%). This was followed by diabetes, (16 cases), kidney disease (15 cases), chronic heart disease (13 cases), and seizure disorder (11 cases). Less than 10 cases had malignancy, HIV, liver disease and only 4 cases had previously had an organ or bone marrow transplant (Figure 3.12a).

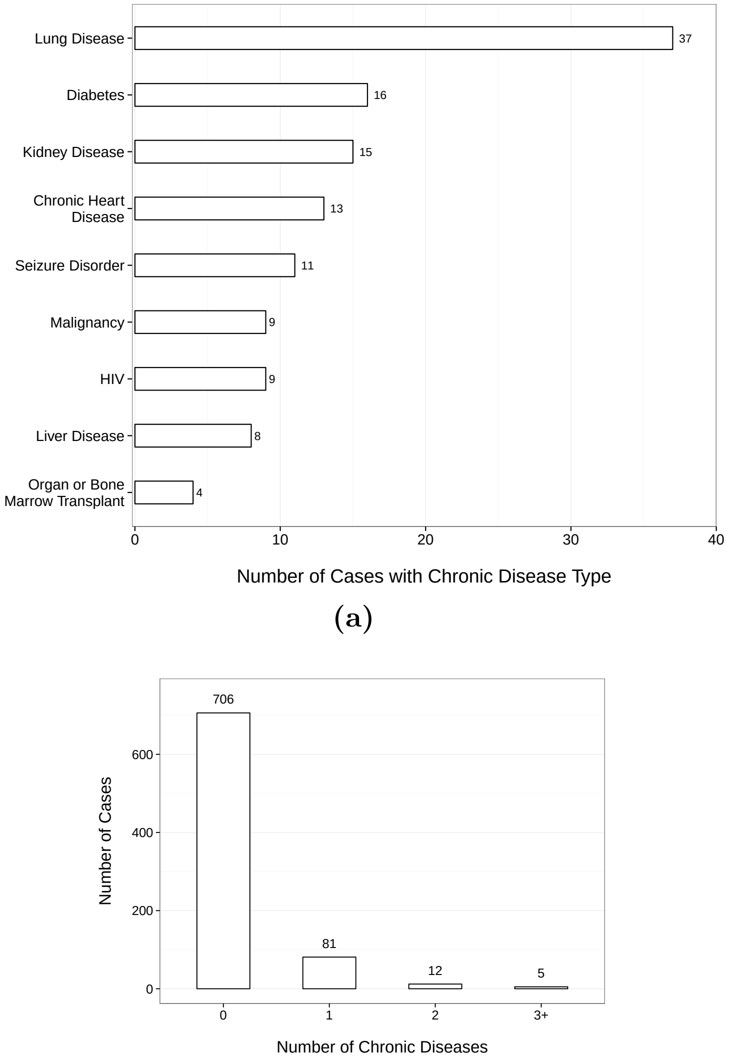
The figure of approximately 8.6% having a chronic disease is lower than the roughly 15% of the Scottish population aged under 65 who have a chronic disease [225] and hence, our results here do not suggest that those with chronic diseases are much more likely to report illness with pH1N1. However, we must note that we assumed that when cases left a response blank regarding chronic diseases, this meant that they did not have that particular disease and this may have biased the estimated percentage downwards. Under the same assumptions, we also investigated if cases with at least one chronic disease were more likely to visit accident and emergency or be admitted to hospital compared with cases with no associated chronic diseases. There was insufficient evidence to suggest that cases with chronic diseases were more likely to visit those health services (OR 1.6, 95% CI: 0.82 to 3.13).

#### Pregnancies

Only 11 of the early pandemic influenza cases were pregnant when they contracted the disease which is not a large enough number of cases to draw any firm conclusions from regarding the severity of their illness or their treatment compared with the general population. However we can note the median number of symptoms reported by these women was fairly consistent with other cases (median of 10 symptoms for the pregnant cases and 9 for all cases). Furthermore, only one of the women received antivirals later than a week after their onset of symptoms (although data was unavailable on whether or not two pregnant women received antivirals), but this person was also admitted to the accident and emergency department soon after they were symptomatic and so she may have been receiving treatment in hospital before being prescribed antivirals.

### 3.4.5 Contacts

There was data on 3,514 contacts of pH1N1 cases in Scotland, that linked to 848 different cases in the cases dataset. Table 3.11 gives an overview of the information available on the contacts that linked to pH1N1 cases. There were slightly more female contacts than male contacts (49.4% vs 46.8%, with the remainder of con-



**(a)**

**(b)**

**Figure 3.12:** Associated chronic diseases for pH1N1 cases.

tacts having unknown gender). Most contacts were aged 15–64 (57.8%) or 5–14 (15.1%) but relatively few were aged 0–4 or 65+, which matches closely to the ages of cases. Deprivation was unknown for over half of contacts but it is still apparent that far more contacts were from more deprived areas, with around a quarter of contacts coming from the two most deprived quintiles, and this was around 56.7% when only considering contacts for which deprivation status was known.

Approximately 78% of the contacts could be grouped into household or non-

| **Variable** | **Group** | ***n* (%)** |
|---|---|---|
| Gender | Female | 1,737 (49.43) |
| | Male | 1,643 (46.76) |
| | Unknown | 134 (3.81) |
| Age Group | 0–4 | 259 (7.37) |
| | 5–14 | 530 (15.08) |
| | 15–64 | 2,030 (57.77) |
| | 65+ | 83 (2.36) |
| | Unknown | 612 (17.42) |
| Deprivation Quintile | Most Deprived | 519 (14.77) |
| | 2 | 365 (10.39) |
| | 3 | 268 (7.63) |
| | 4 | 175 (4.98) |
| | 5 | 232 (6.60) |
| | Unknown | 1,955 (55.63) |
| Type of Contact | Household | 2,036 (57.94) |
| | Non–household | 703 (20.01) |
| | Unknown | 775 (22.05) |
| Symptoms | No | 748 (21.29) |
| | Yes | 302 (8.59) |
| | Unknown | 2,464 (70.12) |

**Table 3.11:** Characteristics of contacts of 2009 pH1N1 cases in Scotland. Only contacts that could be linked to cases in the cases dataset were included.

household contacts using the assumptions we mentioned in Section 3.2.3. Out of all linked contacts, 57% were household and 20% were non-household. Notably, deprivation status for contacts was known much more frequently for household contacts compared with non-household contacts (56% vs 34%). Data regarding whether or not contacts reported being symptomatic was largely missing (70% missing). Thus, that variable could not really be used for any meaningful analyses.

To investigate if there was a bias in cases with certain characteristics reporting contacts, we constructed a predictive model using logistic regression where reporting or not reporting contacts was the binary response variable. The covariates included in the prediction model are shown in Table 3.12. The time period for when a case had onset of symptoms (or another date such as date of illness reported if that was unavailable) was included in the model to see if that had an influence on whether or not cases had contacts. This variable was split into three periods to roughly encompass the beginning, middle and latter part of the containment phase.

In general, the results indicate that the contacts appeared to missing for cases in a fairly random manner (or at least unpredictable manner using the available data and model). However, we note that one OR in the model was statistically significant but this result simply tells us that when a case was missing all responses regarding symptoms, they were also much less likely to report a contact (OR 0.1, 95% CI 0.05–0.19). This result is also vividly illustrated in Figure 3.13 (centre-right plot) where we can observe that over 70% of cases with an unknown number of symptoms had no contacts.

### Case-Contact Relationships

The distribution of the number of contacts per case can be seen in Figure 3.14, and from there, we can see that the distribution is highly skewed with only a small number of cases reported a large number of contacts. The median number of contacts per case was 3 with an IQR of 2–5 contacts, but notably, one case reported 54 contacts, which was by far the most contacts reported by any case (the next highest was 26 contacts). The plots in Figure 3.13 also allow us to explore if cases with different characteristics were reporting different numbers of contacts. There were no clear differences by gender, by deprivation or by numbers of chronic diseases, and for number of symptoms, the only clear difference can be seen when comparing those with an unknown number of symptoms as mentioned previously. However, there was some variation in numbers of contacts by age group. A kruskal-Wallis test on numbers of contacts against age group produced a significant result and pairwise comparisons using Dunn's Tests showed a number of differences by age. Cases aged 0–4 had significantly less contacts than those aged 5–14 but significantly more contacts than those aged 65+, and cases aged 65+ also had significantly less contacts than cases aged 15–64.

| | **OR** | **95% CI** |
|---|---|---|
| *Time Period* | | |
| Before 25 May 2009 | 1.00 | – |
| 25 May to 21 June 2009 | 2.26 | 0.64 – 8.02 |
| After 21 June 2009 | 1.74 | 0.48 – 6.28 |
| Unknown | 2.02 | 0.57 – 7.23 |
| *Symptoms* | | |
| 0–5 | 1.00 | – |
| 6–10 | 1.12 | 0.75 – 1.66 |
| 11+ | 1.12 | 0.71 – 1.76 |
| Unknown | 0.10 | 0.05 – 0.19 |
| *Chronic Disease* | | |
| 0 | 1.00 | – |
| 1+ | 0.72 | 0.44 – 1.17 |
| Unknown | 1.08 | 0.75 – 1.56 |
| *Gender* | | |
| Female | 1.00 | – |
| Male | 1.03 | 0.77 – 1.37 |
| *Age* | | |
| 0–4 | 1.00 | – |
| 5–14 | 1.35 | 0.98 – 1.87 |
| 15–64 | 1.00 | – |
| 65+ | 0.42 | 0.14 – 1.29 |
| Unknown | 0.65 | 0.12 – 3.40 |
| *Deprivation Quintile* | | |
| 1 (Most Deprived) | 1.00 | – |
| 2 | 1.28 | 0.84 – 1.96 |
| 3 | 1.34 | 0.84 – 2.15 |
| 4 | 1.00 | – |
| 5 | 1.36 | 0.84 – 2.22 |
| Unknown | 1.11 | 0.69 – 1.79 |

**Table 3.12:** Odds ratios and 95% confidence intervals from the predictive model for which cases had at least one reported contact. Note that the time period refers to the earliest known date between the date of onset of symptoms, the date reported and the date the case was confirmed.

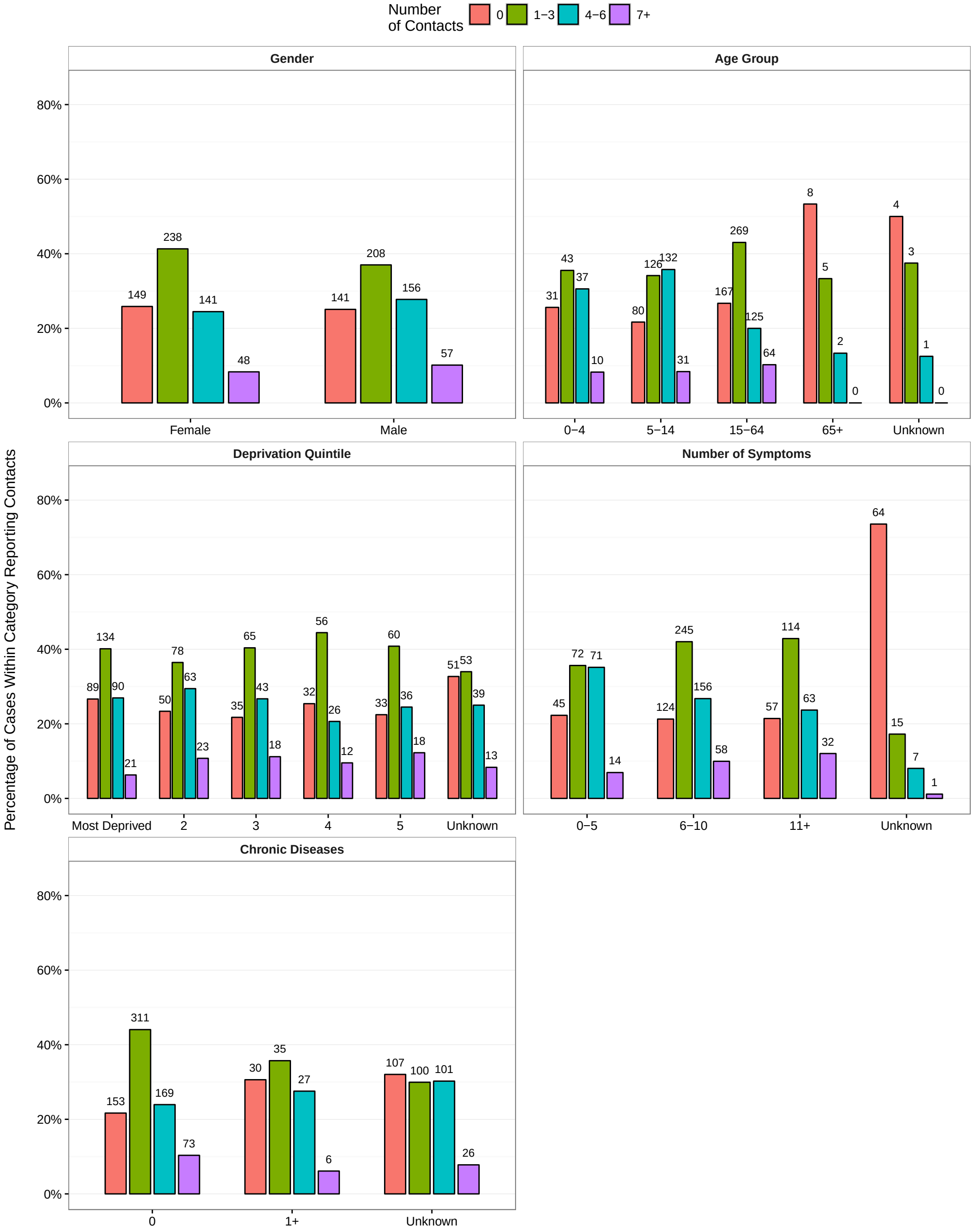


**Figure 3.13:** Barplots of numbers of contacts per case grouped by gender of case, age group, deprivation quintile, number of symptoms and chronic disease. Note that the percentages on the $y$-axis scale give percentages within groupings for each variable reporting a certain number of contacts (for example, percentages of female cases having 0, 1–3 contacts, etc), and the counts above bars show the numbers of cases having 0, 1–3 contacts, etc.

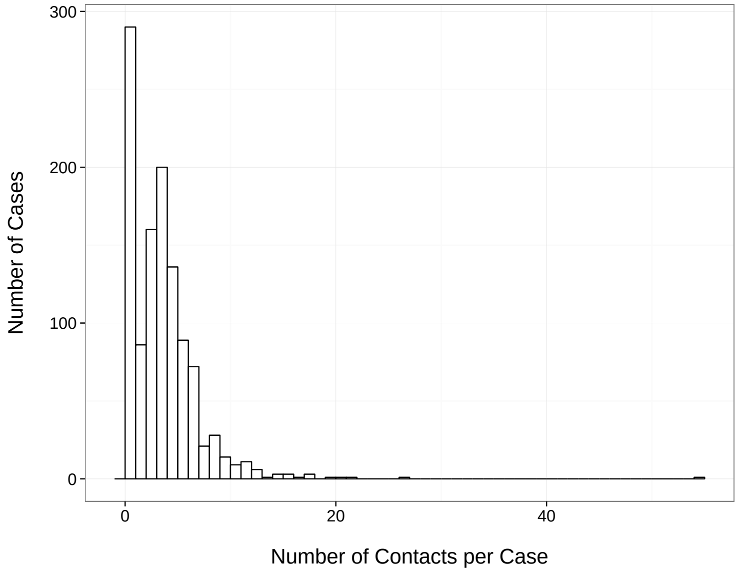


**Figure 3.14:** Distribution of numbers of contacts per case.

We further explored case-contact relationships by looking at ages of cases against the ages of their contacts. Similarly, we also looked at gender of cases against the gender of their contacts and also deprivation of cases against deprivation of their contacts. We found that cases belonging to any age group had most contacts aged 15–64. More specifically, cases aged 0–4, 5–14 and 15–64 had an average of almost two contacts aged 15–64, but cases aged 65+ had only an average of 0.67 contacts aged 15–64 and typically had the least amount of contacts (Table 3.13). For all other case-contact age combinations, averages of less than one contact per case were found.

Table 3.14 shows that cases had more female contacts on average regardless of whether the case was a male or female, but males generally had slightly more contacts. The numbers on the diagonal of Table 3.15 are largest and this shows that cases tended to have contacts that lived in areas belonging to the same deprivation quintile as them. Furthermore, the results also indicate that cases reported extremely few contacts from different deprivation quintiles to them. One reason for this finding is that many of the contacts reported by cases would have been household contacts. A supplementary table showing more complete results on case-contact relationships such as the numbers used to produce the mean numbers of contacts is given in Appendix 3.A.

| | **Contact Age** | | | | |
|---|---|---|---|---|---|
| **Case Age** | 0-4 | 5-14 | 15-64 | 65+ | Unknown |
| 0-4 | 0.36 | 0.64 | 1.86 | 0.10 | 0.52 |
| 5-14 | 0.27 | 0.64 | 1.90 | 0.07 | 0.45 |
| 15-64 | 0.19 | 0.34 | 1.74 | 0.07 | 0.60 |
| 65+ | 0.00 | 0.07 | 0.67 | 0.13 | 0.27 |
| Unknown | 0.00 | 0.12 | 0.62 | 0.00 | 0.50 |

**Table 3.13:** Mean number of contacts by ages of cases and contacts.

| | **Contact** | | |
|---|---|---|---|
| **Case** | Female | Male | Unknown |
| Female | 1.45 | 1.34 | 0.14 |
| Male | 1.60 | 1.55 | 0.09 |

**Table 3.14:** Mean number of contacts by gender of cases and contacts.

| | **Contact Deprivation** | | | | | |
|---|---|---|---|---|---|---|
| **Case Deprivation** | Most Deprived | 2 | 3 | 4 | 5 | Unknown |
| Most Deprived | 1.31 | 0.07 | 0.04 | 0.05 | 0.01 | 1.33 |
| 2 | 0.06 | 1.22 | 0.10 | 0.01 | 0.03 | 1.79 |
| 3 | 0.13 | 0.12 | 1.19 | 0.02 | 0.06 | 1.86 |
| 4 | 0.06 | 0.12 | 0.04 | 1.03 | 0.06 | 2.02 |
| 5 | 0.02 | 0.01 | 0.09 | 0.06 | 1.32 | 1.81 |
| Unknown | 0.24 | 0.28 | 0.17 | 0.07 | 0.06 | 1.96 |

**Table 3.15:** Mean number of contacts by deprivation quintile of cases and contacts.

## 3.5 Discussion

This analysis looked at all known reported cases of 2009 pandemic H1N1 influenza in Scotland between 25th April to 9th July 2009 that also provided detailed data from questionnaires in addition to the routinely collected dates of report or onset of symptoms. One of the main purposes of this analysis was to compare our findings on the characteristics of cases in Scotland against what has been reported elsewhere. The demographic characteristics showed that there were no major differences in numbers of reported cases by gender, but most reported cases were children and young adults. This is in line with previous reports in the UK, although the median age of 17 found here was higher than the median ages of 12 [155] and 15 [240] found previously in the UK. The higher median age found for reported cases in Scotland indicates that the outbreak was community based rather than school based. In a more global comparison, the median age of early cases in Scotland was higher than in Peru (median 13 years) [250] but lower than what was found in Western Australia and Victoria as well as initial case series from Europe and the United States (median 20–25 years) [200].

It is possible that some of the estimates with lower median ages could reflect differential testing patterns over time. This is particularly the case if school children were targeted for testing as this group is known to spread influenza at higher rates due to their mixing patterns compared with adults. In fact, one of the first recognised clusters of cases in the UK occurred in a school in the West Midlands [18]. We also found very few cases aged over 65. This may be because older individuals have been exposed to similar previous influenza pandemics giving them some level of extra protection against 2009 pH1N1 due to cross-reacting antibodies [28]. In contrast to this, there has been no real evidence to suggest that vaccination with recent seasonal nonadjuvanted or adjuvanted influenza vaccines conferred any additional protection against pandemic influenza [149]

The majority of patients reported fairly mild illness with the most common symptoms being similar to those experienced during seasonal influenza outbreaks. These symptoms include fever, coughing, general fatigue (malaise), sore throat and muscle or joint aches. The proportions of early cases in Scotland that reported these common symptoms were similar to what was found previously in the UK by the Health Protection Agency (HPA) in England (whose role was taken over by Public Health England since 2013) [155].

Most cases reported six or more symptoms, with a median of nine symptoms per case which is higher than the median number of symptoms reported in an Australian study [51]. However, one possibility for their lower median is that they only considered eleven different symptoms compared with the eighteen symptoms that we considered. A much more reliable comparison on numbers of symptoms could only be made if both studies considered the exact same symptoms. The Australian study compared pandemic influenza cases with seasonal influenza cases and found that the median reported number of symptoms was the same for both groups. Hence, this could suggest that pandemic influenza did not cause a substantially greater degree of discomfort compared with seasonal influenza.

Asymptomatic infections are possible but a systematic review found that there was not much evidence to suggest that asymptomatic cases cause much transmission of influenza [283]. However, the authors of that review do stress that much more research is needed on the transmission potential of asymptomatic cases. Only four of the early cases in Scotland reported no symptoms despite being laboratory-confirmed for pandemic influenza. The actual number of asymptomatic infections is not known but will almost certainly be much higher as asymptomatic cases will inevitably be missed by surveillance measures if individuals do not meet the case definitions described in Section 3.2.1.

One notable aspect related to symptoms is that there were higher than usual proportions of cases reporting diarrhoea and nausea/vomiting compared with seasonal influenza cases which is a feature also found in an analysis of the first few hundred (FF100) cases in the UK [240]. We note that the FF100 data also included eighteen Scottish cases which would also have been included in our study. As only around 2% of the cases in our study were also in the FF100 study, the reason for both studies producing these same findings cannot be attributed to these individuals being included in both studies.

The reasons for the higher proportions of cases reporting diarrhoea and nausea/vomiting are not clear but one study on the use of oseltamavir did find that some patients suffered from gastrointestinal upset and these symptoms were the most frequently reported in their review [377]. However, we found no significant difference in the likelihood of cases reporting diarrhoea or nausea/vomiting based on whether or not they received oseltamivir. The proportions of cases reporting diarrhoea varied around the world which illustrates a lack of consistency for this finding. In the USA and France, the proportions of cases reporting diarrhoea were considerably smaller than in our study [177, 190], but in Spain, the proportions reporting this symptom were higher [350]. In general, the proportions of cases in Scotland reporting symptoms such as nausea/vomiting and muscle/joint ache were high compared with these other countries.

When making comparisons between groups (age groups, gender and deprivation) on counts data such as numbers of symptoms and the delay before receiving antivirals, it should be noted that we could also have used Poisson regression. However, as we were focussing on univariable relationships for those analyses, we believed it was sufficient to use non-parametric statistical tests to draw conclusions. Here we found that females tended to report more symptoms than males and that those aged 15–64 reported more symptoms than infants aged 0–4.

We found that males did not differ much from females in terms of the types of symptoms reported, which is consistent with results from a surveillance report in Colombia [55]. Our lack of significant differences found between males and females could be due to the choice of using Bonferroni corrections for multiple testing which imposes a stricter cut-off level for statistical significance. However, corrections for multiple testing were used as it is more appropriate for multiple comparisons. This brings the benefit of producing less spurious results which could arise when looking at a large number of symptoms. An earlier surveillance report for the UK by Pebody et al. [286] found that females were more likely to report vomiting and were also more likely to report diarrhoea, but they only supplied odds ratios and did not state if they took into account multiple testing. We also found that females were more likely to report nausea/vomiting but did not find a significant difference in the likelihood of reporting diarrhoea between males and females.

Different symptom patterns were found for individuals belonging to different age groups. Infants up to four years old were more likely to report runny nose while being less likely to report sore throat, chills, shortness of breath, muscle/joint ache and headache compared with adults aged 15–64. Although statistically significant differences were found, we speculate that information relating to infants may have been completed by parents or guardians and hence, may offer less accuracy than self-reported symptoms. School-aged children (ages 5–14) were more likely to report fever but less likely to report muscle/joint ache and shortness of breath. These results are fairly similar to the results of an analysis of the FF100 cases where they looked at only two age groups (under 16 and 16+) [240]. A study from the USA by Jhung et al. [190] found that children were more likely to report vomiting compared with adults but we did not find sufficient evidence for this.

When looking for differences in the specific types of symptoms reported by gender, age group and deprivation, logistic regression was used to obtain ORs and these were interpreted as relative risks (RRs) would be. However, these results should be interpreted with caution as this would only have been wholly appropriate for symptoms that were reported rarely. This is because the OR only approximates the RR when the outcome is rare and for more common outcomes, the risk is overestimated when using the OR; as the incidence of an outcome increases, this overestimation becomes larger [77, 208]. In this situation, alternative methods can be considered such as methods which adjust the OR for common outcomes [242, 399].

We used PCA to investigate how groups of symptoms clustered among the cases. The procedure split the cases into three main clusters which broadly categorised cases according to their severity of illness – at least as defined by the average numbers of symptoms reported by cases within those groups. The symptoms that caused the most separation between groups were sneezing and runny nose which were reported by all cases in the most severe group and not reported by any cases in the least severe group. This finding loosely suggests that cases that do not report these two symptoms generally report less symptoms overall. However strong conclusions were not possible from using PCA as we only considered the first two PCs which only accounted for slightly under a third of the variance in the symptoms data.

Most cases received antivirals, with the vast majority getting oseltamivir (Tamiflu) and only a few receiving zanamivir (Relenza). No cases received Amantadine or Rimantadine which were not recommended for treating pandemic influenza [253] as the current pandemic strain showed some resistance to these treatments [188]. The FF100 analysis showed that treatment with antivirals reduced the length of symptoms but the antivirals must be administered in a timely fashion; it is recommended that antivirals are taken within 48 hours of symptoms onset. Less than 40% of cases received antivirals within this time frame and on average, it took around three days to receive antivirals. This signals a need to raise awareness about the importance of using antivirals promptly. Antiviral use decreased slightly with increasing age but there was no significant association between receiving antivirals and age.

The case-hospitalisation rate was 2.7% and 5.5% attended accident and emergency departments. Higher rates of hospitalisation were found in earlier studies in the UK which could be due to people being more cautious in an earlier phase of the pandemic rather than any real difference in severity of illness. Morever, an earlier study in the US by Jain et al. [185] found that those with underlying medical conditions were more likely to be hospitalised. We also found that those with at least one associated chronic disease had an increased odds of attending these departments but our result did not reach statistical significance; this is most likely due to having only 98 patients with at least one chronic disease. Hence, our results, along with the findings of other studies, still indicate that there is probably a need for targeted early intervention in those considered to be at higher risk of complications following influenza infection.

An analysis of the contacts of cases offered only limited value as much of the data was incomplete or difficult to extract meaningful information from due to data errors, inconsistencies and the fact that much of it was entered in free-text form. While reasonably sophisticated methods for filling in missing data with *plausible* values do exist such as multiple imputation [345], these methods were deemed to be not appropriate for the contacts data due to the large quantities of missing data for variables.

Not all cases reported contacts and for these cases, this likely points to a data issue regarding follow-up on their contacts rather than a true result of those cases having no contacts. A further issue with the contacts data was that some contacts related to cases that were not in the cases dataset that we analysed and so they could not be used in our analysis here. It is probable that many of the contacts later became cases but this could not be identified from the available information on contacts.

Out of the contacts that could be categorised into household or non-household contacts, around three quarters were household contacts and this implies that household contacts are much more likely to be reported by cases compared with non-household contacts. One possible cause for this may be recall bias where cases may be more likely to remember their household contacts compared with their non-household contacts. It is also important to note that some contacts were assumed to be household contacts based on terms describing family members such as ‘brother’, ‘wife’, ‘mother’ and some of these will likely be wrongly classified.

The numbers of contacts for cases differed by age group of the case but not by deprivation. Higher proportions of cases aged 15–64 had 0–3 contacts and on the whole, seemed to have less contacts than cases aged under 15. One possible explanation for this could be that most of the cases aged 15–64 have left school and perhaps, have moved away from their parents’ households, thus causing them to have less frequent contact with family and friends. Regardless of the age of the case, most of their contacts were aged 15–64. Since the 15–64 age range covers the largest age range, this finding is expected. For infants and children, their contacts aged 15–64 would most likely have been their parents while for cases aged 15–64, their contacts aged 15–64 are likely to be less obvious – their contacts could be partners, friends or work colleagues. The quality and completeness of the contacts data was not good enough to produce evidence for these speculations.

A higher proportion of cases as well as their contacts came from more deprived areas and cases predominantly reported contacts that were from areas with a similar deprivation status to them. One reason for this finding would have been because a large proportion of reported contacts were household contacts and a contact's postcode (and hence, deprivation) was known much more often for household contacts compared with non-household contacts. There was insufficient evidence to show that deprivation affected whether or not cases received antivirals, treatment delay or the severity of the illness in terms of the number of symptoms reported. The disproportionate amount of cases from more deprived areas may result from differences in mixing patterns or could merely reflect reporting or testing biases rather than actual differences in infection transmission. A large majority of reported cases were from the GGC healthboard which generally has more areas of higher deprivation compared with other parts of Scotland [138]. A speculation we could make here is that there may have been a detection bias towards cases from Glasgow and hence, a bias for finding cases from more deprived areas.

There are several more limitations inherent in this study on top of those already mentioned. As the data comes from questionnaires and interviews, we can never fully discount the possibility of memory recall errors from cases, especially when recalling for instance, dates of symptoms onset. Data recording errors are another issue which has to be considered. During the earlier phases of a pandemic, this is likely to be more of a problem as there is tremendous pressure on staff to deal with cases very quickly. Additionally, some of the data regarding symptoms was imputed as previously described in Section 3.2.2, but the quantities of data imputed were not excessive. This data was also imputed in such a way as to take into consideration how questionnaires are completed by cases. Thus, we believe the data imputations would not have introduced significant errors in the results but we did not test how different results would have been if the analysis was performed with only the non-imputed data. As the data uses only laboratory-confirmed cases, many cases may have been missed; particularly asymptomatic cases. However, initially all suspected cases were tested and this means that more cases would have been found than if testing had been restricted to only highly probable cases.

To conclude, the analysis undertaken here has provided an overview of the descriptive epidemiology around early cases of 2009 pandemic influenza in Scotland and adds to the knowledge base on the disease. Missing data was accounted for by making sensible assumptions around the data generation process and corrections were made for multiple testing to limit the amount of spurious associations found in analyses. By and large, cases reported fairly mild illness, similar in severity to seasonal influenza. In contrast to seasonal influenza, pandemic influenza was more prevalent in younger individuals and rarer for older individuals. The types of symptoms reported by cases did not differ much between gender or deprivation but more differences were observed by age group.

This analysis provides information which could be useful in efforts to model outbreaks of the disease in Scotland as well as assisting in policy decisions. However there are many more key transmission parameters related to infectious disease outbreaks that we have not covered here. One of the most important parameters is the *reproductive number* of a disease which describes the average number of individuals that each case infects. This is useful for predicting how quickly an epidemic will grow and in turn, this information can be used to guide decisions on if interventions are required to control an outbreak. The issue is that estimating this parameter can be troublesome unless perfect data on who infected whom is available and this is almost never available for any outbreak, let alone an outbreak on the scale of a pandemic. Depending on what is known about a disease, only certain methods may be suitable for estimating the reproductive number. In the next chapter, we review some of the different ways that this parameter can be estimated.

# 3.A Appendix: Chapter 3

| Case Group | Contact Group | Contacts | Cases | Mean Contacts per Case | Max Contacts per Case |
|---|---|---|---|---|---|
| *Age Group* | | | | | |
| 0–4 | 0–4 | 44 | 121 | 0.36 | 5 |
| | 5–14 | 77 | 121 | 0.64 | 15 |
| | 15–64 | 225 | 121 | 1.86 | 23 |
| | 65+ | 12 | 121 | 0.10 | 2 |
| | Unknown | 63 | 121 | 0.52 | 12 |
| 5–14 | 0–4 | 98 | 369 | 0.27 | 5 |
| | 5–14 | 237 | 369 | 0.64 | 20 |
| | 15–64 | 701 | 369 | 1.90 | 8 |
| | 65+ | 25 | 369 | 0.07 | 2 |
| | Unknown | 166 | 369 | 0.45 | 10 |
| 15–64 | 0–4 | 117 | 625 | 0.19 | 4 |
| | 5–14 | 214 | 625 | 0.34 | 5 |
| | 15–64 | 1089 | 625 | 1.74 | 16 |
| | 65+ | 44 | 625 | 0.07 | 2 |
| | Unknown | 375 | 625 | 0.60 | 13 |
| 65+ | 0–4 | 0 | 15 | 0.00 | 0 |
| | 5–14 | 1 | 15 | 0.07 | 1 |
| | 15–64 | 10 | 15 | 0.67 | 3 |
| | 65+ | 2 | 15 | 0.13 | 1 |
| | Unknown | 4 | 15 | 0.27 | 4 |
| Unknown | 0–4 | 0 | 8 | 0.00 | 0 |
| | 5–14 | 1 | 8 | 0.12 | 1 |
| | 15–64 | 5 | 8 | 0.62 | 3 |
| | 65+ | 0 | 8 | 0.00 | 0 |
| | Unknown | 4 | 8 | 0.50 | 2 |
| *Gender* | | | | | |
| Female | Female | 835 | 576 | 1.45 | 15 |
| | Male | 773 | 576 | 1.34 | 11 |
| | Unknown | 83 | 576 | 0.14 | 11 |

***Continued on next page***

| **Case Group** | **Contact Group** | **Contacts** | **Cases** | **Mean Contacts per Case** | **Max Contacts per Case** |
|---|---|---|---|---|---|
| Male | Female | 902 | 562 | 1.60 | 29 |
| | Male | 870 | 562 | 1.55 | 25 |
| | Unknown | 51 | 562 | 0.09 | 6 |
| *Deprivation Quintile* | | | | | |
| Most Deprived | Most Deprived | 438 | 334 | 1.31 | 11 |
| | 2 | 24 | 334 | 0.07 | 4 |
| | 3 | 12 | 334 | 0.04 | 3 |
| | 4 | 18 | 334 | 0.05 | 6 |
| | 5 | 5 | 334 | 0.01 | 2 |
| | Unknown | 446 | 334 | 1.34 | 16 |
| 2 | Most Deprived | 13 | 214 | 0.06 | 3 |
| | 2 | 262 | 214 | 1.22 | 7 |
| | 3 | 21 | 214 | 0.10 | 6 |
| | 4 | 3 | 214 | 0.01 | 1 |
| | 5 | 7 | 214 | 0.03 | 4 |
| | Unknown | 383 | 214 | 1.79 | 9 |
| 3 | Most Deprived | 21 | 161 | 0.13 | 3 |
| | 2 | 19 | 161 | 0.12 | 7 |
| | 3 | 191 | 161 | 1.19 | 6 |
| | 4 | 4 | 161 | 0.02 | 1 |
| | 5 | 9 | 161 | 0.06 | 2 |
| | Unknown | 300 | 161 | 1.86 | 18 |
| 4 | Most Deprived | 7 | 126 | 0.06 | 4 |
| | 2 | 15 | 126 | 0.12 | 6 |
| | 3 | 5 | 126 | 0.04 | 3 |
| | 4 | 130 | 126 | 1.03 | 8 |
| | 5 | 7 | 126 | 0.06 | 3 |
| | Unknown | 254 | 126 | 2.02 | 43 |
| 5 | Most Deprived | 3 | 147 | 0.02 | 1 |
| | 2 | 2 | 147 | 0.01 | 1 |
| | 3 | 13 | 147 | 0.09 | 7 |
| | 4 | 9 | 147 | 0.06 | 2 |
| | 5 | 194 | 147 | 1.32 | 8 |

***Continued on next page***

| **Case Group** | **Contact Group** | **Contacts** | **Cases** | **Mean Contacts per Case** | **Max Contacts per Case** |
|---|---|---|---|---|---|
| | Unknown | 266 | 147 | 1.81 | 21 |
| Unknown | Most Deprived | 37 | 156 | 0.24 | 5 |
| | 2 | 43 | 156 | 0.28 | 5 |
| | 3 | 26 | 156 | 0.17 | 6 |
| | 4 | 11 | 156 | 0.07 | 3 |
| | 5 | 10 | 156 | 0.06 | 3 |
| | Unknown | 306 | 156 | 1.96 | 17 |

**Table 3.16:** Numbers of contacts for cases by different age groups, gender and deprivation quintile. The mean and maximum numbers of contacts per case for these groupings are also given.

# Chapter 4

# Methods of Estimating Reproductive Numbers

In Chapter 3, we gave an overview of the characteristics of 2009 pandemic influenza cases in Scotland during the early phase of the epidemic. Whilst it is important to understand how the disease is affecting different groups of individuals, there are also many other crucial features of an epidemic which must be estimated using mathematical and statistical methods; this is what we turn our attention to in this chapter. Specifically, this chapter reviews some of the literature on the different methods of estimating a vitally important parameter in infectious disease outbreaks. This parameter is known as the the *reproductive number* and it is often used to summarise the rate of spread of an infectious disease.

Covering all of the methods of estimating the reproductive number would be a truly monumental task due to the huge quantity of methods out there. Hence, for this literature review, the intention is not to provide a comprehensive list of all the different methods of estimating the reproductive number but rather to outline a variety of methods which can be used depending on the data and information available regarding the epidemic. The aim here is to describe methods of estimating the reproductive number which are applicable to the types of data that are most commonly collected during different types of disease outbreaks. Since having some understanding around mathematical models for infectious diseases is beneficial for understanding some of the other methods of estimating the reproductive number, we give a basic introduction to those models here and also show how they can be used to derive the reproductive number as well.

For selecting methods, the thought process was to include methods which can be applied to different sizes of outbreak and different phases of outbreak – especially the early phase which can be a hugely important phase as failure to control an outbreak early can cause an outbreak to grow to sizes that are far more difficult to manage. Whilst some methods can only be applied to the early phase where numbers of cases usually grow exponentially, other methods can only be applied at the end of an outbreak since they require data which can only be observed at the end. Perhaps the most useful methods are those which can be applied continually throughout an outbreak which is vital for surveillance and planning interventions; we term these as *real-time* methods. Finally, an important aspect of this review is that some of the methods explained here will be applied to Scottish data on 2009 pandemic influenza in the next chapter, and thus this chapter acts as reference material for some of the analysis carried out in Chapter 5.

## 4.1 Introduction

When conducting epidemiological studies of disease outbreaks, it is most often the case that data is at a premium and hence, many important epidemiological parameters need to be estimated from only the limited amount of data that is available from for instance, disease surveillance systems. Outbreak data are usually based upon counts of cases and the most obvious way of detecting cases comes through individuals exhibiting symptoms of the disease and subsequently reporting this to health services. These cases may also be stratified by some measures of interest such as age, sex, and socioeconomic status.

Another major problem with epidemiological data is that it is usually imperfect; for example, when looking at an infectious disease such as influenza, the number of people who actually report their illness is only a small minority compared with all of the people who actually contract influenza. One reason for this is that not all infected individuals will actually exhibit clinical symptoms and thus, may not even realise that they have been infected; these are known as *asymptomatic cases*.

Commonly the notification rates of disease are better during periods where the incidence of infection is high. In particular, the media plays a vital role towards how an epidemic unfolds and can heavily influence notification rates [95]. Furthermore, notification rates tend to vary with factors such as age where we might expect notification to be highest among young infants and the elderly and lowest among teenagers and young adults.

With all of these issues in mind, researchers have developed a plethora of different methods of estimating parameters and these methods are usually specifically tailored to what data is available for the disease. However, this is an on-going area of research and huge efforts still need to be made in order to improve the accuracy and efficiency of our estimations. This can be achieved both through the continued development of current methods and also via the creation of completely new methods. The ultimate goal is to achieve accurate and timely estimation which is absolutely vital as these estimated parameters can often be the basis for making important decisions on how to control infectious diseases [124].

### 4.1.1 Important Epidemiological Parameters

Several parameters are useful when analysing infectious diseases. These include the *incubation period* which is the delay between infection and the onset of symptoms [43]; the *latent period* which is the delay between infection and the beginning of being able to transmit the infection to other susceptible individuals; and the *infectious period* which describes how long an infected individual may transmit the virus. It should also be noted that a *recovery rate* can often be derived from the infectious period as well. The statistical estimation of the latent, incubation, and infectious periods are no simple task with a major problem being the collection of adequate samples of data. For example, the detailed data on pandemic influenza cases used in Chapter 3 was insufficient to make reasonable estimates on any of these parameters.

It could have been simple to derive a maximum possible incubation period for cases using the earliest possible exposure date to the disease which we could take as the date of onset of symptoms for the first case in Scotland. However, that method would obviously not produce useful maximums for cases that were diagnosed at times much later than the initial case and in general, it is difficult to estimate the earliest possible exposure time for diseases that are transmitted from person-to-person unless we have a reasonably good idea of when cases were in contact with each other. We may have also been able to derive a range for the incubation period using cases where the date of infection reported or date of confirmation was before the self-reported date of onset of symptoms but this only happened for less than ten cases. A further issue with the data was that the accuracy in the dates was questionable as some of the cases had a date of report over a month before their date of onset of symptoms for influenza. For these reasons, it would also have been difficult to estimate the latent period for cases. In fact, if we assume that cases are only able to transmit infection when they have symptoms then the latent period would become identical to the incubation period. As there was no data on which cases transmitted the disease to other cases and no data on viral loads [38, 216] for cases over time, this makes it very difficult to estimate the infectious period or the recovery rate as well.

Estimates of these parameters may be based upon detailed statistical analyses of transmission, volunteer infection studies or be more speculative in nature [9]. The period of time that individuals are in each of these disease states is variable. For example, the incubation period has previously been assumed to approximately follow a lognormal distribution [265, 324]. Furthermore, in mathematical models, it is often essential to be able to estimate the person-to-person transmission probability (commonly denoted as $\beta$ in mathematical models) and this factor is used to describe an individual's chances of becoming infected after contact with an infectious individual. Ideally, $\beta$ should take into account various factors such as epidemiological, environmental and social factors which may all affect the transmission rate. All of these parameters will of course, vary between individuals, and this is why we would estimate average values for these parameters.

### Serial Intervals

A key parameter which needs to be estimated during epidemics is the *serial interval distribution* (or *generation interval distribution*) of a disease, which specifies the time-scale at which infections accumulate. Many authors use these two names interchangeably. However, some authors will make a subtle distinction between the generation interval and the serial interval – the generation interval can be defined as the time interval between the infection of a primary case and one of its secondary cases, and the serial interval can be defined as the times between occurrences of similar observable events for a primary case and one of its secondary cases; most commonly the onset of clinical symptoms [352]. Thus, these two terms are only equivalent under certain conditions and can differ particularly when there are many asymptomatic infections and asymptomatic secondary infections.

These intervals have also previously been referred to as *transmission intervals* and have also been defined in a less verbal but more mathematical form as the sum of the average latent and average infectious period [9]. Fine [118] argued that this mathematical definition would only be identical to the other definitions if the contact frequency and infectiousness of a case is independent of the time since infection and for most infectious diseases, this would not be true. In the recent literature, this mathematical definition and the term, transmission interval, have been less common and hence, we will, by and large, use the term serial intervals as just defined.

For diseases such as influenza and severe acute respiratory syndrome (SARS), the serial interval is often observed by either contact tracing of cases to discover who acquired infection from whom or by observing time intervals between the symptoms onset of a first case and subsequent cases in a household. The average serial interval is variable for different diseases; for instance, it has been found to be 8.4 days for SARS in Singapore [372] and 3.6 days for influenza [72]. A number of methods utilise the serial interval distribution in estimating reproductive numbers as we shall go on to exemplify. The central role that the serial interval distribution plays in deriving reproductive numbers for some methods is a further reason for why it can be highly important to estimate the parameters that characterise the shape of the distribution.

### Reproductive Numbers

Perhaps the most important epidemiological parameter which has to be estimated is the average number of people that a primary case infects for a particular disease. This is also known as the *basic reproductive number* (note that sometimes "reproduction" is used instead of "reproductive") or $R_0$ when the population is completely susceptible. This can happen, for instance, before the introduction of a novel infectious disease that nobody has immunity from. Evidence of its importance can be seen by the fact that it is included in almost all academic papers that use some mathematical modelling in studying the spread of infectious diseases [163].

The value of $R_0$ is governed by the interaction between the infectious organism, its host and the environment. Factors which affect this include the route of infection, the duration of the infectious period, the type and frequency of contacts between individuals in a population and also the probability that a contact between an infected individual and a susceptible person will actually result in infection. Another reason why the estimation of $R_0$ is seen as vitally important is that it can be used as a tool to help guide government policy decisions such as whether or not to vaccinate individuals at birth for certain infectious diseases or if schools should be closed temporarily to limit transmission [110].

For the case where part of the population is immune from an infectious disease, either through past exposure to the disease or vaccination, we estimate the *effective reproductive number* (instead of $R_0$) which is abbreviated as $R$. This can also be tracked over a time period which is extremely important for surveillance purposes as it allows us to observe how an epidemic is progressing. This is in contrast to $R_0$ which cannot really be tracked over time for all diseases as for some diseases, individuals gain partial or full immunity following recovery from infection. The effective reproductive number at a particular time point can be denoted as $R_t$

where $t$ is the value of $R$ at time $t$. The concepts of $R$ and $R_0$ and the distinction between them is illustrated in Figure 4.1, which shows an example of $R_0 = 2$ and $R = 2$ (each infective individual infects on average, two others).

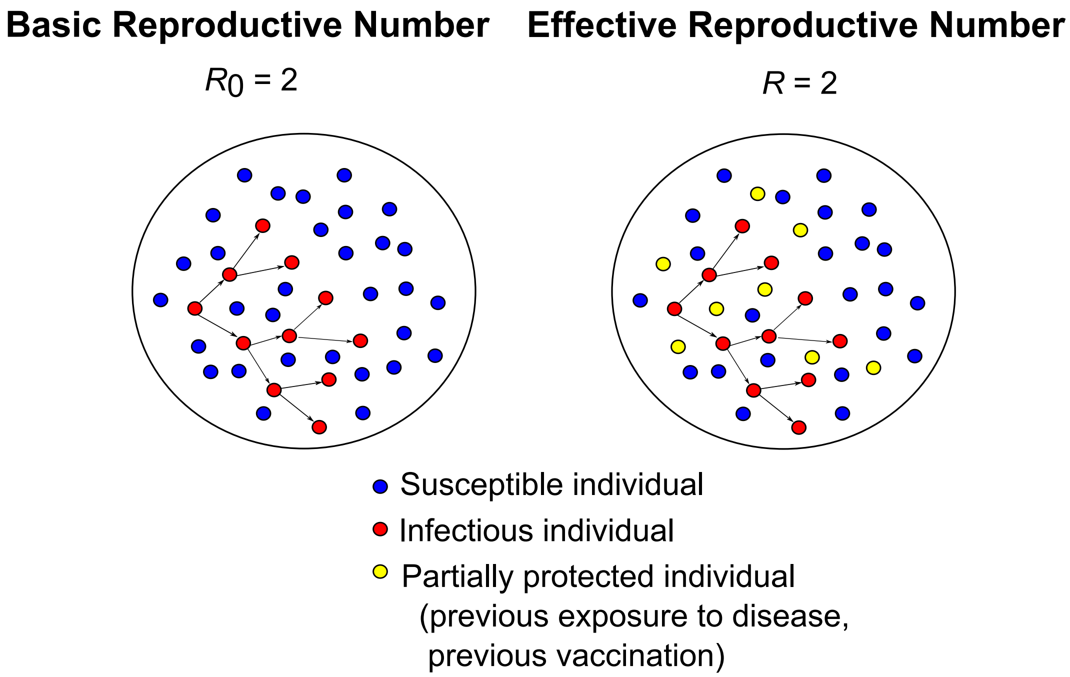

Basic Reproductive Number
$R_0$ = 2
Effective Reproductive Number
$R$ = 2
Susceptible individual
Infectious individual
Partially protected individual
(previous exposure to disease,
previous vaccination)

**Figure 4.1:** Illustration to show the distinction between basic and effective reproductive numbers ($R_0$ and $R$). In the example, each case infects on average, two other susceptible individuals. For the effective reproductive number, yellow circles represent partially protected individuals who could be immune to the disease or at least, less likely to be infected compared with fully susceptible individuals (blue circles).

One of the strengths of $R$ is its intuitive appeal with regards to the fact that it has a threshold value of 1. When $R > 1$, we can think of this as each person transmitting the disease to more than one person and so the the disease will spread into the population. When $R < 1$, the disease should eventually die out as long as this persists long-term, and finally in the case that $R = 1$, there will theoretically be no growth or decline in the number of cases. Thus, infection would be maintained in the population (*endemic*). Both $R$ and $R_0$ are especially useful as they allow us to assess the risk of an epidemic and also gives an idea about the scale of the effort that is necessary to bring the infection under control.

## 4.2 Mathematical Models

We start our review on methods of estimating $R$ by discussing mathematical models of infectious diseases [168, 197, 245] as the general concepts involved for some of the basic models fit into many other methods of estimating $R$. Hence it is necessary to have some understanding around them before proceeding to discuss some of those other methods. Moreover, it is only necessary to discuss the basic models here but discussions on far more complex models can be found in, for example, Keeling and Rohani [198].

Mathematical models are a useful way of providing some predictions and insight into an infectious disease. Wallinga and Lipsitch [371] even stated that mathematical models of transmission have become invaluable in planning for the control of emerging infectious diseases. The idea behind these models is that they assign an *infection status* to all members of a population. A set of equations is formed with the intention of capturing the change in the numbers of individuals (or commonly proportions of a population) in each of the infection states over time. Creating a model usually requires us to make many assumptions on the disease process, and the model is usually formulated depending on the precision or generality required as well as the data that is available. This is because it is often too strenuous to account for all factors of the infectious disease. Take for example, an airborne infection such as influenza. An accurate model would need to take into account the variations in social interaction, variations in temperature and climate and also variability due to genetic factors.

### 4.2.1 SIR Model

It is best to introduce mathematical models of infectious diseases through one of the simplest models – the Susceptible-Infectious-Recovered (SIR) model [204] excluding any complications such as demographic variations. This means that we do not concern ourselves with births, deaths or migration and assume that their effects on the disease process in a particular scenario are all negligible. This would only be plausible in a situation where a low level of disease is introduced to a large naive population, and the epidemic ends sufficiently quickly such that demographic factors do not have sufficient time to exert any significant influence on the process. Furthermore, the model also assumes homogeneous social mixing whereby individuals of all ages, genders and ethnic backgrounds will contact each other at exactly the same rate, which may be reasonable in the case of households or schools, but would perhaps not be reasonable when considering a country-wide population.

The SIR model classifies individuals of a population into three distinct classes; individuals are initially susceptible ($S$), then they become infected as well as infective ($I$) and finally they make a full recovery and gain lifelong immunity from the infectious disease or die ($R$). For an epidemic to start, we initially require at least one infected individual. The SIR process is represented schematically in Figure 4.2.

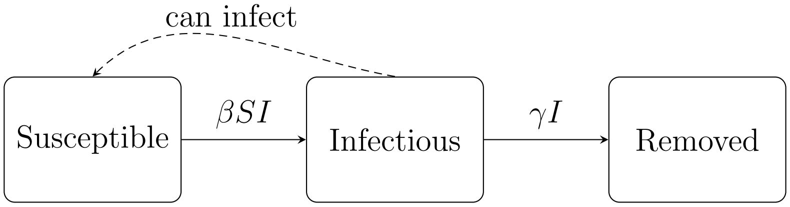


**Figure 4.2:** Flow diagram of the SIR model. Individuals progress from susceptible to infectious at a rate $\beta$ and then to removed/recovered at a rate $\gamma$.

The net rate of spread of infection is assumed to be proportional to the product of the density of susceptible people and the density of infectious individuals; this concept is known as the *mass action principle* [9]. This is an important assumption for the SIR model and with this, we can say that at each time step, only a proportion $\beta$ of the contacts between susceptibles and infectives actually result in infection. During each time step, the infectious class grows by the number of newly infected individuals. However, at the same time, some infectives will recover or die and will cease to be infective. These people progress to the removed/recovered stage at a rate $\gamma$. This consequently means that at each time step the recovered class increases by the same amount as the infected class decreases. Note that all of the rates, $\beta$ and $\gamma$, are assumed to be constant over time. A mathematical description of the SIR model can be constructed using a series of differential equations.

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta SI \\
\frac{dI}{dt} &= \beta SI - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{aligned}
\tag{4.1}
$$

where $\beta$ and $\gamma$ are as described before; $S$, $I$ and $R$ (not to be confused with the effective reproductive number used frequently throughout this chapter) represent the proportions of susceptible, infectious and removed individuals in a population respectively. Here $\gamma$ is the transition rate from the infectious stage to the removed stage and as the rates in this model are assumed to be constant, the average duration that an individual is infectious is $\frac{1}{\gamma}$, based upon an exponential distribution for the duration of infection. It must be noted that because we are working with proportions, $S + I + R = 1$ and consequently, $R$ can be solved as $1 - S - I$. The required initial conditions for Equation (4.1) are $S(0) > 0$, $I(0) > 0$ and $R(0) = 0$.

### Parameter Estimates

For a novel infectious disease where we assume that initially nobody has immunity to the disease, we can usually find $S$ from demographic data such as mid-year population estimates [180, 257], and $I$ will initially be a small number, perhaps one as new infectious diseases may commence as a result of one individual spreading it. This leaves us with the parameters $\beta$ and $\gamma$ to solve.

Despite the simplicity of this model, we cannot easily get an exact analytical expression for the dynamics of $S$ and $I$ though time due to the nonlinear transmission term $\beta SI$. Thus solutions for these models are commonly found using numerical methods such as *Runge-Kutta* methods [48] for finding numerical approximations to ordinary differential equations and $\beta$, $\gamma$ can be found by fitting the SIR model to the data. One method of finding optimal values for $\beta$ and $\gamma$ is by fitting a least squares model; that is by finding the values for the parameters which will minimise the square of the errors between the observed data and the model predictions. Once $\beta$ and $\gamma$ are found, we can compute our parameter of interest, $R_0$. For an epidemic to happen, the number of infected individuals has to be increasing so $\frac{dI}{dt} > 0$ and thus, we also have

$$\beta SI - \gamma I > 0$$

$$\frac{\beta SI}{\gamma} > I$$

$$\frac{\beta S}{\gamma} > 1$$

At the start of an epidemic, we may assume that everybody is susceptible (or at least almost everybody). Hence $S \approx 1$, and setting $S = 1$ leaves $\frac{\beta}{\gamma} = R_0 > 1$. Another intuitive way to look at $R_0$ is to consider it as the transmission rate multiplied by the average infectious period which gives a measure of the maximum reproductive potential for infected individuals.

$$R_0 = \beta \times \frac{1}{\gamma} = \frac{\beta}{\gamma} \tag{4.2}$$

### Critical Proportion

As a value of $R_0 > 1$ means the infection will spread into the population, $R_0 = 1$ can be thought of as a critical value. Thus, for an infectious disease, it is also possible to calculate a *critical proportion* which is the proportion of susceptibles which must be protected from the disease (for instance, by vaccination) in order to bring $R_0$ down to a level where it will eventually die out. For an epidemic defined with $R_0 > 1$, the critical proportion can be derived.

$$p_c = 1 - \frac{1}{R_0} \tag{4.3}$$

In Equation (4.3), the term $\frac{1}{R_0} = \frac{\gamma}{\beta}$ is the *relative removal rate*. Taking the scenario of using vaccination as the means of protection, if the relative removal rate is small, then a large proportion of susceptibles need to be vaccinated. Conversely, if the relative removal rate is large then only a small proportion of susceptibles would require vaccination. This is the intuition behind the derivation of the formula for $p_c$. Since vaccines are not 100% effective (this topic is discussed in Chapters 6 and 7), we can adapt the equation to include an efficacy rate $\epsilon$ [147, 273].

$$p_c = \frac{1}{\epsilon}\left(1 - \frac{1}{R_0}\right) \tag{4.4}$$

### 4.2.2 Including Demography and a Latent Phase

Not all infectious diseases will follow a classical SIR pattern and so it must be adapted to suit the characteristics of a particular disease. This may involve adding or removing classes of hosts within the compartmental framework of the model. For example, we can add a class for individuals who are infected but not yet infectious (*latent*); this gives us what is known as a a Susceptible-Exposed-Infectious-Recovered (SEIR) model. Many common infectious diseases including influenza are described reasonably well by this type of model and hence, we will concentrate on explaining this model (see Keeling and Rohani [198] for explanations of many other variations of mathematical models for epidemics).

In the SEIR model, individuals are initially susceptible, then they become infected with an initial inoculation with a very small amount of pathogen units. The pathogen will subsequently reproduce rapidly within the host but during this phase the pathogen amounts are too low for active transmission to other susceptible individuals. Therefore individuals are infected but not yet infectious and this is what is known as the exposed stage. After this latent period is over, individuals in the exposed class then become infectious and are thus, classified as infectious in the mathematical model, where they can transmit the disease to other susceptibles for a period of time. Finally, in this model, individuals will make a full recovery and gain lifelong immunity or die, and can thus, be classified as removed/recovered. Figure 4.3 shows the SEIR model in a schematic fashion.

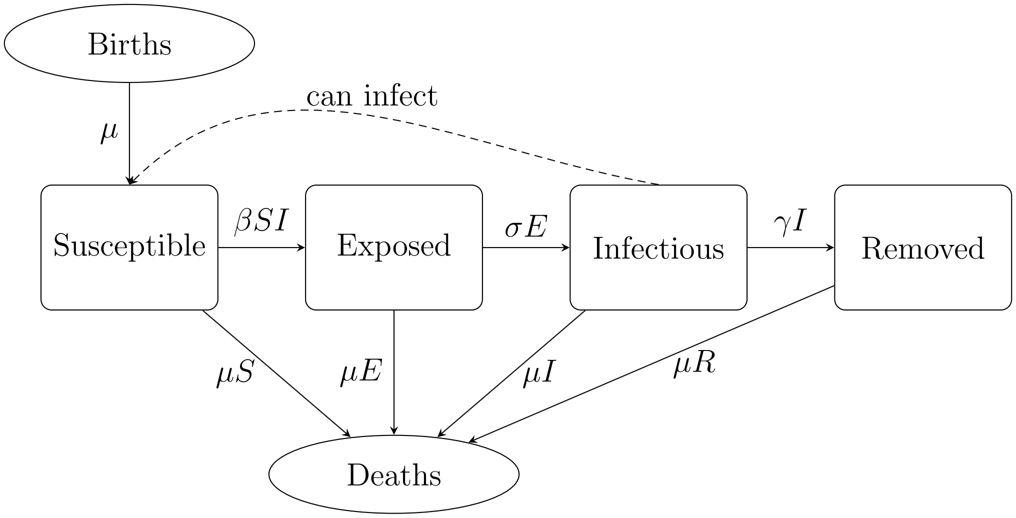


**Figure 4.3:** Flow diagram of the SEIR model with simple demography. The model is similar to the SIR model (as shown in Figure 4.2) except that individuals can also be classed as *exposed* and they leave this stage at a rate $\sigma$. In this model, newborns enter the susceptible population at a rate $\mu$ and deaths from natural causes occur at an equivalent rate $\mu$.

For the SEIR model, we also introduce some demography by adding a natural host lifespan into the differential equations where everybody is assumed to live exactly to a certain age and then promptly dies. It is important to note that this lifespan is usually the expected lifespan of a human and is independent of the disease. Therefore, in the susceptible class, a proportion $\beta$ of the contacts between susceptibles and infectives result in infection where they will endure a latent phase, but now some susceptibles are also lost due to natural mortality and some are gained due to natural births.

The exposed class will increase by the number of susceptibles who became infected, then some individuals will move to the infectious class at a rate $\sigma$ and some will also be lost in this class due to natural mortality. Similarly the infectious class receives individuals from the exposed class and then loses some due to natural mortality and also through individuals transitioning to the removed class a rate $\gamma$. Finally the removed class gains the individuals from the infectious class who have fully recovered but will also lose some individuals due to natural mortality. The differential equations for the SEIR model with births and mortality included are

$$\begin{aligned}\frac{dS}{dt} &= \mu - \beta SI - \mu S &&= \mu - (\beta I + \mu)S \\ \frac{dE}{dt} &= \beta SI - \mu E - \sigma E &&= \beta SI - (\mu + \sigma)E \\ \frac{dI}{dt} &= \sigma E - \mu I - \gamma I &&= \sigma E - (\mu + \gamma)I \\ \frac{dR}{dt} &= \gamma I - \mu R\end{aligned} \tag{4.5}$$

where $\mu$ is a natural mortality rate (from which the natural lifespan is $\frac{1}{\mu}$ years). In order to ensure that the model remains fairly simple, $\mu$ also represents the crude birth rate in the differential equation $\frac{dS}{dt}$ so that the total population size remains constant through time. $\beta$ and $\gamma$ remain as they were for the SIR model, and $\sigma$ is the rate at which individuals move from the exposed class to the infectious class. Again, some of these parameters including the natural mortality rate can be found from the demographic and case data, leaving $\beta$, $\sigma$ and $\gamma$ as the parameters that need to be estimated by fitting the model to data. Due to the addition of the exposed class, we now have $S + E + I + R = 1$, and $R_0$ is now

$$R_0 = \frac{\sigma}{(\mu + \sigma)} \times \frac{\beta}{(\mu + \gamma)} = \frac{\beta\sigma}{(\mu + \gamma)(\mu + \sigma)} \tag{4.6}$$

where $\frac{1}{(\mu+\sigma)}$ is the average amount of time spent by an individual in the exposed class and $\frac{1}{(\mu+\gamma)}$ is the average amount of time spent in the infectious class. To further interpret Equation (4.6), the ratio $\frac{\sigma}{(\mu+\sigma)}$ is the fraction of people that progress from the exposed stage to the infected stage; the fraction of people that become infected spend an average of $\frac{1}{(\mu+\gamma)}$ units of time infecting susceptible individuals, and again, we can recall that at the beginning of an epidemic $S \approx 1$.

We can also relate Equation (4.6) to the equation for $R_0$ in the simple SIR model. If we ignore births and natural mortality by setting $\mu = 0$, then Equation (4.6) becomes $\frac{\beta}{\gamma}$ and this is exactly the same as Equation (4.2), which also ignores births and natural mortality. Hence, introducing the exposed stage has no impact on the overall transmission potential of infected individuals unless other external factors such as births and deaths are introduced into the model. In essence, adding the exposed stage into the SIR model has the effect of creating a time delay before individuals become infectious. In a disease outbreak, the time delay is an important factor to account for in the model as it gives additional time for diagnosing as well as treating cases which could have a major impact on the numbers of individuals that become infected.

### Example: SEIR Model for the 1918 Influenza Epidemic

We mentioned earlier that the SEIR model fits reasonably well to the characteristics of influenza. One example of where this model has been applied to this disease is in a study by Massad et al. [236] who created an SEIR model for the 1918 Influenza epidemic in São Paulo, Brazil. They used this model to estimate the value of $R_0$ and also found a theoretical proportion that required vaccination to bring the value of $R_0$ below unity.

Even within the structure of a typical infectious disease model such as SEIR, the models are often customised depending on what is known about the disease – particularly with regards to the transmission dynamics of the disease and the parameters in the model. Therefore, the SEIR model by Massad et al. [236] was altered from the way that we previously outlined in Equation (4.5) by adding in a number of extra factors and choosing to work with actual numbers of individuals rather than proportions as well. In their model, individuals in the exposed and infected stages are both treated as being able to infect others. Furthermore, they add in a disease-induced mortality rate; a birth rate which is not equal to the natural mortality rate; asymptomatic cases; an incubation period; and also a carrying capacity of the population (a population size that the environment can sustain indefinitely). In keeping with earlier notation as far as possible, the system of differential equations to describe their SEIR model is

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta\left(\frac{E+I}{N}\right)S - \mu S + rN\left(1-\frac{N}{K}\right)\\
\frac{dE}{dt} &= \beta\left(\frac{E+I}{N}\right)S - (\mu+\sigma+\kappa)E\\
\frac{dI}{dt} &= \sigma E - (\mu+\alpha+\gamma)I\\
\frac{dR}{dt} &= \kappa E + \gamma I - \mu R
\end{aligned}
\tag{4.7}
$$

where $N = S + E + I + R$ is the total population size; $\beta$, $\sigma$ and $\gamma$ are the same as described previously; $\mu$ is the natural mortality rate; $r$ is the birth rate; $\kappa$ is the rate at which asymptomatic individuals move from the exposed class directly to the removed/recovered class without entering the infectious stage; $\alpha$ is the disease-induced mortality rate which only occurs during the infected stage; and $K$ is the carrying capacity of the population. With these extra additions, the expression for $R_0$ is

$$R_0 = \frac{\beta(\mu+\alpha+\gamma+\sigma)}{[(\mu+\sigma+\kappa)(\mu+\alpha+\gamma)]} \tag{4.8}$$

A heuristic for this derivation is to consider that infected individuals spend an average of $\frac{1}{(\mu+\sigma+\kappa)}$ time units in the exposed stage while infected individuals spend an average of $\frac{1}{(\mu+\alpha+\sigma)}$ time units in the infective stage. In addition to this, we can think of the possibilities for transmitting infection in two parts – all infected individuals will enter the exposed stage (keeping in mind that in this model, those in the exposed stage also transmit infection at an equivalent rate to those in the infectious class $\beta$) and then some of these individuals will transition directly to the removed stage, while the remainder – the fraction $\frac{\sigma}{(\mu+\sigma+\kappa)}$ – will enter the infective stage before making the transition to the removed stage. Therefore, we could write $R_0$ as

$$R_0 = \beta\left[\frac{1}{(\mu+\sigma+\kappa)}\right] + \beta\left[\frac{\sigma}{(\mu+\sigma+\kappa)} \times \frac{1}{(\mu+\alpha+\gamma)}\right]$$

which simplifies down to Equation (4.8). Massad et al. [236] used a combination of local demography data as well as data on weekly numbers of new cases and previous research on influenza by Longini Jr et al. [226] in order to find the required parameters. In fact the only parameter that they estimated by fitting to data was $\beta$. They showed that their model predicted the actual number of new weekly cases and deaths with a high level of accuracy (predicted deaths was almost identical to the official recorded number while the total number of cases was within a few percent of the official recorded number) and derived $R_0 = 2.68$. Therefore, the critical proportion was $p_c = \frac{1}{R_0} = 0.63$, meaning that 63% of the population required vaccination in order to bring $R_0$ below unity.

### 4.2.3 Transmission Heterogeneities

For many infectious diseases, individuals with differing characteristics may be more or less likely to transmit a disease to other susceptibles. As a consequence, this makes certain factors epidemiologically relevant and we must include this in a model to gain a more accurate estimation of our parameter of interest $R_0$. Traits such as age, sex and genetic composition can influence an individual's susceptibility to an infectious disease and in addition to this, these factors can also be related to an individual's pattern of social contact with other individuals [163]. An example of where it may be important to include heterogeneities is with an infectious disease that is transmitted respiratorily. Here we would assume that the disease has a higher transmission rate in enclosed areas such as schools and for this reason, we could give school children a higher transmission parameter.

To introduce heterogeneities in disease transmission, we will look at a simple model where individuals do not actually gain lifelong immunity, but after infection they become susceptible once again; this is known as a Susceptible-Infectious-Susceptible (SIS) model. This gives the simplification of being able to eliminate the differential equation for removed/recovered individuals, and as long as we have the differential equations for the infected class, we can calculate the remaining susceptibles by using $S = 1 - I$ (supposing $S$ and $I$ are proportions of the population).

In our model, we only have two distinct groups as shown in Figure 4.4 – individuals who have a higher rate of transmission (H) and individuals with a lower rate of transmission (L). Furthermore, we will also disregard demographic information.

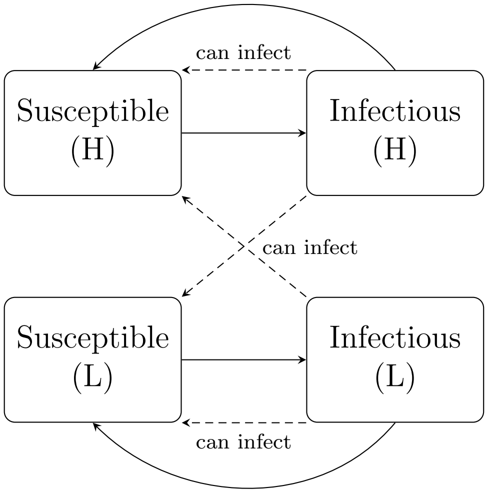


**Figure 4.4:** Flow diagram of the SIS model with heterogeneities in transmission rates. In this model, there are high risk and low risk groups for susceptible and infectious individuals who transmit infection and get infected at different rates.

As mentioned previously, we only need the differential equations for the infectious class for the SIS model. These will essentially be the same as the infectious differential equation for our first SIR model in Equation (4.1), but with the introduction of two separate risk groups we now have two differential equations for this class and for each differential equation we have two transmission parameters.

$$\begin{aligned}\frac{dI_H}{dt} &= \beta_{HH}S_HI_H + \beta_{HL}S_HI_L - \gamma I_H \\ \frac{dI_L}{dt} &= \beta_{LH}S_LI_H + \beta_{LL}S_LI_L - \gamma I_L\end{aligned} \tag{4.9}$$

In Equation (4.9), $\beta_{HH}$ is the rate at which high risk infectives transmit to high risk susceptibles; $\beta_{HL}$ is the rate at which high risk infectives transmit to low risk susceptibles; $\beta_{LH}$ denotes the transmission coefficient from low risk infectives to high risk susceptibles; and finally, $\beta_{LL}$ gives the transmission rate of low risk infectives to low risk susceptibles. This means that we now need to estimate a total of four transmission parameters along with the recovery rate parameter. The most convenient way of including heterogeneities is to use a *Who Acquires Infection from Whom* (WAIFW) matrix which contains all the different transmission parameters.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{HH} & \beta_{HL} \\ \beta_{LH} & \beta_{LL} \end{pmatrix} \tag{4.10}$$

A commonly assumed type of social mixing pattern is *assortative mixing*. This describes a situation where individuals with the same or similar characteristics are more likely to mix with each other. For instance, individuals in distinct age groups may have most social contact with other individuals in their own age group. In the context of this particular model, individuals in the high risk group would be more likely to mix with high risk susceptibles while those in the low risk group would be more likely to mix with low risk susceptibles. We would set the transmission parameters in the following way for assortative mixing: the transmission coefficient $\beta_{HH}$ would be highest and we would expect $\beta_{LL}$ to be the second highest term. Finally we would assume that the terms $\beta_{HL}$ and $\beta_{LH}$ would be the lowest and we could even further assume that $\beta_{HL} = \beta_{LH}$. This would mean that high risk individuals transmit infection to low risk individuals at the same rate as low risk individuals transmit to high risk individuals. We can also use social contact surveys to obtain an idea of how often individuals of various age groups contact other individuals of different age groups. The data from the surveys allow us to create contact matrices and in turn, this could be used as the WAIFW matrix [373].

Generally, $R_0$ is higher when including heterogeneities in transmission than when assuming homogeneous mixing [1, 26]. To calculate $R_0$, we use an eigenvalue approach in which we calculate the eigenvalues of the matrix

$$\boldsymbol{R} = \begin{pmatrix} \frac{\beta_{HH} n_H}{\gamma} & \frac{\beta_{HL} n_H}{\gamma} \\ \frac{\beta_{LH} n_L}{\gamma} & \frac{\beta_{LL} n_L}{\gamma} \end{pmatrix} \tag{4.11}$$

The matrix $\boldsymbol{R}$ is the matrix of coefficients (known as the *Next Generation Matrix*) where we have a common recovery rate $\gamma$ for all infected individuals and we also use an approximation $S_H = n_H$ and $S_L = n_L$ which is plausible during the initial phase of an epidemic where the relative change in the number of susceptibles is very small. Diekmann et al. [88] states that the expected number of secondary cases produced by a typical infected individual during their entire period of infectiousness in a completely susceptible population is mathematically defined as the *dominant eigenvalue* of a positive linear operator which is termed as *the next-generation operator.*

This operator is linear as we only take into account the initial stage of an epidemic where we disregard the depletion of susceptibles. The linear operator maps generations of infectives into each other, distributed over all the different characteristics of the individuals, and the dominant eigenvalue of this next-generation operator gives us information on whether or not the infected population will grow or decline in size. Eigenvalues fit the scenario for infectious diseases perfectly as they have a threshold value of one and can be interpreted as the average contribution to the next generation [163]. Here $R_0$ in this population can also be thought of as the *spectral radius* of our next generation matrix $\boldsymbol{R}$ and is found by taking all the absolute values of the eigenvalues of $\boldsymbol{R}$ and then choosing the largest value – the dominant eigenvalue.

Although our example only contained two risk groups, we can split the individuals of a population into more than two groups (for example, high, medium and low risk groups which could be made up of individuals belonging to three different age groups) and in the case of having $j$ groups, we would get the following differential equation for the infectious classes

$$\frac{dI_i}{dt} = \sum_j \boldsymbol{\beta} S_i I_j - \gamma_i I_i \tag{4.12}$$

where $\boldsymbol{\beta}$ is our matrix used to parameterise the transmission between groups of individuals, the same way as before in Equation (4.10). Moreover $R_0$, is still found using the eigenvalue approach in this general case as well.

## 4.3 Statistical Approaches

The methods used to estimate reproductive numbers depend heavily on the data available on infectious disease outbreaks, and here we wish to cover methods that are applicable with some of the most commonly collected types of data. In addition to this, the data that is available is also dependent on the stage of an outbreak at which we want to estimate the parameter. For instance, one method that we discuss looks at estimation using final size data and thus, the reproductive number can only be estimated retrospectively with that method. In contrast to this, some methods require data on daily counts of case notifications but can only be applied relatively close to the start of an epidemic where case counts grow at an exponential rate.

Other methods that use daily counts of cases allow for the reproductive number to be tracked as an epidemic is unfolding which has obvious advantages for surveillance. If counts of cases are available and an outbreak can be broken down into smaller clusters, then we would also be able to derive the sizes of those clusters which opens up another avenue for estimating the reproductive number; we describe one method that uses this *outbreak size* data. Finally, another type of data that is commonly collected for infectious diseases but can be costly and takes time to collect is data on serological markers of infection and so we outline one method that uses that data. To begin with, we look at at a method that can be used to estimate the reproductive number during the vital early phase of an epidemic.

### 4.3.1 Estimation Assuming Exponential Growth

For new infectious diseases such as SARS in 2003 and pandemic influenza in 2009, the number of cases tends to grow exponentially in the initial phase of an epidemic as most of the population is susceptible to the disease. Estimation of the reproductive number in the early phase allows us to obtain estimates before mitigation strategies have been put in place and also before any sizeable depletion of susceptibles. The rate of exponential growth in this early phase can be denoted as $r$ and is defined as the *per capita change in number of new cases per unit of time* by Wallinga and Lipsitch [371]. This can be obtained from numbers of daily case notifications during the initial phase of an epidemic.

Before the concept of $R_0$ saw widespread use in epidemiology, it was heavily developed in studies of ecology and demography by Dublin and Lotka [93] where it was used to describe the expected number of offspring born to each female during her life. The ideas of the growth of a population and the spread of disease are very similar and this is the reason why Wallinga and Lipsitch [371] start by relating the *Lotka-Euler equation* for the human population to the scenario for infectious diseases and so we begin with the equation

$$b(t) = \int_{a=0}^{\infty} b(t-a)n(a)\,da \tag{4.13}$$

Here $t = 0$ would be the present time and in the infectious disease context, $b(t)$ can be thought of as the rate of new infections at time $t$; $a$ denotes the time since infection (or *age of infection*); and $n(a)$ refers to the expected rate of creating secondary cases at time $a$ since infection. For more details on how notation differs between ecology and epidemiology, see the supplementary material for Wallinga and Lipsitch [371]. In the initial phase of an epidemic, we can assume that the number of cases increases exponentially with a growth rate $r$. One method of finding $r$ is by fitting a regression model (e.g. Poisson or negative binomial) to the daily numbers of cases (see McBryde et al. [237] for an example of this for 2009 pandemic influenza in Australia). We can write $b(t)$ as

$$\begin{aligned} b(t) &= b(t-a)e^{ra} \\ b(t-a) &= \frac{b(t)}{e^{ra}} \end{aligned} \tag{4.14}$$

Substituting this term for $b(t-a)$ into Equation (4.13) gives a new equation for $b(t)$.

$$\begin{aligned} b(t) &= \int_{a=0}^{\infty} b(t)e^{-ra}n(a)\,da \\ 1 &= \int_{a=0}^{\infty} e^{-ra}n(a)\,da \end{aligned} \tag{4.15}$$

Equation (4.15) gives the standard Lotka-Euler equation. As we have previously mentioned, $n(a)$ is the expected rate of creating secondary infections. Hence the reproductive number, $R$, is obtained by integrating $n(a)$ over all times since infection.

$$R = \int_{0}^{\infty} n(a)\,da \tag{4.16}$$

we can normalise the rate $n(a)$ to a probability distribution for serial intervals.

$$g(a) = \frac{n(a)}{\int_0^\infty n(a)\,da} = \frac{n(a)}{R} \tag{4.17}$$

We know $a$ is the time since infection and thus, the serial interval distribution is exactly the equation for $g(a)$. Now rearranging Equation (4.17) to get $n(a) = g(a)R$ and substituting this into Equation (4.15) gives

$$\begin{aligned} 1 &= \int_{a=0}^{\infty} R\,g(a)e^{-ra}\,da \\ \frac{1}{R} &= \int_{a=0}^{\infty} e^{-ra}g(a)\,da \end{aligned} \tag{4.18}$$

Equation (4.18) is the Laplace transform of the function $g(a)$ and if we set $z = -r$ (negative of the growth rate), we can get the moment generating function (MGF) of the serial interval distribution.

$$M_{g(a)}(z) = \int_{a=0}^{\infty} e^{za}g(a)\,da \tag{4.19}$$

Noting that Equations (4.18) and (4.19) are essentially the same, the MGF can be used to simplify Equation (4.18).

$$\begin{aligned}
\frac{1}{R} &= M_{g(a)}(z) \\
R &= \frac{1}{M_{g(a)}(z)} \\
R &= \frac{1}{M_{g(a)}(-r)}
\end{aligned} \tag{4.20}$$

$R$ is the inverse of the MGF of the serial interval distribution with a growth rate $r$. Moreover, the shape of the serial interval distribution uniquely characterises the relationship between $R$ and $r$. This means that to calculate $R$ using the method presented by Wallinga and Lipsitch [371], we only need two things – the shape of the serial interval distribution and the rate of exponential growth in the initial phase of an epidemic. Previous research studies that investigated the time intervals between primary to secondary transmission of a disease can be used to get an idea around the shape of the serial interval distribution (see Cowling et al. [72] and Hirotsu et al. [171] for examples on influenza) and as we mentioned before, the exponential growth rate can be found by fitting an exponential curve (e.g. Poisson regression) to the data on counts of cases. We can explain this method further through a few specific examples.

### SIR Model

The SIR model assumes that the serial interval distribution follows an exponential distribution. We define $\gamma$ as the rate of leaving the infectious stage (assumed constant). This implies that the mean serial interval is $T_c = \frac{1}{\gamma}$ and we have for the probability distribution and MGF

$$\begin{aligned}
g(a) &= \gamma e^{-\gamma a} \\
M_{g(a)}(z) &= \frac{\gamma}{\gamma - z} = \frac{\gamma}{\gamma + r}
\end{aligned} \tag{4.21}$$

Using Equation (4.20), $R$ is

$$\begin{aligned} R &= \frac{1}{\frac{\gamma}{\gamma+r}} = \frac{\gamma+r}{\gamma} \\ R &= 1 + r\frac{1}{\gamma} = 1 + rT_c \end{aligned} \tag{4.22}$$

Furthermore this definition can be related to derivation for $R_0$ in Equation (4.2). The exponential growth rate $r$ is equivalent to the transmission rate $\beta$ minus the rate of leaving the infectious stage $\gamma$, i.e. $r = \beta - \gamma$. Hence we have

$$\begin{aligned} R &= 1 + \frac{(\beta-\gamma)}{\gamma} \\ R &= 1 - 1 + \frac{\beta}{\gamma} = \frac{\beta}{\gamma} \end{aligned}$$

which is exactly equivalent to Equation (4.2).

### SEIR Model

We can extend the SIR model to an SEIR model fairly easily by adding in a rate for leaving the exposed stage. The serial interval distribution for an SEIR model would be assumed to follow a convolution of two exponential distributions. In fact, Equation (4.20) can be generalised for the case of an infection which has $n$ stages.

$$R = \frac{1}{M(-r)} = \frac{1}{M_1(-r) \times M_2(-r) \times \ldots \times M_n(-r)} \tag{4.23}$$

Here $M_1, M_2, \ldots, M_n$ are the MGFs for the duration between successive events in the infection cycle. So for an SEIR model, $M_1$ would use the MGF for the exposed stage and $M_2$ would use the MGF for the infectious stage. In this case, we can call the rate of leaving the exposed stage $\sigma$ while $\gamma$ would denote the rate of leaving the infectious stage and both rates would be assumed constant. This gives

$$\begin{aligned} R &= \frac{1}{M_1(-r) \times M_2(-r)} = \frac{1}{\left(\frac{\sigma}{\sigma+r}\right)\left(\frac{\gamma}{\gamma+r}\right)} \\ &= \frac{1}{\frac{\sigma\gamma}{(\sigma+r)(\gamma+r)}} = \frac{1}{\frac{\sigma\gamma}{\sigma\gamma+\sigma r+\gamma r+r^2}} \\ &= \frac{1}{1+\frac{\gamma}{r}+\frac{\sigma}{r}+\frac{\sigma\gamma}{r^2}} = 1+\frac{r}{\sigma}+\frac{r}{\gamma}+\frac{r^2}{\sigma\gamma} \\ &= \left(1+\frac{r}{\sigma}\right)\left(1+\frac{r}{\gamma}\right) \end{aligned} \tag{4.24}$$

We can note that the mean serial interval is $T_c = \frac{1}{\sigma} + \frac{1}{\gamma}$ in this SEIR model.

### Delta Distribution

A *Delta Distribution* is described by the function

$$g(a) = \begin{cases} 1 & \text{if } a = \mu \\ 0 & \text{otherwise} \end{cases} \tag{4.25}$$

This suggests that all secondary infections occur at time $T_c$. This imposes the assumption that there is no variation in the generation intervals which would almost certainly not be true in practice but nevertheless, it may sometimes be sufficient to describe the serial interval distribution of an infectious disease when there is very little variation in serial intervals. Wallinga and Lipsitch [371] suggest that this equation is best used to obtain an upper bound for $R$ rather than to obtain an accurate measure of $R$. We obtain $R$ as

$$\begin{aligned} M_{g(a)}(z) &= e^{\mu z} = e^{T_c(-r)} \\ R &= \frac{1}{M(-r)} = \frac{1}{e^{T_c(-r)}} = e^{rT_c} \end{aligned} \tag{4.26}$$

Note that since $r > 0$ and $T_c > 0$ are required conditions, $R$ is always greater than one here.

### Empirical Distributions

In the case that serial intervals do not clearly follow a standard probability distribution, we can simply observe the distribution via a histogram. For this method, it is required to define the category bounds of the histogram of the observed durations by $b_0, b_1, \ldots, b_n$, and also the observed relative frequencies of observed serial intervals by $y_1, y_2, \ldots, y_n$. Now each of the bars of the histogram can be described by a uniform distribution. For our example here, we will denote the upper bound of the bar by $b_2$ and the lower bound by $b_1$. Therefore each bar has the following distribution and MGF

$$
\begin{aligned}
g(a) &= \frac{1}{b_2 - b_1} \qquad \text{for} \quad b_1 \leq a \leq b_2 \\
M_{g(a)}(z) &= \frac{e^{b_2 z} - e^{b_1 z}}{(b_2 - b_1)z} = \frac{e^{b_2(-r)} - e^{b_1(-r)}}{(b_2 - b_1)(-r)}
\end{aligned}
\tag{4.27}
$$

When calculating $R$ we need to attach weights onto each of the bars which are represented by the relative frequencies of the bars (as described earlier by $y_i$). So therefore $R$ is

$$
\begin{aligned}
R = \frac{1}{M(-r)} &= \frac{1}{\sum_{i=1}^{n} y_i \frac{(e^{-rb_i} - e^{-rb_{i-1}})}{-r(b_i - b_{i-1})}} \\
&= \frac{-r}{\sum_{i=1}^{n} y_i \frac{(e^{-rb_i} - e^{-rb_{i-1}})}{(b_i - b_{i-1})}} \\
&= \frac{r}{\sum_{i=1}^{n} y_i \frac{(e^{-rb_{i-1}} - e^{-rb_i})}{(b_i - b_{i-1})}}
\end{aligned}
\tag{4.28}
$$

## 4.3.2 Real-Time Estimation

During the course of an epidemic, it is vital to obtain estimates of $R$ as the epidemic is developing. This allows timely decisions to be made, gives an idea of the scale of the effort required to bring a situation under control, and if mitigation strategies have already been implemented, then we have the possibility of observing whether or not the measures taken have been effective. It is for this reason, that perhaps the most important methods of estimating $R$ are those methods which can be utilised in real-time. Several methods have been developed in this context but we will focus on the method introduced and applied to SARS by Wallinga and Teunis [372]. This method is useful as we often have the types of data required for it and it imposes relatively few assumptions regarding the disease transmission process compared with some other methods.

Wallinga and Teunis provide a method which allows us to manipulate the time series of cases (using perhaps dates of symptom onset), along with a known (or assumed) distribution for the serial intervals, into a time series of estimated reproductive values on each day, denoted as $R_t$. As a simple example of what their method allows, we can consider an epidemic with just 3 cases $(c_1, c_2, c_3)$ who have dates of symptom onset $(t_1, t_2, t_3)$ with $t_1 < t_2 < t_3$. Using the serial interval distribution, Wallinga and Teunis' method gives a way of calculating the probability that $c_1$ infected $c_2$ and the probability that $c_1$ infected $c_3$. The expected value of $R$ for case $c_1$ at time $t_1$ is the sum of these probabilities. A practical example of where this method can be seen is in Nishiura et al. [268] where this method was used to investigate an outbreak of pneumonic plague in China.

Their method requires us to think about *infection networks* (or *transmission trees*) but these infection networks have certain restrictions as each case can only have exactly one primary case and cannot infect themselves. Therefore, in our infection network, infection is transmitted from one case to another and in principle, we would draw a link or an arrow between the two cases to show that they form a transmission pair. The method of Wallinga and Teunis is particularly useful when we have either no information or incomplete information relating to who infected whom as this method uses a likelihood-based estimation procedure which infers who acquired infection from whom and in essence, reconstructs the entire transmission tree based on the probabilities of infection for each case.

In a mathematical sense, we will have some epidemic curve $\mathbf{t} = (t_{(1)}, t_{(2)}, \ldots, t_{(m)})$, where $t_{(i)}$ could be the date of onset of symptoms of the $i$th person with infection. The individual who infects person $i$ (infector of person $i$) is denoted by $v_{(i)}$. The infection network is $\mathbf{v} = (v_{(1)}, v_{(2)}, \ldots, v_{(m)})$ and this vector specifies the infector of each case. We can denote the entire set of all infection networks which are consistent with the epidemic curve $\mathbf{t}$ by $V$. If we have an outbreak with reported dates of onset of symptoms for $n$ cases, of which $q$ come from outside the population (imported cases) then we have $n - q$ individuals who have a primary case within our population. This means that we have $(n-q)^{n-1}$ different network structures as for $n - q$ non-imported cases, there are $n - 1$ possible primary cases, assuming there is no temporal pattern. We can also note that $v_{(i)} = 0$ for imported cases as the infectors of imported cases do not come from within the population being considered. The fact that secondary cases can only have been infected by a primary case who acquired disease before them means that many of those network structures are assigned zero probability.

We assume that serial intervals for the population follow some distribution with probability density function $w(\tau|\theta)$ where $\tau$ is the serial interval and $\theta$ is a vector of parameters characterising the probability distribution. Note that $w(\tau|\theta) = 0$ for $\tau < 0$ (and perhaps $\tau \leq 0$ for some diseases). As an illustration, we can consider influenza, for which serial intervals have previously been found to closely follow a Gamma distribution [237] and also a Weibull distribution [73]. These two distributions can be characterised by shape and scale parameters and hence $\theta$ would contain two parameters for both of these distributions. We can define a likelihood function for the probability of observing an epidemic curve $\mathbf{t}$, given the parameters for $\theta$ and $\mathbf{v}$.

$$L(\mathbf{v}, \theta|\mathbf{t}) = \prod_{i:v_{(i)} \neq 0} w(t_{(i)} - t_{v(i)}|\theta) \tag{4.29}$$

Summing this likelihood over the set $V$ of all possible infection networks gives

$$L(V, \theta|\mathbf{t}) = \prod_{i:v_{(i)} \neq 0} \sum_{j \neq i} w(t_i - t_j|\theta) \tag{4.30}$$

In order to obtain the *relative likelihood* that a person $k$ was infected by a person $j$, we take the *likelihood ratio.*

$$p_{kj} = \frac{w(t_k - t_j|\theta)}{\sum_{i \neq k} w(t_k - t_i|\theta)} \tag{4.31}$$

This works intuitively as we are taking the likelihood of a person $k$ being infected by a case $j$ and dividing this by the likelihood that person $k$ was infected by any possible infectious case $i$. Note that in the summation we have $i \neq k$ to clearly show that cases cannot infect themselves. Equation (4.31) can be used to estimate an individual case reproduction number for one case $j$ which we can denote by $R_j$. This is estimated by summing all the $p_{kj}$ for a case $j$ which are the probabilities that case $j$ infected all other cases $k$. If there are a total of $n$ cases, of which $q$ are imported cases, then for each case $j$, we only sum probabilities for $n - q$ cases as the $q$ imported cases have been infected outside of the population and thus, do not contribute to $R_j$.

$$R_j = \sum_{k=1}^{n-q} p_{kj} \tag{4.32}$$

The expected number of secondary infections at a timepoint $t$ is the mean of all the individual reproductive numbers for those cases $j$ with onset of symptoms at time $t$; if the number of cases at time $t$ is $N_t$, then we have

$$R_t = \frac{1}{N_t} \sum_{j \,\in\, t_j = t} R_j \tag{4.33}$$

This is the estimate that we are usually most interested in when monitoring disease outbreaks. A more complete explanation around the mathematical derivations for the likelihood functions can be found in the appendix section of the study by Wallinga and Teunis [372].

### Limitations on Real-Time Tracking

Although this was termed as a 'real-time' method of estimating $R$, we must note that it only allows *nearly* real-time tracking of the $R$. This is because accurate estimates of the instantaneous reproductive number can only be made after a delay. It is necessary to wait until we can observe the number of persons a given primary case has infected; that is, we must wait a sufficient amount of time such that all secondary cases generated by that primary case have become infected, acquired symptoms and reported their infections. This lag period is slightly longer than the generation time, plus the incubation period plus the reporting delay. Hence, the delay can take days, weeks, months or years depending on the type of disease under consideration [221]. For diseases with a very long incubation period, this method could not really be termed as a real-time method due to the long lag period.

Another limitation of the method comes when an epidemic is censored as there will be unobserved secondary cases from cases that acquired infection near the censoring time. However, a simple method for predicting the late number of secondary cases when an epidemic is censored at an end time $T$ has been proposed before. The idea behind these predictions is that we can directly use the serial interval distribution to predict the late number of secondary cases and then use these predicted numbers to adjust the values of $R_t$ near the censoring time $T$ (more details can be seen in Cauchemez et al. [56]).

## 4.3.3 Estimation Using Outbreak Size Data

Farrington et al. [111] outline a method for estimating $R$ using data on the sizes of individual and distinct outbreaks of an infectious disease. Here we will briefly summarise their method of using outbreak size data to estimate $R$ by reproducing their result found when applied to measles data from the USA between 1997–1999 (see Orenstein et al. [274] for more details on the data). Their method draws upon some theory around *branching processes* [12].

### Branching Processes

In probability theory, the *offspring distribution* can be viewed using the theory of branching processes which are well suited for the purpose of epidemiological surveillance as they require only data on counts of cases [111]. The simplest form of branching process posits that the number of offspring is independent across time steps and is known as the *homogeneous Galton-Watson branching process*. In their original application, Galton-Watson processes were used to study the extinction of family names [379]. For our application, branching process theory provides a simple estimation method for the reproductive number where it is suggested that each individual in a population from some generation $n$ produces some random number of individuals in the next generation $n + 1$. The number of individuals produced gives rise to a distribution called the *offspring distribution* and is usually specified by some probability distribution (for example, Poisson). This process is repeated until the epidemic dies out (assuming it does). In an epidemiological sense, these individuals would correspond to the number of people who are infected with some particular disease and in turn, they may infect some random number of susceptible individuals with some probability in one time step. The most important result for us here would be the *offspring mean*, which would represent the reproduction number $R$ when applied in the context of infectious disease outbreaks.

Since this method requires the condition of extinction, it is best suited for the purpose of analysing epidemics with $R < 1$ as this is the requirement for extinction of an epidemic. In essence, it can be used to monitor situations where an epidemic was dying out but is perhaps, in danger of becoming endemic once again. An example of where this situation has arisen is with measles in the UK where vaccine uptake began to decline after 1998 due to discussions about alleged side effects of the vaccine; specifically a link between the measles, mumps and rubella (MMR) vaccine and autism. The claims of this link were later refuted [112, 354]. As a consequence of the controversy, larger outbreaks started to resurface, and it was shown that the epidemic was rising to an almost critical state ($R \approx 1$) [187].

### Example: Measles in the USA 1997–1999

We denote $x$ as the size of an outbreak generated from $s$ initial cases. Note that $x$ is the size of the outbreak including the $s$ initial cases. Then the distribution of outbreak sizes (or offspring distribution), $X$, comes from the power series family [25].

$$\mathbb{P}(X = x; s) = b(x, s)\frac{\lambda^{x-s}}{A(\lambda)^x} \tag{4.34}$$

where $b(x, s)$ is a constant and $A(\lambda)$ is a power series function, $A(\lambda) = \sum_{r=0}^{\infty} a_r \lambda^r$, with $a_r \geq 0$. Assuming the offspring distribution is Poisson, then $a_r = \frac{1}{r!}$ and we have $A(\lambda) = \sum_{r=0}^{\infty} \frac{1}{r!} \times \lambda^r = e^{\lambda}$. The total outbreak size follows what is known as the *Borel-Tanner distribution* [143], where $b(x, s)$ from Equation (4.34) is $\frac{sx^{x-s-1}}{(x-s)!}$.

$$\begin{aligned} \mathbb{P}(X = x; s) &= \frac{sx^{x-s-1}}{(x-s)!} \times \frac{\lambda^{x-s}}{(e^{\lambda})^x} \\ &= \frac{sx^{x-s-1}\lambda^{x-s}e^{-x\lambda}}{(x-s)!} \end{aligned} \tag{4.35}$$

In their application to outbreak sizes for measles in the USA, an outbreak was defined as including at least two cases including the initial case. Outbreaks ranged in size from 2–33 and all 41 outbreaks started with $s = 1$ initial case; details are summarised in Table 4.1. Since all outbreaks started with one initial case Equation (4.35) becomes

$$P_x = \mathbb{P}(X = x; s = 1) = \frac{x^{x-2}\lambda^{x-1}e^{-x\lambda}}{(x-1)!}, \quad x = s, s+1, \ldots \tag{4.36}$$

Note that $P_x$ was included as a simpler notation to make it easier to refer back to the result of Equation (4.36) later.

| **Outbreak Size ($x$)** | **Number of Outbreaks of size $x$ ($m_x$)** |
|---|---|
| 2 | 13 |
| 3 | 9 |
| 4 | 5 |
| 5 | 5 |
| 6 | 2 |
| 8 | 2 |
| 9 | 1 |
| 11 | 1 |
| 13 | 1 |
| 15 | 1 |
| 33 | 1 |

**Table 4.1:** Numbers of measles outbreaks of different outbreak sizes in the USA between 1997-1999 (see Farrington et al. [111]). Note that all outbreaks start from one initial case and all outbreaks are at least of size 2 including the initial case.

### Likelihood Function

If we denote $m_x$ as the frequency of outbreaks of size $x$ as in Table 4.1, and impose the condition of outbreaks having minimum size $X = 2$ cases, then the log-likelihood function is

$$l(\lambda; x, s) = \sum_{x=X}^{33} m_x \log P_x - \log\left(1 - \sum_{x=1}^{X-1} P_x\right) \sum_{x=X}^{33} m_x \tag{4.37}$$

Summing through the outbreak sizes $x$ with frequency $m_x$ from Table 4.1 gives log-likelihood kernels (log-likelihood function without additive constants, i.e. terms that do not depend on $\lambda$) of $166\log(\lambda) - 207\lambda - 41\log(1-e^{-\lambda})$ for the USA measles data. The best estimate $\hat{\lambda}$, is obtained by finding the value of $\lambda$ that maximises the log-likelihood.

### Profile Likelihood Confidence Interval

A 95% confidence interval for $\hat{\lambda}$ can be found from the profile log-likelihood [366]. If we have $G^2 = 2[l(\hat{\lambda}) - l(\lambda)]$, then the 95% CI consists of values for $\lambda$ for which $G^2 < 3.84$) (where the value 3.84 is the 95th percentile of a $\chi^2(1)$ distribution). This is equivalent to the values of $\lambda$, for which $l(\lambda) \geq l(\hat{\lambda}) - 1.92$.

Figure 4.5 shows the profile log-likelihood for $\lambda$. The MLE for $\lambda$ is $\hat{\lambda} = 0.66$ with a 95% CI of 0.55–0.78. Hence these can be taken as the estimates of $R$ for measles in the USA between 1997–1999.

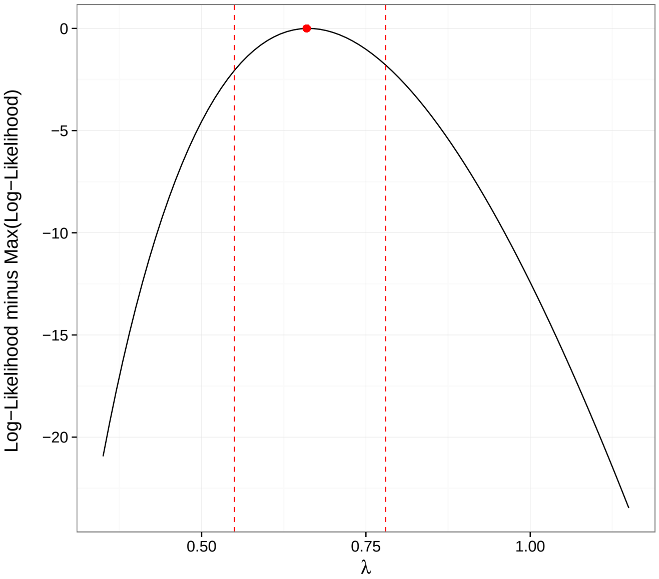


**Figure 4.5:** Log-likelihood profile for $\lambda$. Data comes from USA measles outbreaks between 1997-1999 and the result is reproduced from Farrington et al. [111]. The MLE is $\hat{\lambda} = 0.66$, 95% CI: $0.55 - 0.78$.

## 4.3.4 Estimation Using Final Epidemic Size Data

Becker [27] describes a method for estimating $R_0$ which requires only final epidemic size data and the application of *martingale* theory [387]. The basic idea around martingales is that for some stochastic process, the expectation of the next value is the same as the latest observed value and therefore, all other prior observed values are not beneficial for predicting the next value of the process. Since this method of estimating $R_0$ requires final epidemic size data, it brings with it the natural disadvantage of only being of use when an epidemic has finished. Hence, the method cannot be used to aid in the decision-making of mitigation strategies during an epidemic. Furthermore, when analysing a country-wide or worldwide epidemic, we generally will not accurately know the final epidemic sizes but it can be used in smaller confined environments such as households, schools and farms where it is much easier to know the initial and final numbers of cases. Despite these limitations, this method can still be used to give an estimation which can be compared against the estimations of $R_0$ found using other methods.

### Epidemic Model

To begin, we must consider a community which initially has $S(0) = n$ susceptible individuals and $I(0) = m$ infective individuals. An assumption here is that during the course of the epidemic, there are no further infections resulting from contact with individuals from outwith this community. We introduce $N(t)$ as the number of individuals who become infected during a time interval $(0, t]$ and initially, $N(0) = 0$. $S(t)$, $I(t)$ and $R(t)$ denote the number of susceptibles, infectives and removed individuals at time $t$, respectively. Finally we define $H(t)$ as the history of the process $(N, I, R)$ up to time $t$. Now, with an infection rate $\beta$ and a recovery rate $\gamma$, the progress of the epidemic can be described as

$$\begin{aligned}
&\mathbb{P}[dN(t) = 1,\, dR(t) = 0 \,|\, H(t)] = \beta I(t)S(t)\,dt \\
&\mathbb{P}[dN(t) = 0,\, dR(t) = 1 \,|\, H(t)] = \gamma I(t)\,dt \\
&\mathbb{P}[dN(t) = 0,\, dR(t) = 0 \,|\, H(t)] = 1 - \beta I(t)S(t)\,dt - \gamma I(t)\,dt
\end{aligned} \tag{4.38}$$

where the first line in Equation (4.38) follows the proportion of susceptibles, the second line follows the proportion of infectives, and the third line follows the proportion of removed individuals (simply calculated by using the rules of complementary events for probability).

### Martingale Method

Firstly we assume that the only information we have is $S(0)$, $I(0)$, $S(T_N)$ and $I(T_N)$, or in other words, we only know the initial numbers of susceptibles and infectives as well as the numbers of susceptibles and infectives at the end of the outbreak. To estimate $R_0$, we introduce two processes which are *zero-mean martingales*, $M_1$ and $M_2$.

$$\begin{aligned} M_1(t) &= N(t) - \int_0^t \beta I(x)S(x)dx \\ M_2(t) &= R(t) - \int_0^t \gamma I(x)dx \end{aligned} \tag{4.39}$$

Becker [27] points out that $S(t)$ and $I(t)$ are often not observable throughout time, so therefore, it is required to construct martingales consisting of only observable quantities. In order to do this we create an indicator function of there being a positive number of susceptibles, $J(x) = \mathbb{I}(S(x) > 0)$. Therefore, $J(x) = 1$ when $S(x) > 0$ and 0 otherwise. Furthermore, we can let

$$B(x) = \frac{J(x-)}{S(x-)} \tag{4.40}$$

where the notation $(x-)$ represents the observation immediately prior to $x$. Note here that $B(x) = 0$ when $J(x-) = 0$. Now a process $M_1^*(t)$ is defined by integrating the function $B(x)$ with respect to $M_1$.

$$M_1^*(t) = \int_0^t B(x)dM_1(x) = \int_0^t B(x)dN(x) - \beta \int_0^t I(x)J(x)dx \tag{4.41}$$

$M_1^*$ is also a zero mean martingale as integrating the function $B(x)$ with respect to another zero mean martingale will result in another zero mean martingale [8, 27]. It should be noted that $\int_0^t I(x)J(x)\,dx = \int_0^t I(x)\,dx$ when $S(t-) > 0$. Moreover, $\int_0^t B(x)\,dN(x)$ can be described in the form of a discrete approximation.

$$\int_0^t B(x)\,dN(x) = \frac{1}{s} + \frac{1}{(s-1)} + \ldots + \frac{1}{S(t-)} \tag{4.42}$$

This equation is completely determined by $S(t-)$. We now construct another zero-mean martingale $M$ which includes $R_0$ as part of the equation (also keeping in mind that $R_0 = \frac{\beta}{\gamma}$ for the SIR model, see Section 4.2.1).

$$\begin{aligned} M(t) &= M_1^*(t) - R_0 M_2(t) \\ &= \int_0^t B(x)\,dN(x) - \beta\int_0^t I(x)J(x)\,dx - R_0 R(t) + R_0\int_0^t \gamma I(x)\,dx \\ &= \int_0^t B(x)\,dN(x) - \beta\int_0^t I(x)J(x)\,dx - R_0 R(t) + \frac{\beta}{\gamma}\int_0^t \gamma I(x)\,dx \\ &= \int_0^t B(x)\,dN(x) - R_0 R(t) + \beta\int_0^t I(x)[1-J(x)]\,dx \end{aligned} \tag{4.43}$$

which is still a zero-mean martingale as any linear combination of martingales still yields a martingale [172]. It is assumed that the final size of the disease outbreak can be observed. Thus, $R$ (eventual number of removed/recovered individuals) and $\int_0^t B(x)dN(x)$ are both known, leaving only one nuisance term $\int_0^t I(x)[1-J(x)]dx$ which can be eliminated by introducing a suitable *stopping time*. This stopping time will be denoted as $T_N$ and is the time when the infection process ends. This can happen in two ways – when there are no more susceptibles left or when there are no more infectious individuals in the community. Therefore $T_N$ can be defined as

$$T_N = \inf\{t \geq 0 \,|\, [S(t) = 0] \vee [I(0) + N(t) - R(t) = 0]\} \tag{4.44}$$

where $\vee$ denotes an 'OR' condition. Since we are dealing with zero-mean martingales, we can equate Equation (4.43) to its mean which is of course, zero and then solve with respect to $R_0$. To do this we first deal with the case when $S(T_N) > 0$ and hence $I(T_N) = 0$. We can note that when $S(T_N) > 0$ then $\int_0^{T_N} I(x)[1-J(x)]\,dx = 0$

since $J(x) = 1$ when $x < T_N$ and also make use of the definition in Equation (4.42) to solve for $R_0$.

$$\begin{aligned}
\hat{R_0} &= \frac{\int_0^t B(x)dN(x) + \beta \int_0^t I(x)[1 - J(x)]dx}{R(t)} \\
&= \frac{\int_0^t B(x)dN(x)}{R(T_N)} \qquad (4.45)\\
&= \frac{1}{R(T_N)} \times \left[\frac{1}{s} + \frac{1}{(s-1)} + \ldots + \frac{1}{[S(T_N) + 1]}\right]
\end{aligned}$$

However, we have one final issue to deal with and that is the case of there being no more susceptibles. If $S(T_N) = 0$, then a cumulative amount of infectious period is *wasted* as some time is spent when there are no more susceptibles to actually infect so it is necessary to correct for this problem. To solve this problem, we introduce a time $T_R$ which is the occurrence time of the last recovery.

$$T_R = \inf\,[t \geq 0 \,|\, I(0) + N(t) - R(t) = 0] \qquad (4.46)$$

If $S(T_N) = 0$ then $T_R > T_N$, which literally means that if there are no more susceptible individuals, then the time of the last recovery must occur after the time when there are no more susceptibles. Here, Becker [27] argues that if $S(T_N) = 0$ then the time $T_N$ is not identifiable as when $S(T_N) = 0$, then $I(x)[1 - J(x)]$ is no longer zero. As a consequence it is no longer possible to solve $R_0$ from Equation (4.43). However De Jong and Kimman [80] state that the problem is purely mathematical and in fact, there is no reason why $T_N$ cannot be observed in this case. If we suppose that $T_N$ can be observed for this case, then we can calculate $R_0$ very similarly to before. First we define $c$ as the amount of infectivity excreted while susceptibles are still present in the community.

$$c = \frac{\int_0^{T_N} I(x)}{\int_0^{T_R} I(x)} \qquad (4.47)$$

Now we can construct a zero mean martingale just like in Equation (4.43) which we denote here as $M_3$.

$$\begin{aligned}
M_3(T_R) &= M_1^* - cR_0M_2 \\
&= \int_0^{T_R} B(x)\,dN(x) - \beta \int_0^{T_R} I(x)J(x)\,dx - cR_0R(T_R) + cR_0 \int_0^{T_R} \gamma I(x)\,dx \\
&= \int_0^{T_R} B(x)\,dN(x) - \beta \int_0^{T_R} I(x)J(x)\,dx - cR_0R(T_R) + c\frac{\beta}{\gamma} \int_0^{T_R} \gamma I(x)\,dx \\
&= \int_0^{T_R} B(x)\,dN(x) - cR_0R(T_R) - \beta \int_0^{T_R} I(x)J(x)\,dx + c\beta \int_0^{T_R} I(x)\,dx \\
&= \int_0^{T_R} B(x)\,dN(x) - cR_0R(T_R) + \beta \int_0^{T_R} I(x)[c - J(x)]dx = 0
\end{aligned} \tag{4.48}$$

Since $c$ is the proportion of infectivity excreted while there are still susceptible individuals left, the last term in Equation (4.48) becomes equal to zero.

$$\begin{aligned}
\int_0^{T_R} I(x)[c - J(x)]dx &= \int_0^{T_R} [cI(x) - I(x)J(x)]\,dx \\
&= c\int_0^{T_R} I(x)\,dx - \int_0^{T_R} I(x)J(x)\,dx = 0
\end{aligned} \tag{4.49}$$

So now we are left with just $M_3(T_R) = \int_0^{T_R} B(x)dN(x) - cR_0R(T_R)$ which can be rearranged and solved for $R_0$ analogous to the way previously in Equation (4.43).

$$\begin{aligned}
\hat{R_0} &= \frac{\int_0^t B(x)dN(x)}{cR(T_N)} \\
&= \frac{1}{cR(T_N)} \times \left[\frac{1}{s} + \frac{1}{(s-1)} + \ldots + \frac{1}{[S(T_N)+1]}\right]
\end{aligned} \tag{4.50}$$

To get an idea of the precision around these estimates of $R_0$, standard errors can also be derived. Details of the mathematics for deriving them can be found in

Becker [27] and Höhle [172]. One example of where this method has been applied was to data on an experiment looking at how vaccination reduced transmission of Aujeszky's disease virus (ADV) in pigs [80]. Using numbers available from the experiment for $S(0)$, $S(T_N)$ and $R(T_N)$ as well as the total numbers of pigs in each group, they found $R \approx 10$ in the unvaccinated group compared with $R \approx 0.5$ in the vaccinated group.

### 4.3.5 Estimation Using Seroprevalence Data

An epidemiological statistic which is of great importance is the proportion of individuals within a population who possess the specific antibodies required to fight a particular infection [9]. These antibodies are detected using various serological techniques based on immunological assay; consequently, this gives rise to the term *seroprevalence data*. These measures may be of a qualitative nature (presence or absence of antibodies) or a quantitative nature (e.g. antibody titres). If antibodies are detected, this tells us that the individual has had the infection previously or that they are currently infected. On the other hand, non-detection of antibodies does not necessarily imply that an individual has never experienced a particular infection.

Serological data can be obtained from surveys and will often be stratified by factors such as age and sex, providing us with important information on specific classes of a population who are more or less at risk of infection. Having data on the proportions of *seropositive* and *seronegative* individuals in a population can allow us to estimate a value for the reproductive number of a disease and for this reason, measuring levels of antibody production is of considerable interest to us. However, one issue with using this kind of data is that it is often only available after a reasonable length of time following the start of epidemics – for example, this study by Miller et al. [244] contains serological survey results for 2009 pandemic influenza in the UK but the study was published online in January 2010 which was well after the beginning of the pandemic. Although obtaining this data may take some time, estimates of $R$ using serological data are very useful as they offer a route to obtaining potentially more accurate estimates of $R$ as the serological data is a far less ambiguous marker for infection compared with diagnosis of cases through, for example, symptoms.

### Combining Seroprevalence Data with Social Contacts Data

For diseases which are transmitted by a respiratory or close contact route such as influenza, it is well known that social contact patterns are highly important in determining the spread of the disease. Surveys regarding numbers of conversational contacts that individuals of differing age groups have with each other have already been carried out before [248]. Although there are obvious concerns around these social contact surveys [170] such as inaccurate reporting of numbers of contacts due to memory recall issues, these surveys offer a conceivable improvement upon assuming mixing patterns such as homogenous or assortative mixing which are often made with a lack of empirical evidence to back it up. The assumption is that the number of potentially infectious contacts is proportional to the self-reported numbers of contacts found in these surveys.

Wallinga et al. [373] introduced a framework for estimating $R$ by combining data on social contacts together with data on the age-specific proportions of individuals immune to disease as taken from serological samples. The social contacts data that they used relates to the number of conversations people of different age groups had and comes from a random sample of people in Utrecht, Netherlands. The first step in their method is to estimate a social contact matrix $\mathbf{M}$ from this data and they do this by using a negative binomial model for numbers of conversational contacts.

A correction is made for the reciprocal nature of conversational contacts – total contacts from individuals belonging to age class $j$ with individuals from age class $i$ must be the same as total contacts from individuals belonging to age class $i$ with individuals from age class $j$. This correction was necessary as the surveys gave data which did not conform to this condition. In fact, this correction would most likely be necessary for any survey conducted for this purpose since numbers of contacts would probably never be reported with perfect accuracy.

They found parameters for $m_{ij}$ (mean number of conversations per week from individuals in age class $i$ reported by people in age class $j$) by maximum likelihood methods. We leave out specific details of estimating the contact matrix which can be found in Wallinga et al. [373] and concentrate more on estimating the transmission parameters in this summary. Note that other methods of estimating contact matrices are possible (see Goeyvaerts et al. [135] for another individual example and Hens et al. [165] for a more comprehensive review).

### Transmission Parameters

Given the mean numbers of contacts between individuals in different age groups $m_{ij}$, we have to find the next generation matrix $\mathbf{N}$ (giving age-specific numbers of potential transmission events per individual) as not every conversation will lead to transmission. This is found as

$$\mathbf{N} = (n_{ij}) = (q\, m_{ij}) \tag{4.51}$$

where $q$ denotes what is known as the *disease-specific infectivity* parameter which is estimated by keeping $m_{ij}$ fixed and searching for the value of $q$ which maximises the log-likelihood of observing the age-specific proportions of a population immune to the disease. Note that using the proportions immune to a disease would only be a suitable proxy measure for the proportion that were infected assuming that immunity was not gained through vaccination or if immunity due to vaccination and natural immunity gained through recovery from infection could be well distinguished such as for Hepatitis B. One way of ensuring that the effects of vaccination do not come into play here is to only consider diseases for which there is no vaccine or to look at proportions immune before the introduction of a vaccine.

In Equation (4.51), the elements of $\mathbf{N}$ essentially give the numbers of infectious contacts (as opposed to conversational contacts) individuals in age group $i$ make with individuals in age group $j$. Finding $q$ involves a longer process. We first define the proportion immune in age class $i$ as $z_i$. The estimated per capita number of infectious contacts that individuals in age class $i$ have during an epidemic is

$$\int \lambda_i(t)\,dt = \sum_j n_{ij} z_j \frac{w_j}{w_i} \tag{4.52}$$

where $\lambda_i(t)$ is the hazard rate of infection for age class $i$ and $w_i$ is the population size for age class $i$ which can be found from demographic data for a population. Therefore, the estimated probability of being infected during the epidemic in age class $i$ is

$$\begin{aligned} z_i &= 1 - \exp\left(-\int \lambda_i(t)\,dt\right) \\ &= 1 - \exp\left(\sum_j n_{ij} z_j \frac{w_j}{w_i}\right) \end{aligned} \tag{4.53}$$

More generally, we have observations on age groups $\tilde{a}$; the number of people tested in an age group is $g(\tilde{a})$ and the expected proportion immune is $f(\tilde{a})$. A binomial distribution with parameters $g(\tilde{a})$ and $f(\tilde{a})$ can be used to describe the numbers of immune people in an age group given by $h(\tilde{a})$. The log likelihood is thus

$$l(q;h) = \sum_{\tilde{a}} \ln \text{Bin}(h(\tilde{a}); g(\tilde{a}), f(\tilde{a})) \tag{4.54}$$

The best estimate of $q$ is found by searching for the value of $q$ which maximises the log-likelihood from Equation (4.54). From this point, we have estimated the parameters that allow us to derive the next generation matrix $\mathbf{N}$ and $R$ can be found as the largest eigenvalue of $\mathbf{N}$. As a further point, Wallinga et al. [373] showed that using the social contacts data provided superior fits to data on mumps and influenza compared with assuming homogeneous and assortative mixing. This comparison was made using the Bayesian Information Criterion [326]. However, they also showed that the estimates of $R$ for mumps were sensitive to the transmission model being used which underlines the importance of choosing an appropriate model.

## 4.4 Discussion

In this chapter we reviewed a number of methods of estimating reproductive numbers. The number of research papers that we looked at was by no means exhaustive, especially considering that more methods of estimating $R$ are always on the horizon. As we discussed, many of the methods can only be implemented when the appropriate types of data are available and some can only be applied during certain phases of an epidemic. Although there are often restrictions on what methods can be used to estimate $R$, it is important to have an awareness around different methods of estimation. This is because it can be beneficial to use more than one method to estimate $R$ if possible as if estimates are consistent between methods then this lends more assurance around the validity of the estimates.

In our review, we were primarily interested in methods which could be applied with types of data that are commonly available from routine surveillance such as daily case notifications and demographic data. However, we found that the serial interval distribution of a disease was essential for a number of methods of estimating $R$ [371, 372]. During epidemics, it is not uncommon for epidemiologists to investigate possible links between cases and in these circumstances, it may be possible to obtain the serial interval distribution for a disease. Although this data can be difficult to collect as it requires cases to be followed up and it can be difficult to find out exactly who infected whom, it may only be necessary to know about the general shape of the distribution as we discussed and this can be obtained from research studies that looked at past disease outbreaks that exhibited similar transmission patterns. An example for this could be to use the serial interval distribution found from a previous influenza outbreak and apply that for estimating $R$ in a new influenza outbreak. For these reasons, we also focussed on methods which use the serial interval distribution in addition to daily case data.

One of the methods that used the serial interval distribution in conjunction with daily case data was the method of Wallinga and Teunis [372]. This method can be utilised for surveillance purposes as for some diseases, it allows near real-time estimations of $R$ as an outbreak is unfolding. This is obviously important as it can aid in decision-making related to mitigation strategies. For example, if the reproductive rate is being tracked in almost real-time and it has been observed that the rate has been rising for a period of time, then this gives a clear sign that interventions would have to be considered urgently. Although it may be possible to obtain similar information by simply observing rises in numbers of cases, looking at the reproductive rate reveals more on precisely how fast an outbreak is growing and can also help to predict the size that an outbreak will grow to in a following period of time. One drawback with this method comes from the fact that it requires the serial interval distribution. If the duration of serial intervals for a disease is generally long, then the method cannot be applied in near real-time as for each case, the method estimates the probability that they infected future cases and these probabilities are based on the serial interval distribution. Therefore, the method requires data on future cases up to the point where there is almost no chance of secondary transmission for each primary case in order to obtain reasonably reliable estimates of $R$. Thus, for a disease such as SARS which has a serial interval of around a week [222], the method works well but it would not work so well for diseases that have much longer and much more variable serial intervals such as tuberculosis [369].

An important consideration for estimating $R$ is how accurate the data is for picking up true cases for a disease. With a disease like influenza, diagnoses may sometimes be made by clinicians based on symptoms and this can potentially pick up false-positives as many other diseases cause similar symptoms. Hence, we also discussed a method of estimating $R$ that made use of serology data [373], which can be expensive and take some time to collect. Despite these disadvantages, serology data offers the advantage of providing a much less ambiguous marker for infection and this may provide better estimates of $R$.

Besides the consideration around the data required for methods, we also aimed to present methods which could be used during different phases of an outbreak. Methods that look at the exponential growth phase of an epidemic can usually only be applied during the initial phase of an epidemic, but on the other side of the spectrum, some methods can only be applied retrospectively. We presented one method that uses very little data to estimate $R$ retrospectively – only data on initial and final size data for populations [27, 172]. This method is most applicable for small-scale outbreaks that are confined within certain locations such as farms or schools. The reason for this is that it is much easier to obtain initial and final size data under these circumstances.

In Table 4.2, we have attempted to summarise the methods that we described in this chapter in terms of the data requirements and the period of an epidemic that they can typically be applied. Note that the methods of estimating $R$ using mathematical models were not included in the summary table as the requirements for those models vary widely depending on the type of model being considered and the parameters that need to be estimated for the model. It is also important to note that although the table provides a succinct overview of the methods we reviewed, it is not possible to properly summarise these methods in such a concise fashion as there are many nuances which cannot be covered – for instance, the branching process method [111] does not necessarily need to be applied at the end of an epidemic but could be applied at the point where a reasonable amount of data on individual outbreak sizes is already available. Furthermore, for the method of Wallinga and Lipsitch [371] which looks at the exponential growth phase, it may be difficult to know when the actual peak of the growth phase occurs as an epidemic can have several peaks in numbers of cases followed by periods of decrease in numbers of cases. The consequence of this is that the estimates of the growth rate and hence, estimate of $R$ from that method, can be sensitive to the choice of the peak time chosen.

For this review, most of the attention was paid to describing different ways of estimating $R$, but we did not go into much detail on how to estimate confidence intervals around these estimates. Confidence intervals can be obtained for estimates using all of the methods we described but there are various ways to derive them. For instance, bootstrap methods [96] can be used to obtain confidence intervals around estimates of $R$ using the method of Wallinga and Teunis [372], and for the method of Wallinga and Lipsitch [371], we can derive confidence intervals around the estimate of the growth rate to subsequently get confidence intervals around estimates of $R$.

In the next chapter, we revisit the early phase of 2009 pandemic influenza in Scotland and use some of the methods reviewed here to estimate the reproductive number in that initial phase. In addition to using established methods of estimating $R$, we also explore the feasibility of incorporating spatial data to derive estimates of $R$. Specifically, we want to investigate if improved estimates of $R$ can be obtained by also using location data on where cases reside as this is one of the routinely collected variables on cases in Scotland.

| Method | Main Reference | Data Requirements | Most Suitable Epidemic Phase | Further Notes |
|---|---|---|---|---|
| Exponential Growth and MGF | Wallinga and Lipsitch [371] | Daily case notifications, serial interval distribution | Early phase | Applicable up to the peak of the exponential growth phase |
| Likelihood-based Estimation | Wallinga and Teunis [372] | Daily case notifications, serial interval distribution | All phases | Can get almost real-time estimates for diseases with short serial intervals |
| Branching Process Method | Farrington et al. [111] | Outbreak sizes | End of outbreak | Requires cases to be grouped into clusters to get outbreak sizes |
| Martingale Method | Becker [27] | Initial and final numbers in the susceptible and removed classes | End of outbreak | More applicable for small outbreaks in confined environments |
| Seroprevalence and Social Contacts | Wallinga et al. [373] | Serology and numbers of social contacts | End of outbreak | Social contacts data is not necessary if contact matrix is assumed |

**Table 4.2:** Summary of the different methods of estimating the reproductive number outlined in this chapter. The methods are described in terms of the data required to apply the methods and the time period of a disease outbreak that the method can be typically applied to.

# Chapter 5

# Estimation of Reproductive Numbers for 2009 Pandemic Influenza in Scotland

To prepare and plan effective responses as an epidemic unfolds, it is vital to obtain estimates of some key epidemiological parameters. In particular, these are crucial for producing effective surveillance systems. In this chapter, we make use of data on case reports of 2009 pandemic influenza in Scotland during the early phase (April to June 2009) to estimate firstly, the serial interval distribution, and then we use the the serial interval distribution to estimate the reproductive number $R$, which is the main goal of our analysis here. In Chapter 4 we outlined several different methods of estimating $R$ but as we explained there, the choice of method which can be used is highly dependent on the data and information available regarding the epidemic. Therefore, many of the methods we introduced previously can be eliminated from consideration.

We chose not to use the mathematical modelling methods such as fitting SIR or SEIR models as they require precise knowledge on parameters such as the recovery and transmission rates of the disease which we do not have. These are often not available for novel disease outbreaks, although if the disease shares similar characteristics to previous epidemics then parameters found from them can sometimes be used to obtain preliminary results. Other methods which we felt could not be considered include the branching process methods as they are more suited to epidemics which are endemic or dying out ($R \approx 1$ or $R < 1$); methods which utilise final epidemic size data obviously cannot be used as we have only early phase epidemic data here; and finally we disregarded the method involving social contact and serological data as serological data was not available in the early phase. We note that social contacts data is available in the UK (see Mossong et al. [248]), but this is not specific to Scotland.

Two methods which we can use here are the methods of Wallinga and Lipsitch [371] and Wallinga and Teunis [372]. The first of those two methods estimates $R$ from an initial growth rate found from the early phase of an epidemic, while the second of those methods attempts to track $R$ over time and is more of a continual surveillance method. The reason for deriving $R$ using more than one method is that estimates can potentially be sensitive to the method used and so it is important to explore the range of estimates produced using different methods.

The dataset we used is incomplete in that it contains a dip in case reports due to a period when laboratory testing of cases was temporarily suspended. Thus, we worked with a total of three datasets – the original dataset and two additional datasets that contain imputed data. We named these datasets as the *original dataset*, the *imputed dataset* and the *augmented dataset* (further details on why and how imputations were made will be explained later in Section 5.1.1).

In Scotland, postcode data on cases is routinely collected as well as the dates of onset of symptoms and so this presented us with an opportunity to explore the effects of adjusting estimates from the method of Wallinga and Teunis [372] by adding in a spatial term to their formula. Our reasoning for including a spatial term is that we believe those cases which live nearer to each other should accordingly have a higher likelihood of being linked while pairs of cases living further away from each other have less chance of being linked. Finally, we created two epidemic simulations models (a basic model along with a more complicated and realistic version) so that we could compare estimations of $R$ made using the Wallinga and Teunis method against our *spatial method*.

## 5.1 Data

Individual level data was available on 1,383 laboratory-confirmed cases of pandemic (H1N1) influenza in Scotland during the early phase of the epidemic. More cases are available in this dataset compared with the dataset used in Chapter 3 (which has 1,157 cases) as this dataset is far less detailed and does not contain, for example, details of symptoms reported or antiviral use. The individual level data available here comprises details of the dates of onset of symptoms, dates of reporting the case to Health Protection Scotland (HPS), whether or not cases travelled internationally within the week before their illness was reported as well as demographic details (age, gender and residence postcodes). The first date of report occurred on the 25th April 2009 and the last reported date was on the 7th of July – at this point the outbreak did not cease but the collection of detailed information on all reported cases was halted as the *containment phase* of the outbreak in Scotland had ended. In addition to this, a subset of 121 cases had given details of who they most likely acquired infection from; these contacts were all cases recorded in the dataset. It is presumed that these 121 cases were infected by their reported contact (although we cannot be absolutely certain of this), and so this allows for a reasonably accurate calculation of serial intervals in this group of cases.

Initially, dates of symptom onset were only available for 704 cases (as opposed to date of report which was known for all 1,383 cases), but the dataset was deterministically matched to a second dataset, which comprised a subset of these cases that had answered detailed questionnaires regarding their illness (see Chapter 3). Given the variables that were available in both datasets, matching was made based on four common variables between the two datasets to give the best chance of accurate and unique matches being found; these were date of birth, healthboard, postcode and gender. Note that name or identifying initials were not available so we could not match on these. This allowed the extraction of dates of symptom onset for a further 212 cases which gave us date of symptom onset for 916 cases. After matching, 467 cases were still missing dates of symptom onset, for which dates could be imputed.

There was also one more issue to be resolved. During the period of 15th to 21st June, the number of reported cases was artificially low (see Figure 5.1). In this period, there were too many reported cases to be handled and laboratory testing was temporarily suspended. Note that the dataset constructed after matching but before further data imputations is referred to as the *original dataset* in this chapter.

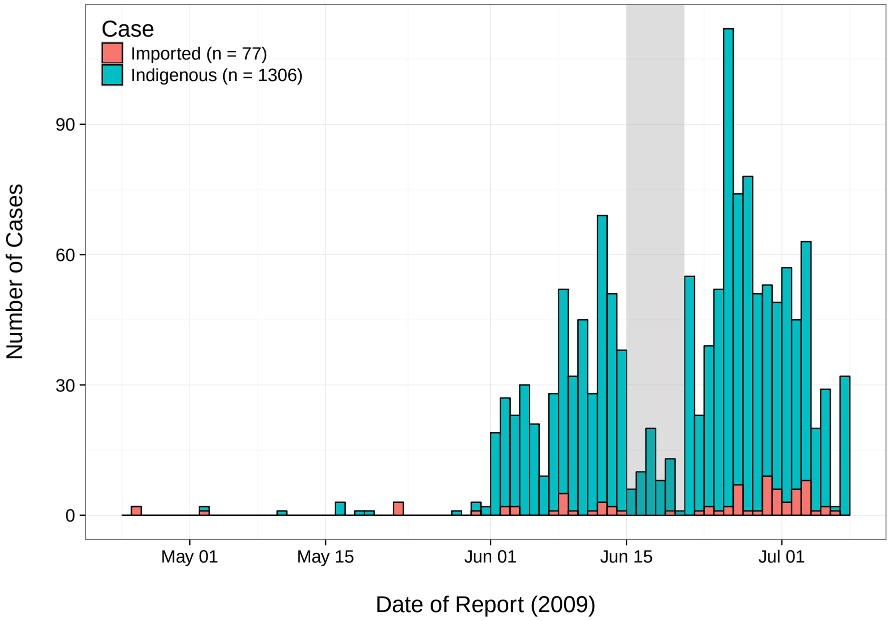


**Figure 5.1:** Number of early phase 2009 pH1N1 indigenous and imported case reports each day in Scotland. The highlighted area between 15th to 21st June 2009 shows a period where case numbers were under-reported as laboratory testing was temporarily suspended due to the high volume of case reports.

### 5.1.1 Imputation of Missing Dates of Symptom Onset

Cases still missing dates of symptom onset had dates imputed using the reporting delay following the method used by White et al. [382]. Firstly, the reporting delay distribution was derived for the 916 cases where we now had a date of symptom onset by taking the time difference in days between the date of report and the date of symptom onset. Figure 5.2 shows the reporting delay distribution, distinguishing between indigenous and imported cases. The reporting delay ranged between

0–21 days, but was predominantly between 2–7 days (around 80% of delays were within 2–7 days).

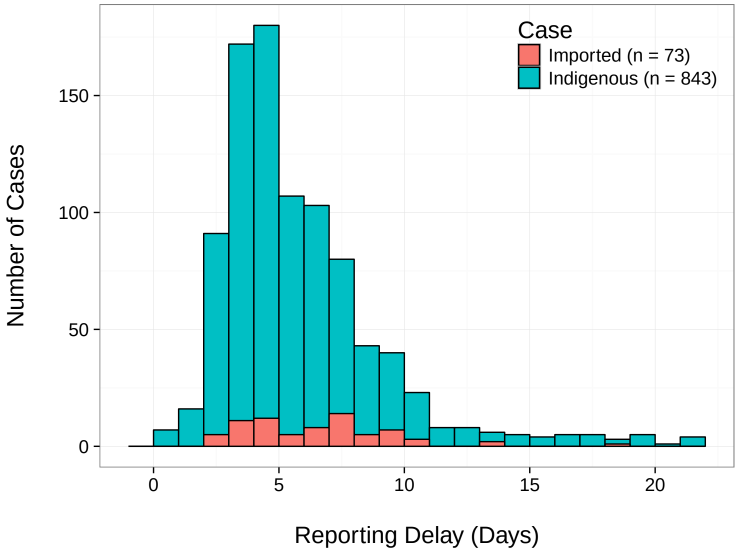
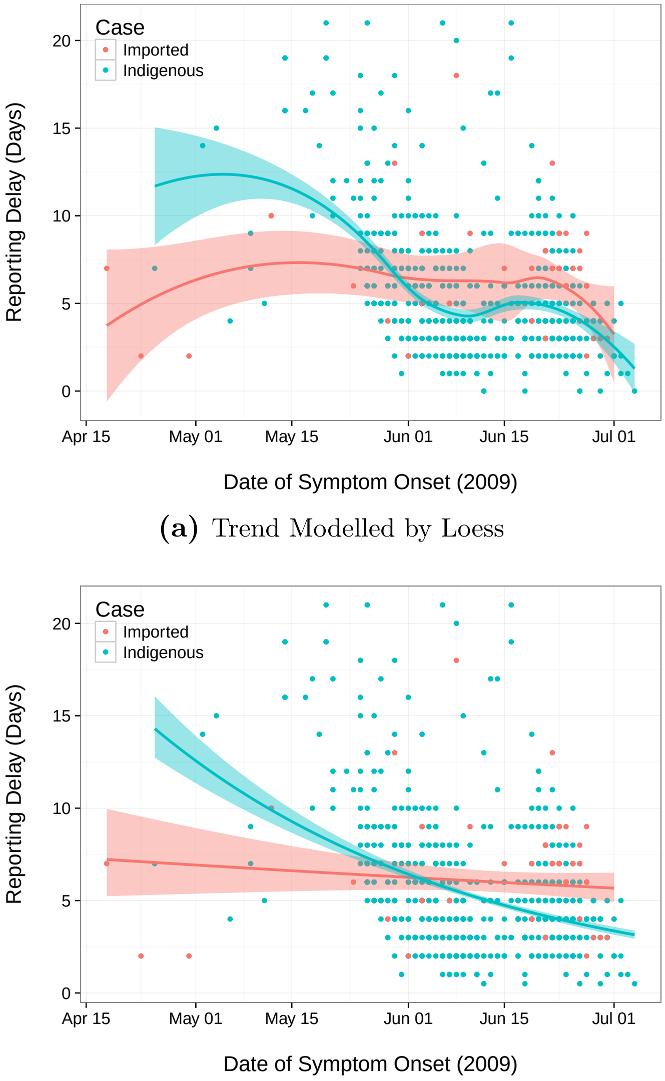


**Figure 5.2:** Reporting delay distribution for early phase 2009 pH1N1 cases in Scotland by indigenous and imported cases. Note that $n$ in the legend refers to the number of cases where the reporting delay was known.

During an epidemic, the reporting of illness may be quicker as the epidemic unfolds and public awareness increases. Figure 5.3a is a plot of the reporting delay against the date which cases had onset of symptoms. The loess curve (local regression) [66] illustrates that the reporting delay shortened as time progressed for indigenous cases while the delay was more constant for imported cases. Therefore, the reporting delay for each case $i$ was modelled using a linear regression with the reporting delay, $\log(d_{t[i]})$ as the response (logs were taken as the reporting delay distribution was skewed as seen in Figure 5.2) and two explanatory variables; the date of report, $r_{t[i]}$, and a variable indicating if a case was imported or not, $b_{t[i]}$. Each case missing an onset date (denoted $o_{t[i]}$), now had a predicted value for their reporting delay given their date of report and whether or not they were an imported cases, which we refer to here as $\hat{d}_{t[i]}(r_{t[i]}, b_{t[i]})$. The modelled trend for imported and indigenous over time can be seen in Figure 5.3b.

Following from this, a random variable $X_{t[i]}$ was generated for each case missing



**(a)** Trend Modelled by Loess

**(b)** Trend Modelled by Log-Linear Regression

**Figure 5.3:** Dates of symptom onset for early phase 2009 pH1N1 cases in Scotland against their reporting delay by indigenous and imported cases. Reporting delays are shown for 916 cases that had a known date of symptom onset. The fitted loess curve (a) illustrates that the reporting delay tended to decrease for indigenous cases as time progressed and the log-linear regression curve (b) shows our predictive model for the reporting delay.

an onset date. Specifically, $X_{t[i]}$ was generated as the exponential of a normally distributed random variable truncated at 21 days (the maximum reporting delay found in our dataset), with mean $\log(\hat{d}_{t[i]}(r_{t[i]}, b_{t[i]}))$ and variance taken as the prediction error from the regression model. The value for $X_{t[i]}$ was rounded to give an integer value representing the reporting delay in days. From here, the imputed date of onset was derived as $\hat{o}_{t[i]} = r_{t[i]} - X_{t[i]}$.

A single imputation was made for each missing date of onset rather than multiple imputation. This decision was made as one of our principle aims here is to compare estimations of reproductive numbers between different methods and a fairer comparison between methods can be achieved if one complete dataset is used between methods. However, we do note that the drawback in this is that variability due to the data imputation process would be better accounted for by using multiple imputation. In fact, this source of variability will essentially be ignored by using a single imputation.

The dataset containing the extra 467 dates of symptom onset is depicted in Figure 5.4 and is referred to as the *imputed dataset* from this point on. As all imputed dates of symptom onset were from the latter part of May 2009 onwards, this clearly shows that missing dates of symptom onset only started to become an issue during the part of the outbreak where numbers of case reports were rapidly increasing.

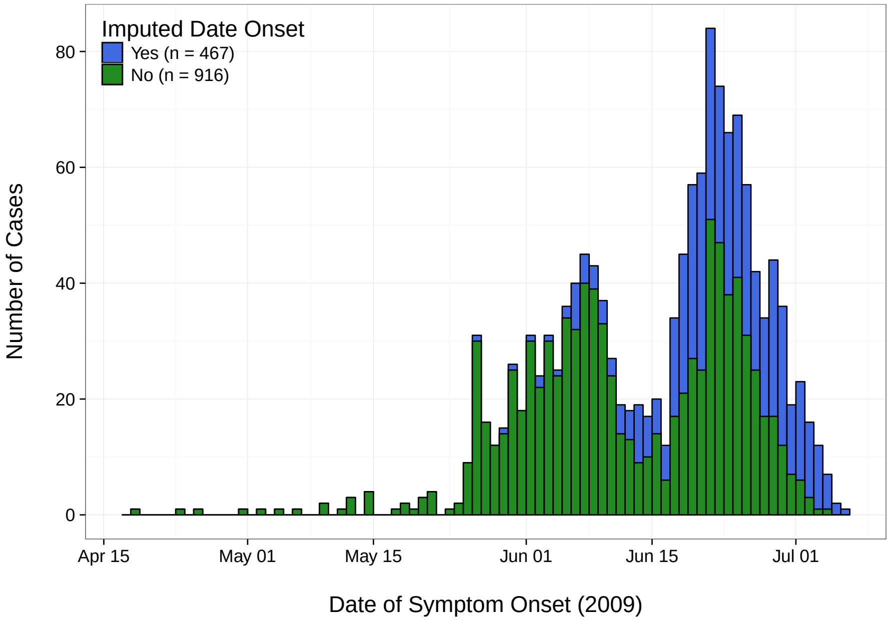


**Figure 5.4:** Dates of symptom onset for early phase 2009 pH1N1 cases in Scotland after imputing dates of symptom onset for 467 cases using the reporting delay.

### 5.1.2 Adjustment for Under-Reported Phase

As mentioned earlier, the six day period between 15th and 21st June 2009 suffered from under-reporting. For each day the number of cases which should have been reported was generated as a Poisson random variable with mean equal to 1.1 times the number of case reports one week earlier. For example, the number of case reports for Monday 15th June was generated using the number of case reports from Monday 8th June. More specifically, on 8th June there were 52 case reports so the number for 15th June would be estimated from a Poisson distribution with mean $1.1 \times 52 = 57.2$.

This process assumes that for the under-reported period, the number of new cases should have followed a similar pattern to the previous week but with some additional growth in the number of cases over time. We restricted the growth to 10% from one week to the next rather than higher values as we wished to err on the side of not overcompensating for the under-reported phase. We also note that too high a rate of growth would make it more likely for there to be more daily cases in this period than were observed in the day of the epidemic peak where there were 84 cases reporting symptoms (after previous imputation of dates of symptom onset), but our assumption was that case counts would build up to the peak.

After estimating the number of cases which should have been reported in this phase, the dates of symptom onset were imputed using the reporting delay as described before. Figure 5.5 shows the number of cases each day with augmented date onsets distinguished. There were still some noticeable dips and peaks in numbers of cases during the period which can be expected due to the stochastic nature of the two-step procedure used here – increasing numbers of reported cases in the under-reported phase and then imputing dates of symptom onset. From here, the dataset including the adjustments for the under-reported phase is called the *augmented dataset*. It is also important to remember that no individual level data is available for these augmented dates of symptom onset.

To summarise, we now have three datasets which can be used in analyses for

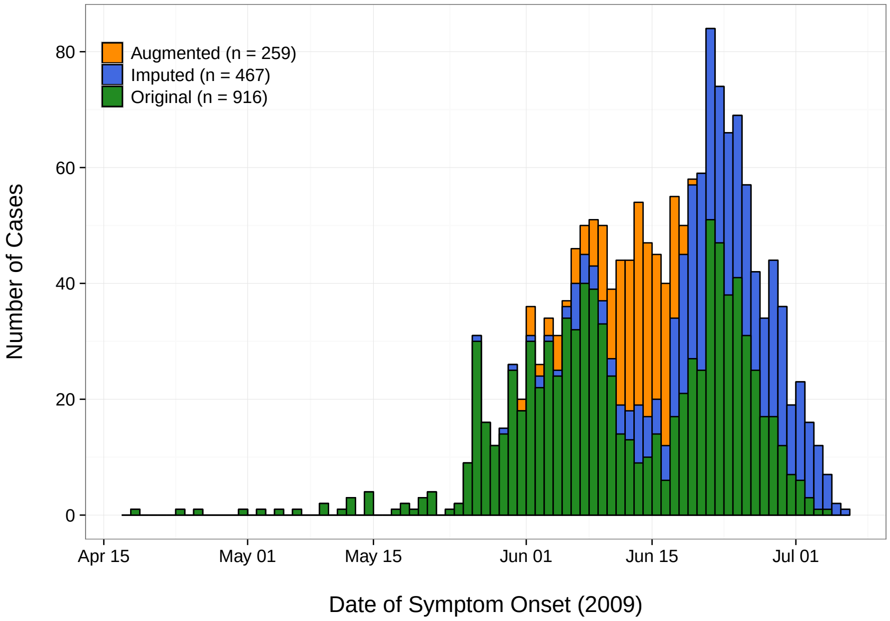


**Figure 5.5:** Dates of symptom onset for early phase 2009 pH1N1 cases in Scotland with additional augmented data for the under-reported phase. For the augmented data, additional case reports were generated between 15th to 21st June 2009 and then reporting delays were generated to derive dates of symptom onset.

estimating reproductive numbers for the early phase of the 2009 pH1N1 outbreak in Scotland. These are listed below.

1. **Original dataset** – before imputations for dates of symptom onsets and before including additional augmented dates of symptom onset for the under-reported phase. Available in all datasets (1), (2) and (3) where the numbers in brackets refer to item numbers in this list. $n = 916$.

2. **Imputed dataset** – after imputations for dates of symptom onset using the reporting delay distribution. Available in datasets (2) and (3). $n =$ 1,383 (original = 916, imputed = 467).

3. **Augmented dataset** – after imputations for dates of symptom onset and including additional augmented dates of symptom onset for the under-reported phase. Only available in dataset (3). $n =$ 1,642 (original = 916, imputed = 467, augmented = 259).

## 5.2 Serial Interval Distribution

Crucial to a number of methods of estimating $R$ is the serial interval distribution [371, 372]. We estimated the serial interval distribution by making use of a number of cases who were followed through a contact tracing exercise where these cases were asked who they had been in contact with over the previous seven days and then subsequently linked to previously reported cases; we shall refer to these cases as the *traced cases*. Thus, for 121 infectees, a presumed primary case was ascertained and the serial interval was calculated as the the interval between the date of symptom onset of the infectee and their presumed infector. Note that all of the cases used to calculate serial intervals already had dates of symptom onset and so they did not have to have a date imputed. If calculations were made using cases with imputed dates then any potential inaccuracies created in the imputation process would also affect the serial interval distribution derived here.

The serial interval ranged from 0–9 days as shown on Figure 5.6, with a mean of 3.44 days. Note that it was assumed that the characteristics of the serial interval do not change over the period considered, which may be reasonable as the data only looks at the early phase of the epidemic. A further assumption was that the set of traced cases consisted of a random sample of cases.

### 5.2.1 Distribution Choice

Distributions that do not allow negative values are suitable candidates for the serial interval data as it is not possible to have a negative serial interval. The fits of several parametric distributions (Gamma, Weibull, Log-Normal, Negative Binomial and Exponential) to the serial interval data were tested. Strictly speaking, the serial interval data comes in a discrete form which would make the Negative Binomial distribution the most applicable. However, it is possible to discretise the continuous distributions and thus, they can also legitimately be used. All of the candidate distributions use two parameters with the exception of the Exponential distribution which only uses a single rate parameter; this makes it less flexible in it's ability to model data compared to the other distributions.

The best fit parameters for the distributions were found by an optimisation procedure to find the maximum log-likelihood [308]. Note that for the Gamma, Weibull and Log-Normal distributions, serial intervals of 0 days were changed to 0.5 days indicating an interval of half a day as those distributions only support positive values. The distribution which fitted the serial interval data best (or offered the most parsimonious approximation to the data) was decided by using the AIC (Akaike information criterion) [5], which accounts for model complexity while considering goodness-of-fit. This is clearly shown in Equation (5.1) where $k$ is the number of parameters that need to be estimated and $L$ is the value for the maximum likelihood. Note that for our candidate distributions, model complexity only affects comparisons of the Exponential distribution against the other distributions since it has one less parameter.

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{5.1}$$

The Gamma distribution with shape parameter, $\alpha = 3.12$ and rate parameter, $\beta = 0.9$ was found to be the best-fitting distribution as it produced the lowest AIC (see Table 5.1); this curve has been superimposed onto Figure 5.6. As can be observed, the vast majority of serial intervals were between 2–4 days, but there was some overestimation in the frequency of one day serial intervals and some underestimation in the frequency of two day serial intervals – otherwise, the Gamma distribution fits the data fairly closely.

| **Distribution** | **Parameters** | **Range** | **Log-Likelihood** | **AIC** |
|---|---|---|---|---|
| Gamma | shape, rate | $(0, \infty)$ | −238.7 | 481.4 |
| Weibull | scale, shape | $[0, \infty)$ | −240.6 | 485.2 |
| Log-Normal | mean, variance | $(0, \infty)$ | −241.5 | 487.1 |
| Negative Binomial | successes, probability | $0, 1, 2, \ldots$ | −245.6 | 495.2 |
| Exponential | rate | $[0, \infty)$ | −270.4 | 542.8 |

**Table 5.1:** Log-likelihood values and AIC for probability distributions fitted to the serial interval data.

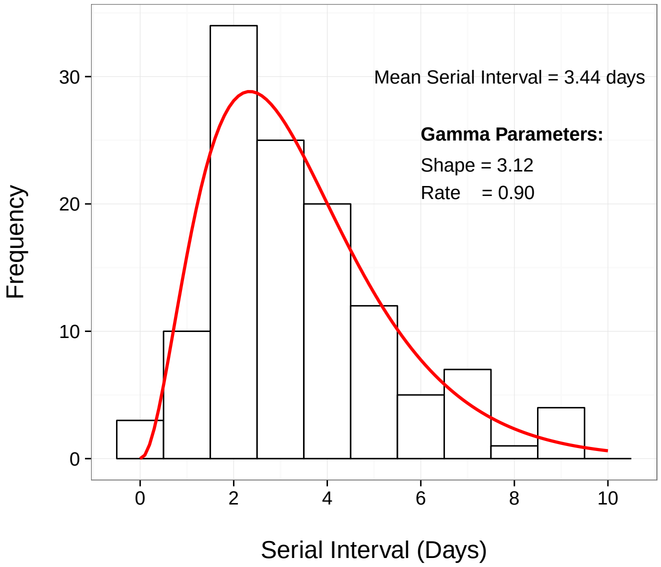
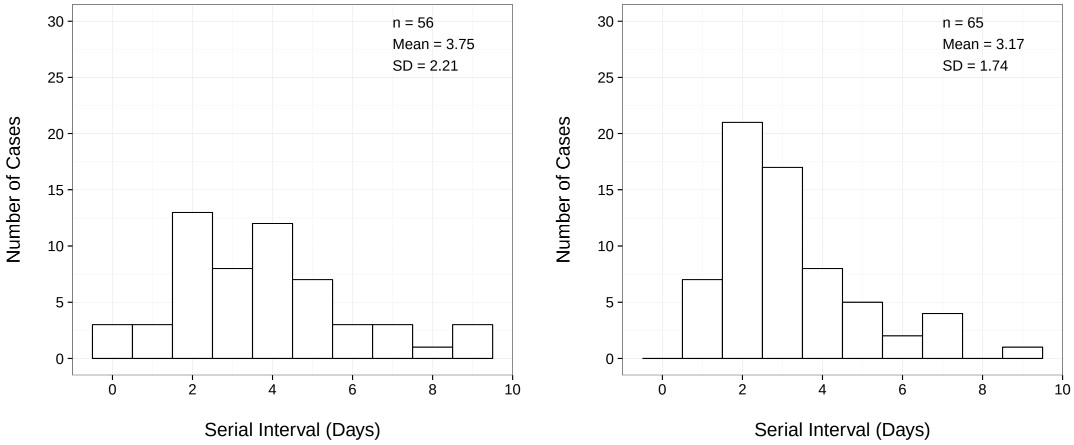


**Figure 5.6:** Serial interval distribution with best fitting Gamma curve superimposed.

## Assumption of Constant Serial Intervals

The methods we use to estimate $R$ in this chapter assume that the serial interval distribution is constant over time [371, 372], which may be questionable. To investigate this assumption, we split the data into two sets – in the first set, the presumed infector had a date of symptom onset on or before the 3rd of June 2009 ($n = 56$), and for the second set, the presumed infector had a date of symptom onset after the 3rd of June ($n = 65$). Note that the cut-off date here was chosen such that it would create two subsets of data that were reasonably equal in size.

As shown on Figure 5.7, the mean serial interval for those with a presumed infector before 3rd June was only slightly higher than those with a presumed infector after that date (means of 3.75 and 3.17 days for the earlier and latter periods respectively). In addition to this, the serial intervals in the latter period were also less variable with the standard deviation being 1.74 days in the latter period while the standard deviation was 2.21 days in the earlier period. A non-parametric Wilcoxon rank sum test also suggests that the serial interval was not significantly different in the first period compared with the second period ($W = 2{,}127$, $p > 0.05$). These results give some assurance that serial intervals stayed relatively constant



**(a)** On or Before 3rd June 2009

**(b)** After 3rd June 2009

**Figure 5.7:** Serial interval distribution split into two sets to investigate the assumption of constant serial intervals through time – those where the presumed infector had date of symptom onset on or before 3rd June 2009 and those with date of symptom onset after that date.

in the time period that we are examining.

### 5.2.2 Discretisation

The serial interval data we have is discrete as we have it as a time difference in days. Hence if we choose to model the serial interval distribution as a continuous distribution such as the Gamma distribution, it is necessary to discretise it. We discretised the distribution in the same way used by Roberts and Nishiura [309].

First the data was considered in the intervals $[0, 0.5), [0.5, 1.5), \ldots, [9.5, 10.5)$ to substitute for the discrete values $0, 1, \ldots, 10$ days. The maximum serial interval was set at ten days, which seemed reasonable as the maximum serial interval found in our data was nine days and thus, we assumed that the probability of secondary transmission after an infection age of more than ten days was negligible. Using the notation of $t$ as the time since infection, we have $t_{\max} = 10$ days, and the distribution was discretised as

$$s_t = \frac{S(t) - S(t-1)}{S(t_{\max})} \tag{5.2}$$

where $s_t$ denotes the discretised probability density for a serial interval, and $S(t)$ is the cdf (cumulative density function) of the Gamma distribution up to time $t$. For example, $s_1$ would be obtained as $s_1 = \frac{S(1.5)-S(0.5)}{S(10.5)}$. Another assumption made is that $s_0 = 0$, meaning that there is no probability for infections to occur on the same day. The discretised probabilities for each serial interval are displayed in Figure 5.8. As mentioned earlier, the chosen Gamma distribution over-predicts the number of cases with a serial interval of one day and under-predicts the number of cases with a serial interval of two days, but looks fairly accurate for the other serial intervals. We also tried discretisation using intervals of $[0, 1], [1, 2), \dots, [9, 10)$ (results not shown). However this resulted in an even larger overestimation in the frequency of one day serial intervals with respect to the observed data, whilst giving similar results otherwise.

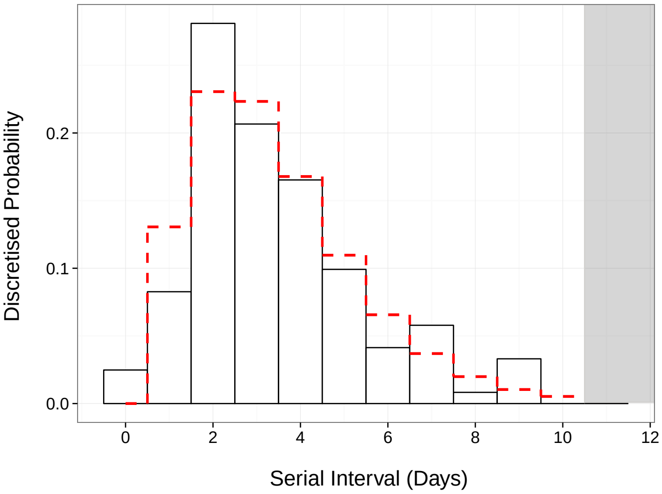


**Figure 5.8:** Discretised Gamma probabilities for serial intervals (shown by the red dashed line) superimposed onto the histogram of actual probabilities for serial intervals given by the data. Probability of secondary transmission after ten days was assumed to be negligible.

## 5.3 Estimation of Reproductive Numbers

### 5.3.1 Exponential Growth

Values of $R$ during the initial phase of an epidemic where case counts are usually growing exponentially can be calculated using the method of Wallinga and Lipsitch [371]. The method requires only knowledge on the shape of the serial interval distribution and the rate of exponential growth, $r$, to estimate $R$. Hence the first parameter that must be estimated is $r$, which can defined as the per capita change in the number of new cases per unit of time. The case reports come in the form of counts and so Poisson regression is ideally suited for estimating $r$ (although linear regression on log counts is also possible).

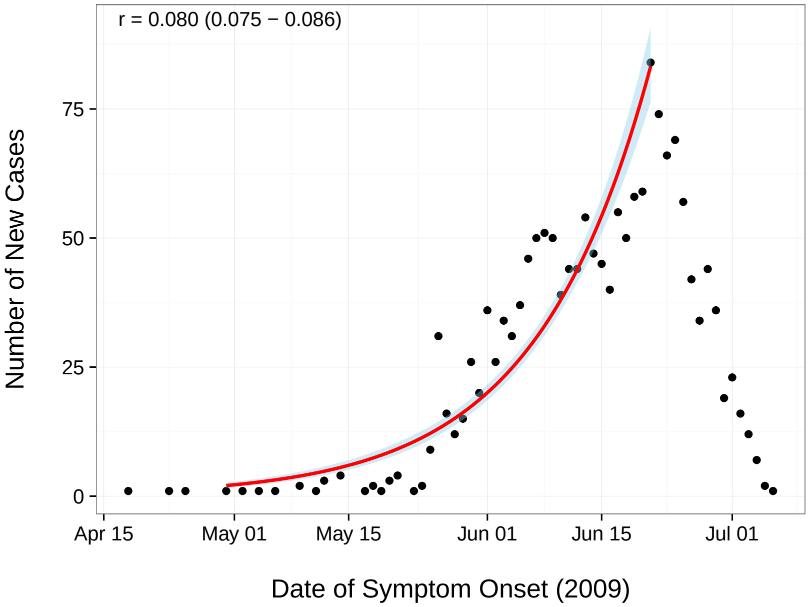


**Figure 5.9:** Exponential growth rate of 2009 pandemic influenza cases in Scotland in the early phase. The data used came from the augmented dataset and the growth rate was derived from a Poisson regression model. The start of the exponential growth phase was set as 30th April 2009 (the time when new cases began to appear almost every day) and the end was set as 21st June 2009 (the epidemic peak).

We chose the beginning of the exponential growth period as the 30th April 2009 since new cases started to appear almost daily from this point, while the end of the exponential growth period was chosen as the 21st June 2009 as this was the day that case numbers peaked (Figure 5.5). If we started the growth period from the earliest date of symptom onset for a case (18th April) as opposed to choosing the 30th of April, it would have been a less accurate reflection of the growth rate in the exponential phase. With an earlier start date, the growth rate would have been underestimated as at the beginning, there was not much secondary transmission of the disease as at that time, imported cases were coming into Scotland sporadically. Precisely, between 18–29 April there were only three cases and two of these were imported. The growth rate estimated from the Poisson regression model was $r = 0.080$ with 95% CI 0.075–0.086 (Figure 5.9).

Once $r$ has been calculated, $R$ can be estimated using the relationship $R = \frac{1}{M(-r)}$ where $M$ is the moment generating function (MGF) of the probability distribution used to model serial intervals (more details on the exponential growth method can be found in Section 4.3 of Chapter 4). However since we are using the discretised serial interval distribution rather than the continuous form, $R$ should be estimated using the formula for empirical distributions [371].

$$R = \frac{r}{\sum_{i=1}^{n} y_i(e^{-rb_{i-1}} - e^{-rb_i})/(b_i - b_{i-1})} \tag{5.3}$$

Here $b_i$ denotes the category bounds for serial intervals, and $y_i$ is the discretised probability related to that interval. Since we always consider intervals of one day, $b_i - b_{i-1} = 1$ and Equation (5.3) simplifies to

$$R = \frac{r}{\sum_{i=1}^{n} y_i(e^{-rb_{i-1}} - e^{-rb_i})} \tag{5.4}$$

Using Equation (5.4) and assuming the discretised Gamma distribution for serial intervals, $R$ for the exponential growth period was estimated as $R = 1.36$ (95% CI: 1.33–1.38). Note that the confidence limits for $R$ were estimated directly from the confidence limits found for the growth rate $r$ and thus, do not take into account uncertainty in the serial interval distribution.

### Sensitivity Analysis

The sensitivity of estimates of $R$ using the exponential growth method was tested against the choice of four variables; the start date and end dates of the exponential growth phase (start dates: 18 April, 30 April and 6 May; end dates: 7 June, 16 June, 21 June and 26 June), the distribution used to model serial intervals (Gamma, Weibull, log-normal and empirical) and the dataset used (original, imputed and augmented). We chose 7th June as one of the end dates for exponential growth as this was the epidemic peak before the drop due to the period of limited lab-testing while the other end dates were selected to test the sensitivity of estimates against small changes away from the observed epidemic peak.

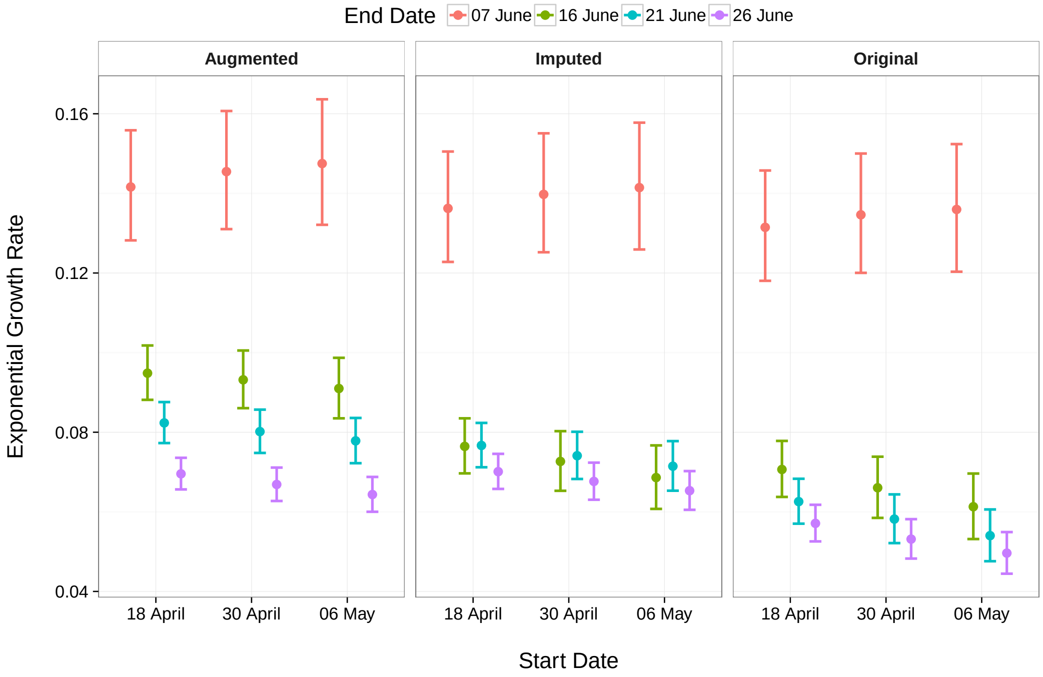


**Figure 5.10:** Sensitivity of growth rate $r$ to choice of dataset as well as start and end dates of the exponential growth phase.

The growth rate $r$ was not affected much when end dates close to the true epidemic peak (21st June) were used but using the much earlier end date of 7th June greatly increased $r$ (Figure 5.10). This is because the numbers of new cases were growing very rapidly in the period up to 7th June, but then the rate of new cases did not maintain that rate of exponential growth. This underlines the importance of trying to be quite certain that a true epidemic peak has been reached before using this method to derive an estimate of $R$ as failure to do so could result in unreliable estimates – in this case, a large overestimate. A further point to note is that much larger confidence intervals around $r$ were obtained when using 7th June as the end date as this meant that far fewer cases would be used in calculations. The dataset used also had a noticeable impact on $r$, which emphasises the importance of our earlier attempts to resolve data issues when using this method for estimating $R$. In general, the choice of start date for exponential growth generally had the least impact on $r$.

Estimates of $R$ by dataset and choice of probability distribution as well as start and end dates for the exponential growth phase are displayed in Figure 5.11. Since the growth rate $r$ has a direct effect on $R$, the same comments regarding $r$ can be repeated for $R$. The most obvious impact of this direct effect can be witnessed with the much larger estimate of $R$ when using 7th June as the end date ($R \approx 1.7$ compared to $R \approx$ 1.3–1.4 for the other end dates). The one additional inclusion here was the choice of probability distribution which had no major impact on estimates of $R$.

### Variability Due to Imputation

We mentioned in Section 5.1 that we decided to use single imputation for comparability reasons but also that a disadvantage of using single imputation for creating datasets was that we would not be able to account for the additional variability in estimates of $R$ due to data imputation. To investigate how much additional variability we could observe, we produced 1,000 augmented datasets by following the same process as described previously. For these datasets, we always used the discretised Gamma distribution for serial intervals with start date as 30th April and end date as 21st June to estimate $R$.

Since this is primarily an exploratory exercise, we simply obtained an average estimate for $R$ over the 1,000 datasets and also noted the lowest estimate for the

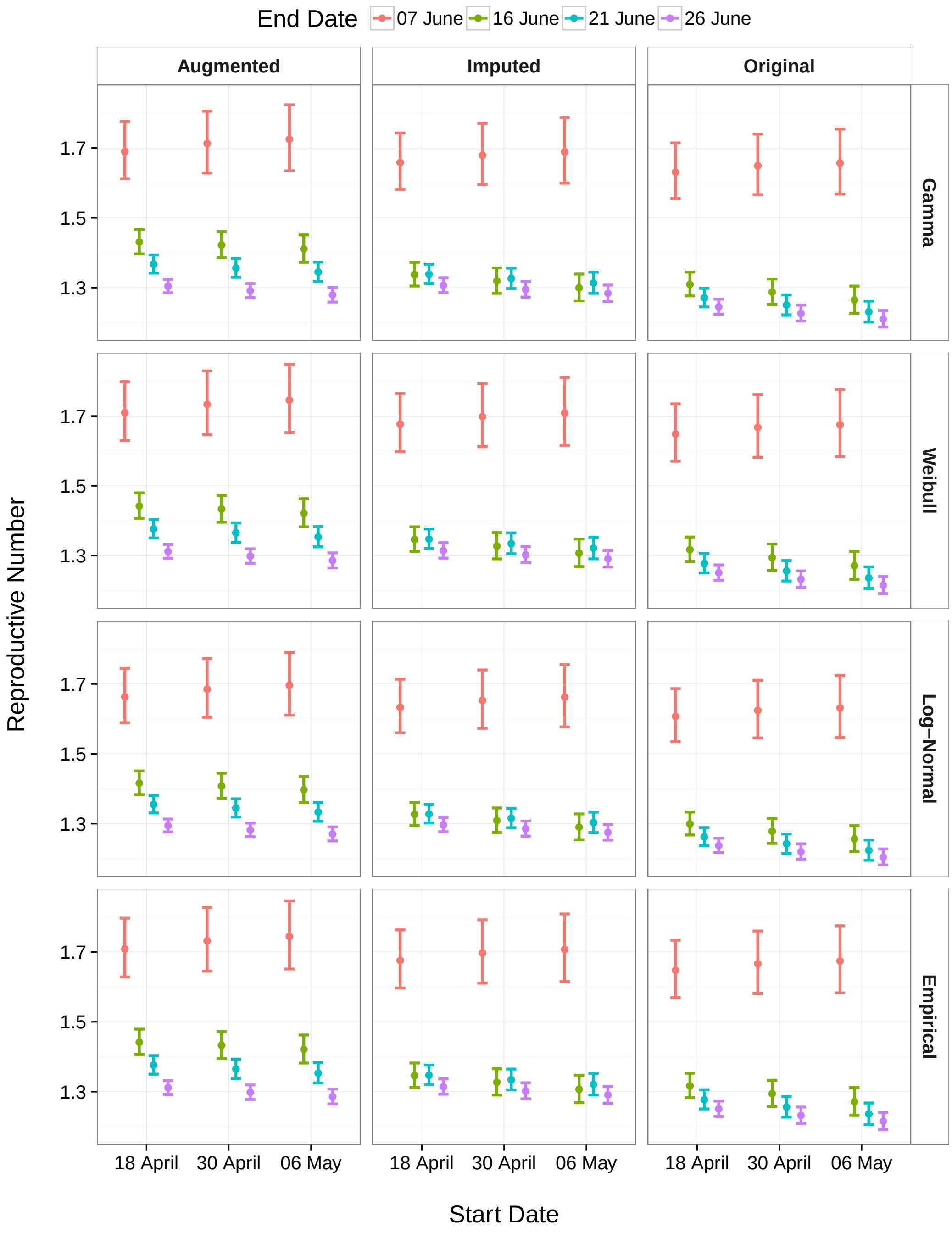


**Figure 5.11:** Estimates of $R$ using the exponential growth method of Wallinga and Lipsitch [371]. Sensitivity of estimates were tested by varying the choice of dataset and probability distribution as well as the start and end dates for the exponential growth phase.

the lower 95% confidence limit and the highest estimate for the 95% confidence limit; this gave us $R = 1.36$ (1.32–1.40). This is not hugely different to the original estimates ($R = 1.36$, 95% CI: 1.33–1.38) and therefore, this demonstrates that the data imputation process did not have a large impact on the estimates of $R$ using the exponential growth method. The reason for this is that the epidemic peak is what seems to have the largest effect on estimates of $R$ as we showed earlier in our sensitivity analysis and the data imputation mainly affects data leading up to the peak rather than the peak itself.

### 5.3.2 Time-Dependent Estimates

We obtained time-dependent estimates of $R$ (denoted $R_t$) for the pandemic influenza data using the method of Wallinga and Teunis [372] (from here on, we refer to this as the "W&T" method). Again, this method only requires knowledge of the serial interval distribution and a series of dates of symptom onset for cases (sometimes called the *epidemic curve*). For more complete details on the method, see Section 4.3.2 in Chapter 4 or the original reference. In brief, the probability that case $i$ was infected by case $j$ can be obtained using the serial interval distribution, $w(\tau|\theta)$ where $\tau$ is the serial interval and $\theta$ is the vector of parameters that defines the distribution. Moreover, let $t_i$ be the date of symptom onset for case $i$ and $t_j$ be date of symptom onset for case $j$.

$$p_{ij} = \frac{w(t_i - t_j|\theta)}{\sum_{k \neq i} w(t_i - t_k|\theta)} \tag{5.5}$$

This is the likelihood that case $i$ was infected by case $j$ relative to being infected by any case $k$. After calculating $p_{ij}$ for all cases, the concept can be turned around and we can look at this as also having the probabilities that each case $j$ infected all other cases. From this, individual reproductive numbers, $R_j$ can be found as $R_j = \sum_{i=1}^{\text{All Cases}} p_{ij}$. If we denote the number of cases with onset of symptoms at time $t$ as $N_t$, the effective reproductive number at time $t$, $R_t$, is worked out as

$$R_t = \frac{1}{N_t} \sum_{j \,\in\, t_j = t} R_j \tag{5.6}$$

Note that imported cases are distinguished in the method. The assumption is that imported cases were not infected by any cases in our data (indigenous or imported) but indigenous cases may have been infected by the imported cases, provided they acquired infection at a later time than the imported case.

### Confidence Intervals

Confidence intervals for values of $R_t$ can be found via simulation. Having obtained values of $p_{ij}$ for each non-imported case $i$, it is possible to simulate the infector of case $i$ based on those probabilities. Specifically, the infector of each case $i$ is simulated based on a multinomial distribution with probabilities $p_{ij}$. Following this, it is possible to get $R_j$ and subsequently $R_t$ for the simulated data using the same process as described before. The 95% confidence interval is computed by repeating this procedure over many simulations (say 1,000 or 10,000 times) and then taking the 2.5th and 97.5th quantiles.

### Correction for Unobserved Secondary Transmission

A method to account for as yet unobserved secondary cases in a set of epidemic curve data was introduced by Cauchemez et al. [56]. This is required as the epidemic curve data comes in the form of dates of symptom onset reported by cases up to a time $T$ and for those cases with onset of symptoms near time $T$, there will be some future cases that they will infect but have not yet been recorded in the current set of data (late secondary cases). As a consequence of this, $R_t$ near time $T$ will always be underestimated. Therefore, the method attempts to correct for this by predicting the eventual number of late secondary cases.

The proportion of secondary transmissions observed and unobserved after $t$ days from each infector is determined from the serial interval distribution shown in Figure 5.8. Since our serial interval distribution dictates that secondary transmission is negligible after ten days from the date of symptom onset, this problem only affects estimates of $R_t$ made at time $t > T - 10$ days; for estimates of $R_t$ made ten or more days before $T$, all secondary transmissions are assumed to have already been observed and thus, $R_t$ is unaffected. The corrections made to a value of $R_t = 1$ between 0–10 days before $T$ are given in Table 5.2. Note that in this period, the confidence intervals for $R_t$ are also adjusted in the same manner so this does not take into account uncertainty in the multiplier used for adjustment.

| **Days before $T$** | **Proportion Observed** | **Proportion Unobserved** | **Adjustment to $R_t = 1$** |
|---|---|---|---|
| 0 | 0.000 | 1.000 | – |
| 1 | 0.131 | 0.869 | 7.659 |
| 2 | 0.361 | 0.639 | 2.769 |
| 3 | 0.584 | 0.416 | 1.711 |
| 4 | 0.752 | 0.248 | 1.329 |
| 5 | 0.862 | 0.138 | 1.160 |
| 6 | 0.928 | 0.072 | 1.078 |
| 7 | 0.964 | 0.036 | 1.037 |
| 8 | 0.984 | 0.016 | 1.016 |
| 9 | 0.995 | 0.005 | 1.005 |
| 10 | 1.000 | 0.000 | 1.000 |

**Table 5.2:** Proportions of observed and unobserved secondary transmissions near the end of epidemic curve data up to time $T$. Proportions are based on the discretised Gamma distribution used to model serial intervals shown in Figure 5.8. The adjustments to $R_t$ are multiplier terms for $R_t$ and were calculated as one divided by the proportion observed, but note that proportions were rounded in the table so the calculation will not produce the exact multiplier shown. Adjustments do not affect estimates of $R_t$ made 10 days or more before $T$ while estimates of $R_t$ less than 10 days before $T$ are affected. Confidence intervals in this period are also adjusted using these multipliers.

### Time-Dependent Estimates for 2009 pH1N1 in Scotland

We chose to obtain time-dependent estimates of $R$ between the 30th of April and the 2nd of July 2009 as this matches the start of the exponential phase period we used before in Section 5.3.1, and the end date was chosen as there were less than ten cases daily after the 2nd of July. Figure 5.12 illustrates the estimates of $R_t$ over the aforementioned time period. Over the whole period from April 30th to July 2nd, the average estimate of $R$ for the augmented dataset was 1.26 (95% CI:

0.65–2.06), which is only slightly lower than estimates found using the exponential growth method (see Section 5.3.1). However, the confidence interval obtained here is much wider than that of the exponential method due to fluctuations in $R_t$ over time.

One notable feature is the explosion in $R_t$ up to a value of around $R_t \approx 4$ towards the latter part of May 2009. Further investigation on this revealed that this was mainly due to a cluster of detected cases in Dunoon. We can also observe that estimates of $R_t$ nearer the beginning of the period were less stable and the confidence intervals were much wider as there were much less case reports compared to the latter part of the epidemic curve.

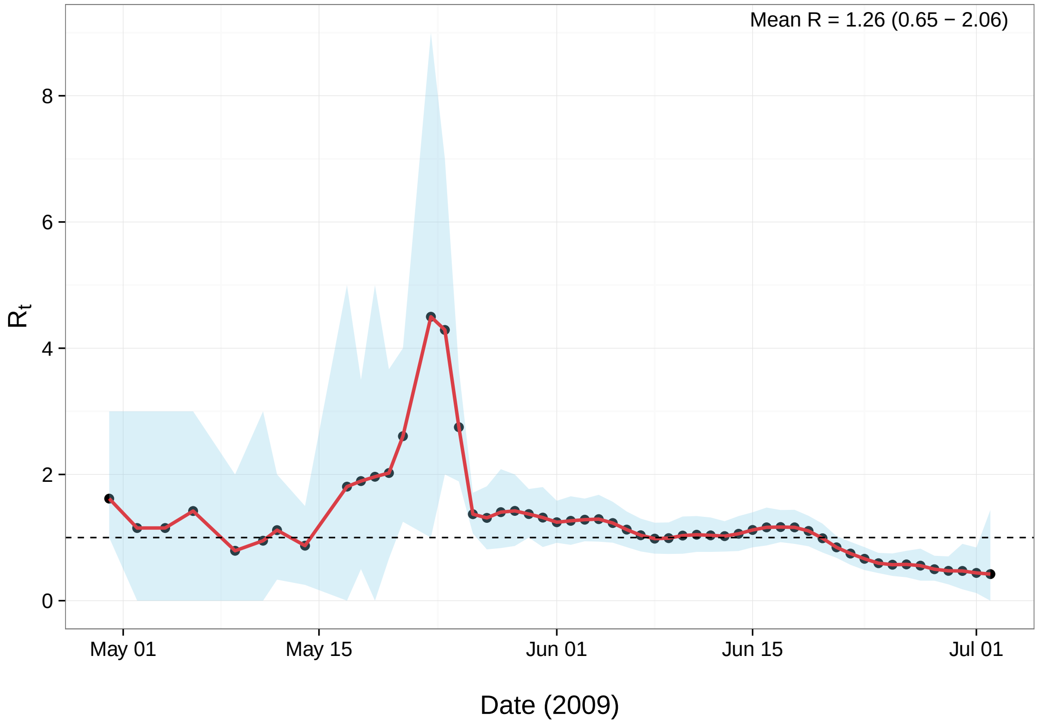
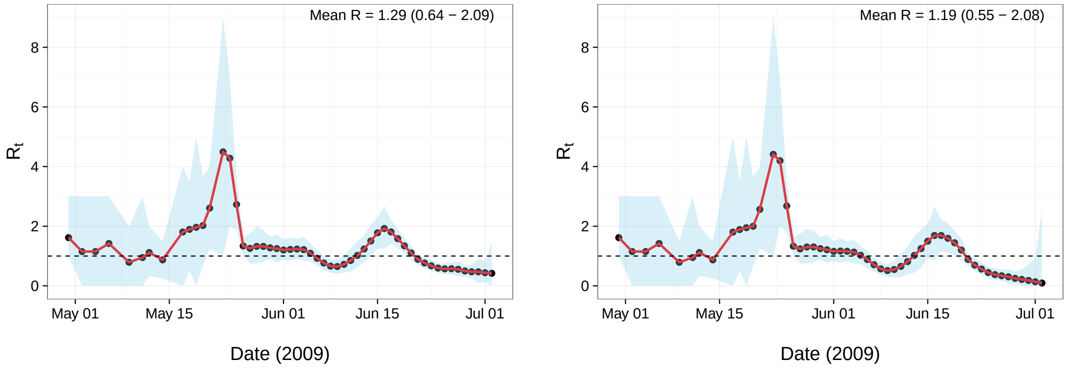


**Figure 5.12:** Time-dependent estimates of the reproductive number for pandemic influenza in Scotland found using the W&T method on the augmented dataset. The mean $R$ was found by averaging estimates from 30th April to 2nd July 2009. Confidence intervals were derived using 10,000 simulations. Note that $R_t$ values derived at times when no indigenous cases were reported were removed as $R_t$ is estimated as zero under those circumstances.

The plots on Figure 5.13 give the estimates of $R_t$ when using the imputed and original datasets instead of the augmented dataset. Both plots show a clear decrease in the estimates of $R_t$ in June, which is due to not correcting for the under-reported phase. Moreover, when using the original dataset, the average $R$ over the period dropped to $R = 1.19$ (95% CI: 0.55–2.08) and the estimates of $R_t$ at the end of the period were lower due to having less cases compared with the other two datasets where data imputations were made.



**(a)** Imputed Dataset

**(b)** Original Dataset

**Figure 5.13:** Time-dependent estimates of the reproductive number using the imputed and original datasets.

## Predictive Distribution

With knowledge of the values of $R_t$ along with the assumed serial interval distribution, the incidence at time $t$, $\hat{I}_t$ can be estimated. To do this, we start by setting the predicted incidence on the first day equal to the actual incidence on that day – in our epidemic ($\hat{I}_1 = I_1 = 1$). Then the values for $\hat{I}_{t;t>1}$, are estimated by multiplying the values of $R_t$ by the probabilities of transmission in the serial interval distribution. Hence in a scenario with one case on day one with $R_{t=1} = 1$ and using the probabilities from Figure 5.8, we would expect the first case to produce around 0.13 new cases on day two. Since we assume no secondary transmission on the same day, we would obtain $\hat{I}_2 \approx 0.13$. To find $\hat{I}_3$, we would also need the value of $R_{t=2}$ in conjunction with the values for $\hat{I}_{t;t<3}$. Similarly, values of $\hat{I}_{t;t>3}$ can be found by extending the summation to account for more values of $R_t$ and $\hat{I}_t$.

Figure 5.14 shows that predicted incidence using estimates of $R_t$ from the W&T method fitted well to the epidemic curve data for all three datasets. However, they missed the epidemic peak by a margin – more specifically, the timing of the epidemic peak was correct but the predicted incidence at this point was lower than the actual incidence. One way of determining the model fit is to use the root mean square error (RMSE) which is calculated as

$$\sqrt{\frac{1}{N_t}\sum_{t}(\hat{I}_t - I_t)^2} \tag{5.7}$$

The model produced RMSE = 5.84, 6.58 and 5.08 for the augmented, imputed and original datasets respectively. The main sources of error appear to come from two sources – clusters of cases detected around May and June 2009 and also the under-reported phase. This perhaps, gives some indication that the data imputations have still not fully accounted for the under-reported phase.

## 5.4 Spatial Model

The W&T method posits that all infection networks are equally likely and the likelihood of infection depends solely on the serial interval which is unlikely in practice. Heterogeneities in transmission can exist for any number of reasons; for example, it is well known that contact patterns vary greatly between individuals of different ages [248]. In Scotland, postcode data on cases are routinely collected allowing us to attempt to provide an extension to the W&T method by including a measure of the distance between the residence locations of cases as well as their serial intervals. The principal assumption we make is that cases living closer to each other are more likely to come into contact with each other, and consequently more likely to infect each other.

### 5.4.1 Locations of Cases

In Section 5.1, it was mentioned that postcode data was available for 1,310 cases out of the total 1,383 for which individual level data was available. Since postcode data is required here, the augmented dataset cannot be used as we do not have

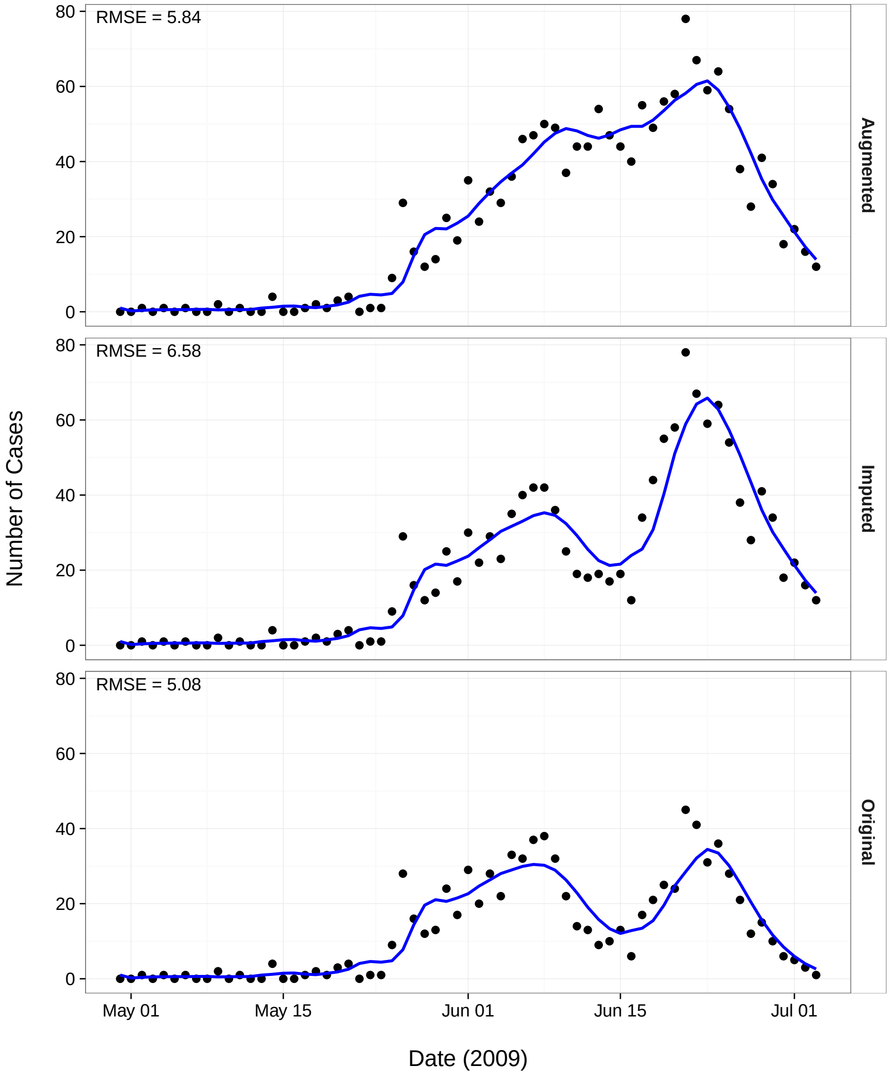


**Figure 5.14:** Predictive distribution of cases for all three datasets based on the predictions of $R_t$ along with the serial interval distribution. RMSE is the root mean square error between the predictive incidence and the actual incidence. Note that imported cases are included in counts of cases but the method cannot predict the arrival of imported cases.

location data for the extra cases added in for the under-reported phase. Thus, the imputed dataset (excluding those 73 cases without a postcode) has to be used instead. We note that for a fair comparison of the spatial method we develop here against the W&T method, it is crucial that the same data is used for both methods.

The postcodes of cases were matched to their Scottish datazone (DZ) [328], which in turn allowed us to obtain easting and northing coordinates related to those DZ centroids. A minority of cases only gave partial postcode information ($n = 11$), and so their locations were matched less precisely. The vast majority of reported cases were from the Glasgow area as illustrated in Figure 5.15, but cases sporadically appeared in the other major Scottish cities such as around Edinburgh, Dundee and Aberdeen. This is to be expected as these locations encompass the cities in Scotland with highest population densities. Imported cases showed up in a larger variety of areas compared with indigenous cases, and a small number of cases were quite isolated from the main clusters. Note that one case with a postcode from England is not shown on the map.

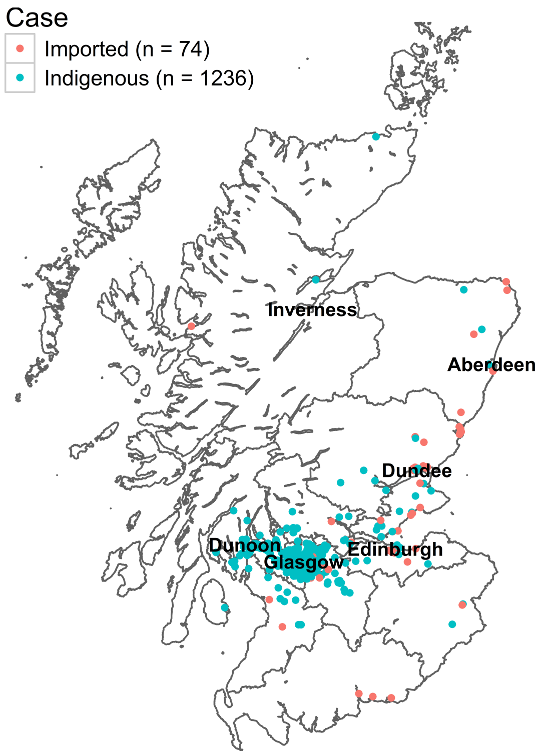


**Figure 5.15:** Locations of early phase 2009 pH1N1 cases in Scotland. $n$ in the legend refers to the numbers of cases with postcode data. Note that only cases in the original/imputed datasets could be used as additional cases in the augmented dataset do not have postcodes. Also note that one case living in England is not shown on the map.

### 5.4.2 Distance Function

The distance between cases was calculated using their easting and northing coordinates. Specifically, the Euclidean metric was used and for a pair of cases $i$ and $j$, the distance is calculated as

$$d_{ij} = \sqrt{(E_i - E_j)^2 + (N_i - N_j)^2} \tag{5.8}$$

where $E$ and $N$ here denote the easting and northing coordinates for cases. As northing and easting coordinates are characteristically associated with measurement in metres, $d_{ij}$ can be divided by 1,000 to get results in terms of kilometres.

The distances between the pairs of cases where the infectee and their likely infector was reported is shown in Figure 5.16. As can be observed, when cases reported their likely infector, it was highly likely that their infector had the same postcode, which we assume to mean living in the same household here. We propose two assertions which should be important to take into consideration when including a distance measure for determining the probability of infection:

1. The probability of infection between a pair of cases decreases as the distance between their households increases.
2. The probability of infection between a pair of cases increases when those two cases live in the same household.

#### Exponential Decay

We assume that the probability of transmission decays exponentially with increasing distance according to the function

$$\exp(-\lambda\, d_{ij}) \tag{5.9}$$

where the parameter $\lambda$ allows control over the rate of exponential decay. The effect of using different values of $\lambda$ on the rate of decay is illustrated in Figure 5.17.

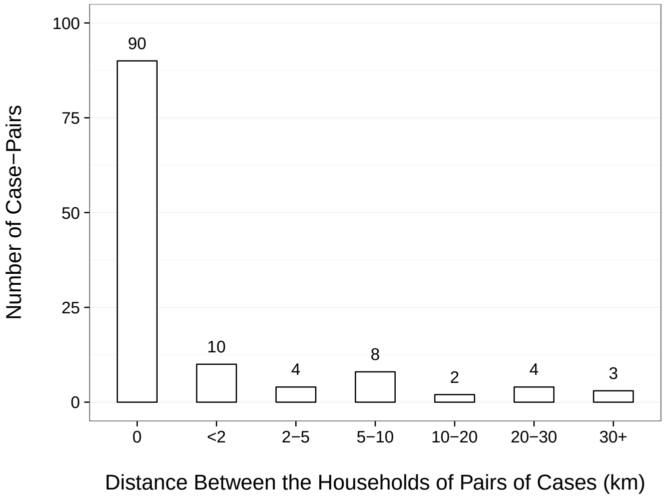


**Figure 5.16:** Distance between pairs of cases where the infectee and their likely infector was reported. We assume that pairs of cases with a distance of 0km reside in the same household.

Clearly, the spatial weight falls to zero extremely quickly as distance increases when using values such as $\lambda = 1$ or 0.5, while using lower values such as $\lambda = 0.15$ or 0.1 still give a reasonable weight to a pair of cases when they are separated by moderate distances such as 15–20km (moderate distances with respect to the geography of Scotland).

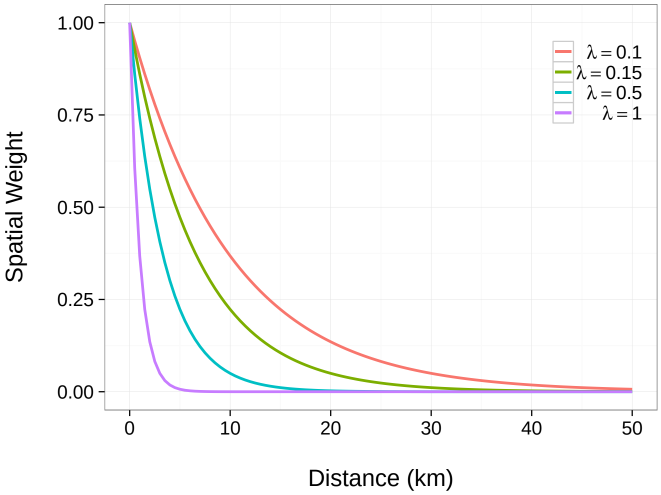


**Figure 5.17:** Rates of exponential decay for different values of $\lambda$ in the function $\exp(-\lambda\, d_{ij})$. Lower values of $\lambda$ give more weight to pairs of cases separated by greater distances.

In addition to this, we can include a parameter to increase the weight for pairs of cases living in the same household, defined as two cases with the same postcode. If this parameter is denoted as $\lambda_0$, then the spatial weight given to a pair of cases with distance $d_{ij}$ is

$$w(d_{ij}) = \lambda_0 \, \mathbb{I}_0(d) + \exp(-\lambda \, d_{ij}) \tag{5.10}$$

where $\mathbb{I}_0(d)$ is an indicator function for a pair of cases living in the same household, i.e.

$$\mathbb{I}_0(d) = \begin{cases} 1 & \text{if } d_{ij} = 0\text{km} \\ 0 & \text{if } d_{ij} > 0\text{km} \end{cases}$$

The effect that $\lambda_0$ has is dependent on the rate of exponential decay $\lambda$. If $\lambda$ is set at a higher value, then then rate of exponential decay is faster as shown in Figure 5.17 and consequently, $\lambda_0$ will cause a larger increase in spatial weight. For example, if we consider a pair of cases separated by 1km and have $\lambda = 0.1$ and $\lambda_0 = 1$, then we get $w(1) = 0 + e^{-0.1} = 0.905$ while a pair of cases living in the same household would have $w(0) = 1 + e^0 = 2$. However, if $\lambda$ is increased to a value of 0.5, then we would get $w(1) = 0 + e^{-0.5} = 0.607$, and so in this scenario the difference between $w(0)$ and $w(1)$ is larger and thus, the impact of $\lambda_0$ would be greater with a larger $\lambda$. Furthermore, if $\lambda_0 = 1$, then this will at least double the weight given to a pair of cases from the same household compared with a pair of cases with non-zero distance as we have $\lim_{d_{ij} \to 0} e^{-\lambda d_{ij}} = 1$.

### Combining Spatial Weights with Serial Intervals

To obtain relative probabilities that case $i$ was infected by case $j$ as opposed to other cases $k$, based on both the spatial weight and the serial intervals, Equation (5.5) can be multiplied by the relative spatial weight.

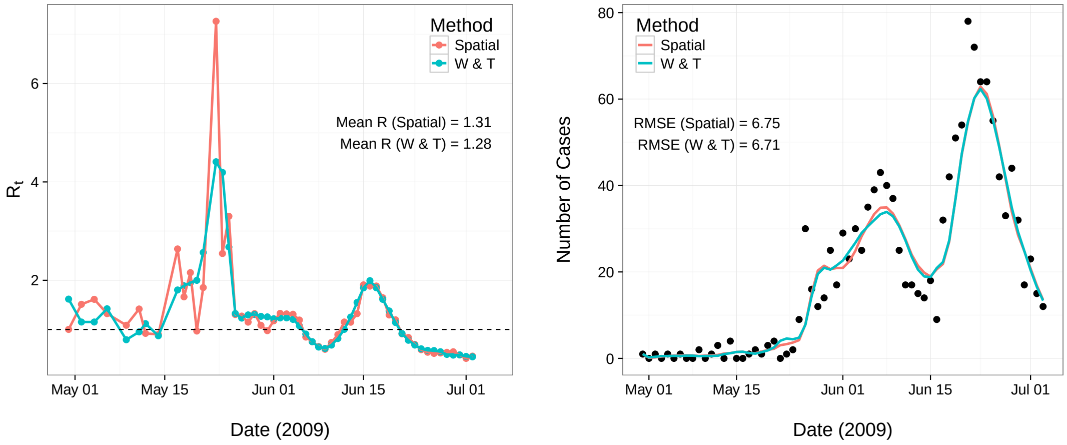
$$p_{ij} = \frac{w(t_i - t_j|\theta) \, w(d_{ij}|\lambda_0, \lambda)}{\sum_{i \neq k} w(t_i - t_k|\theta) \, w(d_{ik}|\lambda_0, \lambda)} \tag{5.11}$$

Since, $w(t_i - t_j|\theta) = 0$ when $t_i - t_j \leq 0$ (negative serial intervals) for the discretised Gamma distribution (see Figure 5.8), including the spatial weight does not have any effect on $p_{ij}$ if case $i$ has date of symptom onset before or on the same day as case $j$ (probability that case $i$ was infected by case $j$ is zero if case $i$ exhibited symptoms before or on the same day as case $j$). In other words, the spatial weight only adjusts the probability of infection for possible transmission networks. Also note that if $\lambda_0$ and $\lambda$ are set to zero, then Equation (5.11) reduces back to the W&T method as given in Equation (5.5).

Estimates of $R_t$ for the imputed dataset found using Equation (5.11) and parameters initially chosen as $\lambda_0 = 1$ and $\lambda = 0.15$ are shown in Figure 5.18a. Compared with the estimates derived from the W&T method, the spatial method predicted a much higher peak in $R_t$ in the middle of May 2009 where $R_t \approx 7$. On the whole, the estimates of $R_t$ differed at the start of the period but were very similar in the latter half of the period. The average $R$ calculated over the period was marginally higher for the spatial method (spatial: $R = 1.31$; W&T: $R = 1.28$ – note that this is slightly different from the estimate found previously in Section 5.3.2 on the imputed dataset as this time we excluded cases without postcodes). Inspection of Figure 5.18b shows that the spatial method seems to be more reactive to the peaks and troughs in the incident number of cases. However, the average predictive performance of both methods over the whole period was very similar when measured using the RMSE (spatial: RMSE = 6.75; W&T: RMSE = 6.71).

### Improving Model Fit

We initially chose parameters for the spatial model as $\lambda_0 = 1$ and $\lambda = 0.15$. These seemed reasonable as a decent amount of weight would be given when the households of pairs of cases were separated by moderate distances and if a pair of cases was from the same household, the weight would be approximately doubled compared with a very small but non-zero distance. However, by utilising the RMSE as a measure of model fit, it is possible to find the best-fitting values of $\lambda_0$ and $\lambda$ to the epidemic curve data. Best-fitting parameters were found by holding the parameters for the serial interval distribution constant and finding the parameters



**(a)** Estimates of $R_t$

**(b)** Predictive Distribution

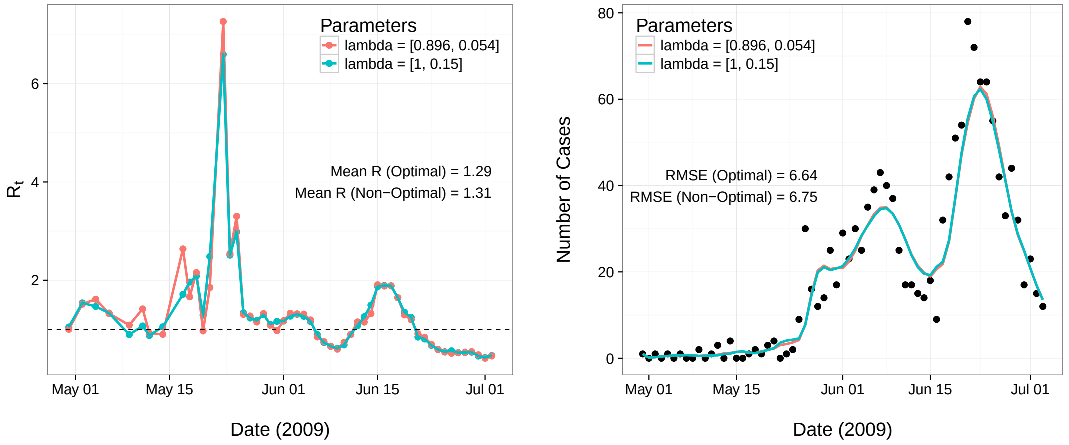
**Figure 5.18:** Estimates of $R_t$ and predictions of incidence for the imputed dataset from the spatial method compared against those from the W&T method. For the spatial method, parameters were chosen as $\lambda_0 = 1$ and $\lambda = 0.15$.

of $\lambda_0$ and $\lambda$ which minimised the RMSE. This was achieved via a bounded optimisation procedure [50] with the conditions $\lambda_0 \geq 0$ and $\lambda > 0$.

The best-fitting parameters suggest that $\lambda_0 \approx 0.9$, meaning that there should be a more modest increase in the spatial weight for pairs of cases with the same postcode, compared with what we had assigned as $\lambda_0$ before ($\lambda_0 = 1$). Furthermore, $\lambda \approx 0.05$ was found for the rate of exponential decay for distance, suggesting that the parameters that were previously chosen did not give enough weight to cases separated by larger distances. RMSE was lower using the best-fitting parameters (RMSE = 6.64) and at the maximum $R_t$ was $R_t \approx 6.5$ (rather than just above 7 with the initially chosen parameters); the average $R$ over the period was slightly lower at 1.29 (Figure 5.19).

## 5.5 Epidemic Simulations

To further test the feasibility of including a spatial element in estimating $R_t$, we created two different simulation models which attempt to simulate influenza-like epidemics in Scotland. The first model is simpler and is intended to create more favourable conditions for estimating $R_t$ using our spatial method while the second



**(a)** Estimates of $R_t$ **(b)** Predictive Distribution

**Figure 5.19:** Estimates of $R_t$ and predictions of incidence for the imputed dataset from the spatial method with best-fitting parameters compared against choosing $\lambda_0 = 1$ and $\lambda = 0.15$. Best-fit was found by minimising the RMSE.

model is more complicated and is supposed to simulate a more realistic scenario for Scotland.

### 5.5.1 Simulation Models

Both simulation models are largely similar and only differ on one key aspect which we will explain in more detail later. Simulations in both models follow the basic structure shown in Figure 5.20 – simulations are initiated with a specified number of cases and ends when a specified end day $T$ is reached. After the first day, further importations arrive each day with rate $r_{\text{imp}}$ which is constant through time. The main purpose of allowing these imported cases is to ensure that the epidemics do not die out too quickly as we want to simulate an epidemic which lasts at least around one to two months. This is particularly an issue in the first days of the epidemic simulations when there are likely to be few cases to generate new infections. Imported cases appear in a random DZ with probabilities based on the population size of DZs, with higher probabilities given to DZs with larger populations (see Figure 5.21 for an illustration of the population of DZs in Scotland). The specific location for imported cases is always set as the DZ centroid.

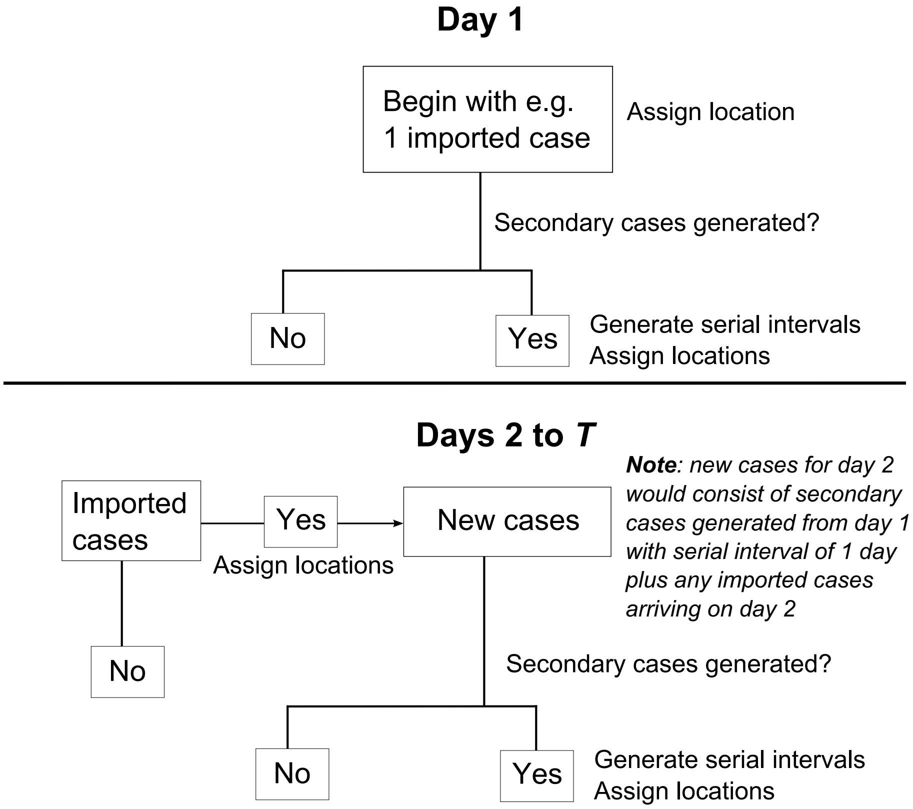

Day 1
Begin with e.g. 1 imported case
Assign location
Secondary cases generated?
No
Yes
Generate serial intervals
Assign locations
Days 2 to T
Note: new cases for day 2 would consist of secondary cases generated from day 1 with serial interval of 1 day plus any imported cases arriving on day 2
Imported cases
Yes
Assign locations
New cases
No
Secondary cases generated?
No
Yes
Generate serial intervals
Assign locations

**Figure 5.20:** Basic outline of simulation procedure. All epidemics begin from a set number of imported cases. After the first day, secondary cases are always generated in the same fashion.

Each primary case infects a random number of other individuals according to a Poisson distribution with mean $R$. Those secondary cases appear after an interval generated from the serial interval distribution. Moreover, secondary cases can be from the same household as the primary case with probability $p_{\mathrm{SH}}$. In our data, 183 cases shared a postcode with at least one other case and if we assume that around half of these cases were infected by a case with the same postcode, then an initial estimate could be $p_{\mathrm{SH}} \approx 0.07$. If a secondary case is from the same household then the eastings and northings for that case can simply be taken from the primary case.

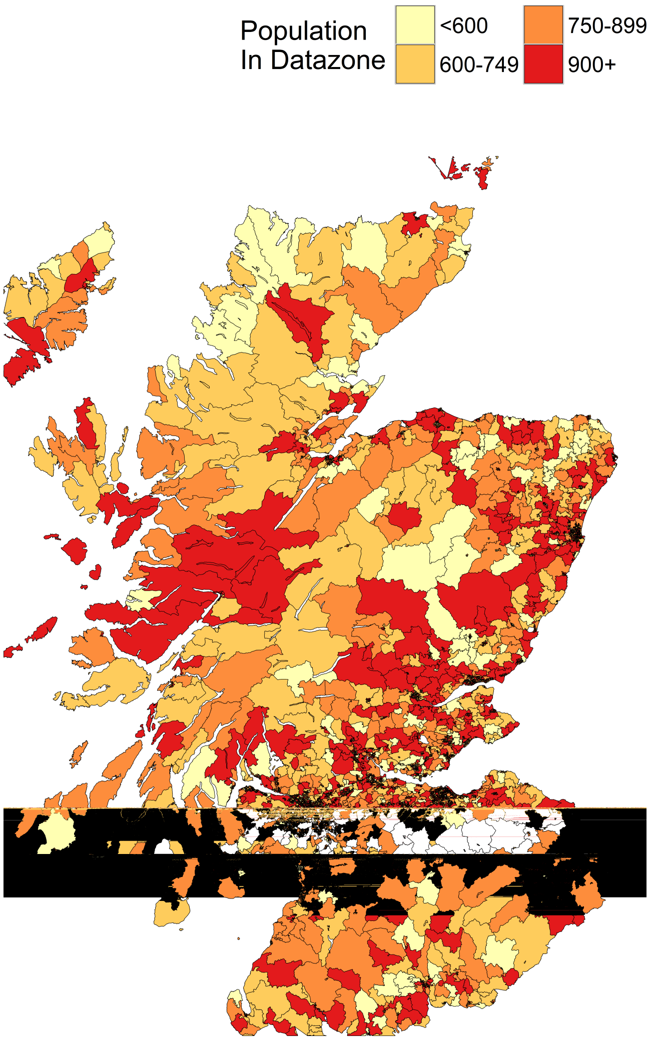


**Figure 5.21:** Population sizes by datazones in Scotland. Note that some datazones on the northern isles are not shown.

### Locations of Secondary Cases

For secondary cases that are not from the same household, further processes are required as a location must be determined for them. A distance $d_{ij}$ is generated from a negative exponential distribution with parameter $\lambda$. Then a circle is generated with radius $d_{ij}$ and centre as the location of the primary case. At this point, the two simulation models differ in how they select the location for the secondary case.

In the simpler model, an angle between 1 and 360 degrees (with 360 degrees indicating north) is selected randomly and the secondary case arises on the line of the circle with the selected angle. This procedure can create unrealistic scenarios as it allows cases to be generated in locations outside of Scotland and even in the sea. However as mentioned earlier, this model is primarily used to test how the spatial method performs under what we believe, should be favourable conditions. Another advantage of the simpler model is that it is not so computationally expensive and so it is feasible to run more simulations with this model.

In the more complicated model, the secondary cases can arise only on any DZ which coincides with the outside line of the circle. Specifically, points are generated on the outside line of the circle (for example, 25 points), and we find which DZs those points lie on and keep the set of unique DZs as possible locations for the secondary case. This process is illustrated in Figure 5.22. In this illustration, a distance of 5km was generated and the secondary case can only appear in one of four candidate DZs. The DZ for the secondary case is chosen randomly with probabilities based on the population sizes of the possible unique DZs for the secondary case.

In the rare circumstance where the points on the line of the circle do not many any DZs, circles must be regenerated until a DZ can be chosen. Once the DZ is chosen, a random set of easting and northing coordinates from within the boundaries of that DZ is selected for the exact location of the secondary case. Selecting locations for secondary cases in this fashion prevents many implausible locations for secondary cases (for instance, cases appearing in the sea or outside of Scotland). Details of the whole simulation procedure for the more complicated model in algorithmic form are given in Appendix 5.A.

### 5.5.2 Simulation Testing

It is first necessary to check that a simulation model is working properly and to do this, we can check that the models are producing the expected epidemic patterns based on chosen parameters. We simulated different epidemics using various parameter values for both, the basic and more complicated simulation models. The parameter values we used were $\lambda = 0.15, 0.35$ (reasonable/low weight given to

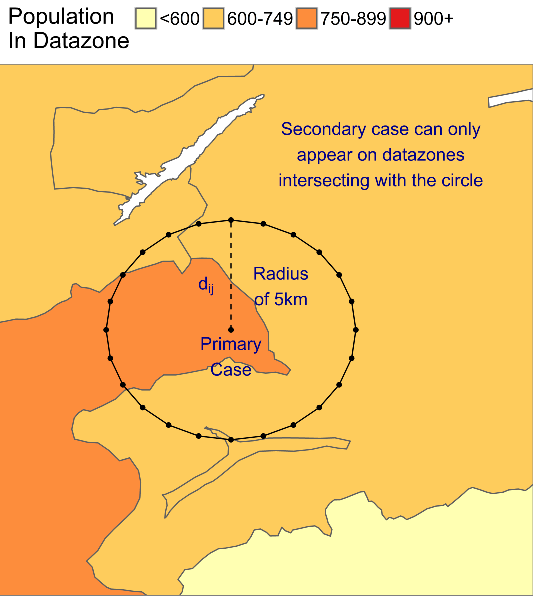


**Figure 5.22:** Illustration of how locations of secondary cases were selected for the more complicated simulation model. The point in the centre of the circle gives the location of the primary case which comes from a randomly sampled point within the boundaries of a DZ, and a distance of 5km was generated for the secondary case in this example. The possible DZs for the secondary case lie on the outside line of the circle (approximated by 25 points as shown). In this example, four DZs are possible, and the most likely DZ is the same as the primary case's DZ due to it having a higher population density compared with the other three candidate DZs which have similar population density to each other.

pairs of cases with moderate distances between their households), same household probabilities of 0.05, 0.5 (low/reasonable chance of same household probability – 0.5 is close to the rough estimate made earlier from the data), and importation rates of 0.5 and 2 (moderate/high importation rates).

Simulations were made using all combinations of these parameters (giving eight different combinations in total), with $R$ fixed at 1.1 and epidemics stopped at the time when 500 cases was reached – we kept the number of cases relatively small as our priority here was to observe the different patterns of epidemic spread produced using different parameter values. The patterns generated from simulations can be seen on Figure 5.23 (basic model) and Figure 5.24 (more complicated model).

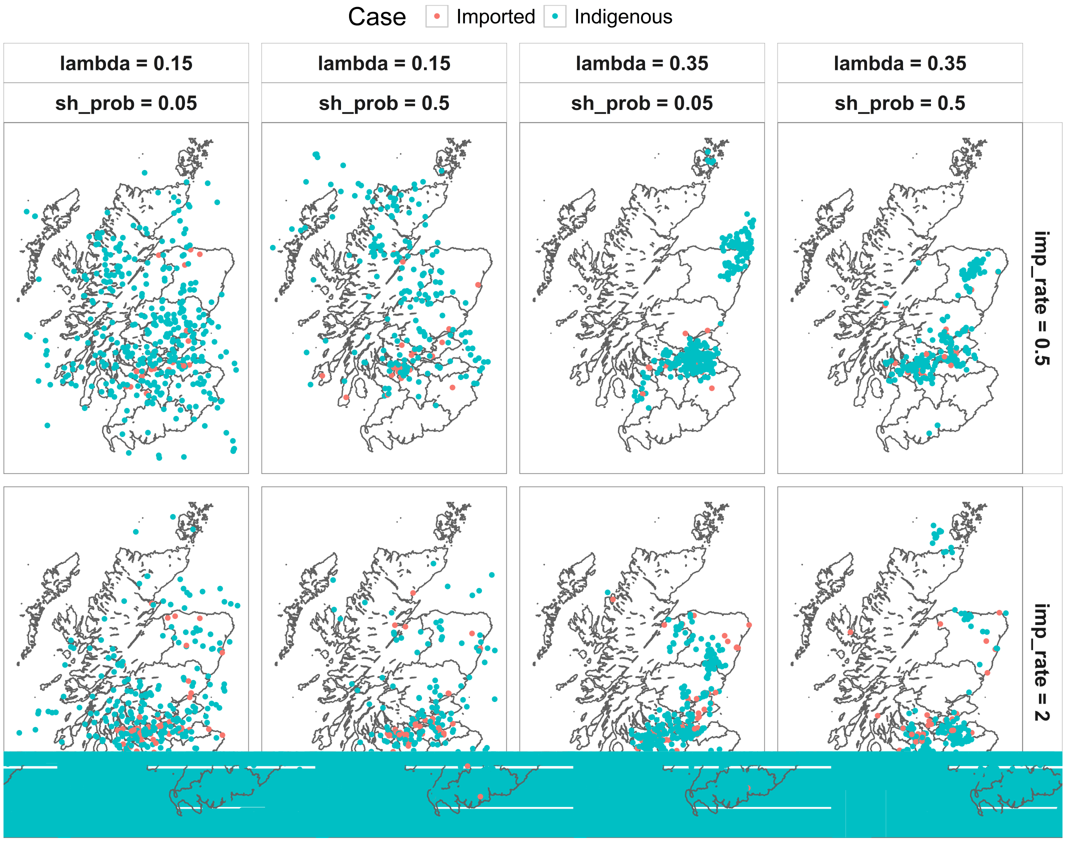


**Figure 5.23:** Typical simulations from the basic model using various values of $\lambda$, same household probability (**sh_prob**) and importation rate (**imp_rate**). All simulations started with two initial imported cases, $R$ was fixed at 1.1, and epidemics were stopped at the time when 500 cases were reached.

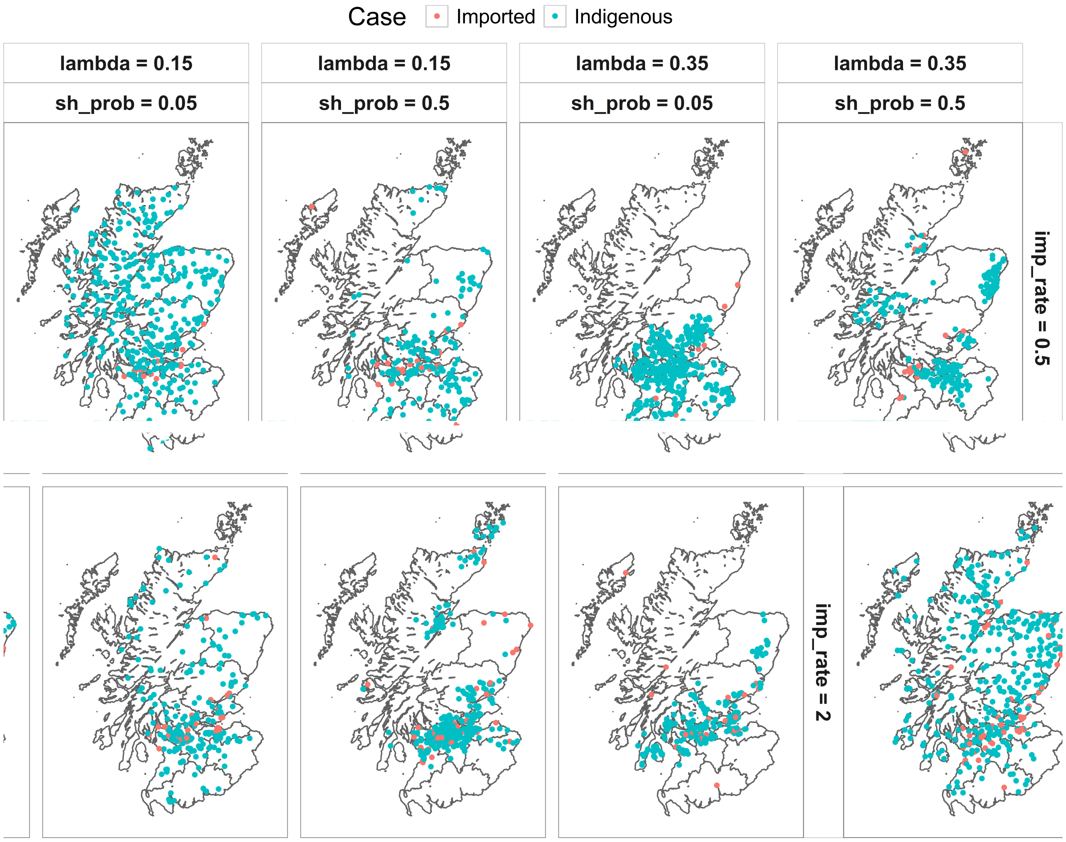


**Figure 5.24:** Typical simulations from the more complicated model using various values of $\lambda$, same household probability (**sh_prob**) and importation rate (**imp_rate**). All simulations started with two initial imported cases, $R$ was fixed at 1.1, and epidemics were stopped at the time when 500 cases were reached.

When simulating using the basic model, it is immediately clear that there was no restriction made for cases to appear within Scotland. In contrast to this, in simulations from the more complicated model, cases were all located within Scotland, and in this respect, our simulation models have worked successfully. It is also evident that setting $\lambda$ to a lower value (in this case 0.15), greatly increases the degree of spread in the locations of cases whilst cases accumulated much closer together when using $\lambda = 0.35$. As expected, using a higher probability for cases being from the same household also induces less spread on case locations. Finally, if the importation rate is higher, then we are likely to get more clusters of cases in distinct locations as a result of imported cases appearing in random DZs. As the simulations produced patterns in line with our expectations, this gives us confidence that the simulation model is working correctly.

### Finding Suitable Parameters

To further ensure that the simulation models produce results as we believe they should, we first conducted some more tests. The aim this time was to investigate if simulations would produce average values of $R$ close to what it is being set at. In the simulations, it is possible for us to track the true value of $R_t$ by examining exactly how many secondary cases are produced by each primary case at each time point $t$ and then calculating the mean $R$ for cases that appear at time $t$. A few variables are likely to have an impact on this: the value of $R$ assigned for simulations, the number of initial cases, and the rate that imported cases arrive.

Therefore, in our first test, we set $R = 1.1$, had one initial case, and had the importation rate set at 0.5 cases per day; simulations were stopped either when 40 days were over or at the time when 1,500 cases were produced if this occurred before 40 days. Note that previously we stopped simulations when 500 cases appeared as the aim there was to get an idea of the patterns produced from simulations but this time we set the stopping point at 1,500 cases as this allows us to check the simulation model when it is used to create epidemics similar to the early phase of the Scottish pandemic influenza outbreak. At the end of a single simulation, the average value of $R$ over the time period was obtained, which should be close to the input value of $R$ set in simulations for a good simulation model. This was repeated 1,000 times so that we could obtain an average for this over the 1,000 simulations. Then we repeated this whole procedure for various other parameter values – $R = 1.2$, 1.3, 1.4, 1.5, initial cases = 2, 3, 5, and importation rate = 1, 2, 3, 5.

The results from the simulation tests are shown in Figure 5.25. Clearly, using a low importation rate such as 0.5 cases per day produces poor results as the average $R$ over 1,000 simulations is far away from the value set for $R$ in simulations. This is because with a low importation rate, epidemics grow too slowly and most of the simulations will reach 40 days and stop without producing a lot of cases. The consequence of having less cases is that the average $R$ will be more variable. In fact, when $R$ is set to a reasonably low value ($R = 1.1$–$1.5$) the importation rate seems to have the most significant impact on the value of $R$ in simulations, whilst the number of initial cases has a lesser impact. On the whole, it appears that for all values of $R$ tested, using two initial cases and setting the importation rate at three cases per day produced consistently good results.

### 5.5.3 Simulation Results

A typical simulated epidemic with $R$ set at 1.2 and running for a period of just over a month is illustrated in Figure 5.26. For the simulation, we used the parameters: $\lambda = 0.3$, same household probability for secondary cases of 0.05 and importation rate set at two cases per day. The true mean $R$, calculated by recording the numbers of secondary infections produced by cases over the period was 1.163, while the mean $R$ estimated using the spatial method over the period was 1.151, and the mean $R$ estimated by the W&T method was 1.157. Note that $\lambda = 0.3$ and $\lambda_0 = 1$ were used in estimations using the spatial method.

From this measure, the W&T method was closer to the true average $R$ than the spatial method, but perhaps a better measure of performance is how well the two methods capture the changes in $R_t$ over time. One way of measuring this is to compare the true $R_t$ with the estimated $R_t$ at time points where there are cases

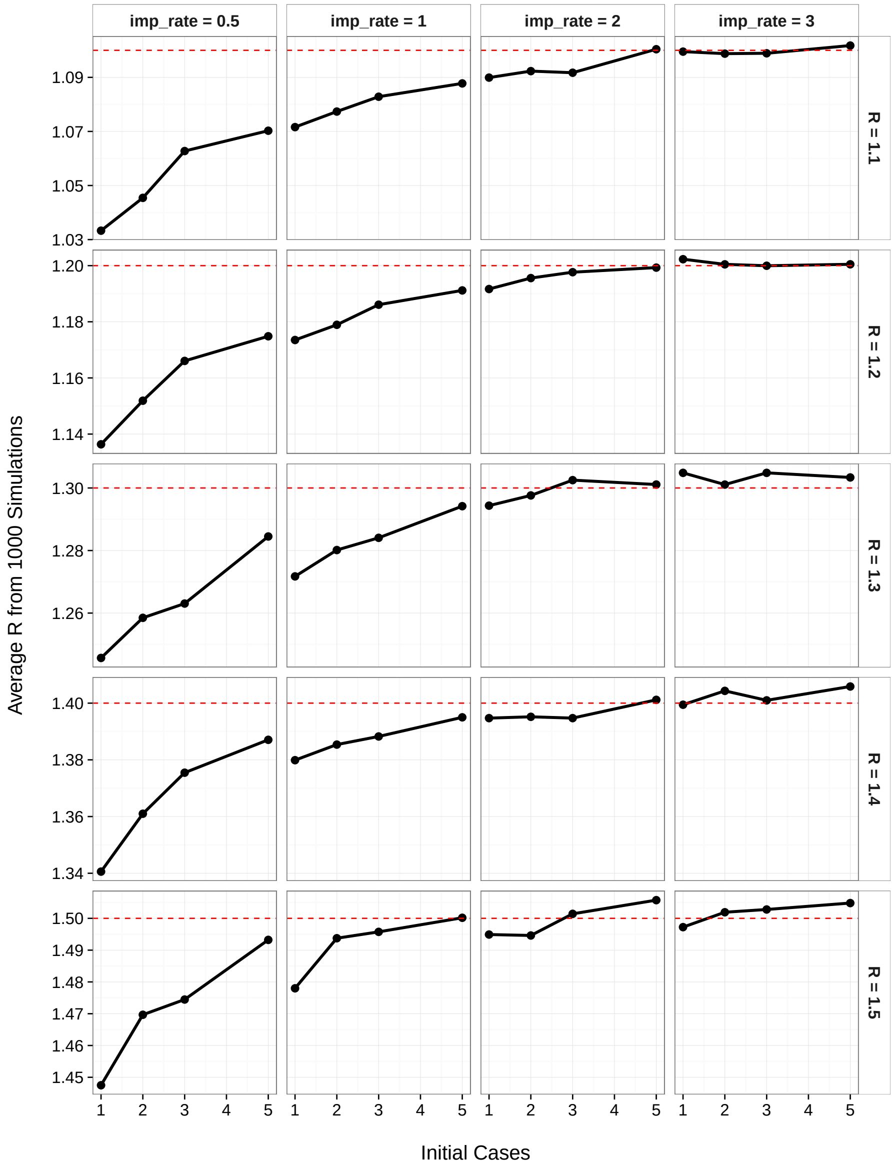


**Figure 5.25:** Results from test runs of the simulation procedure. The average value of $R$ over 1,000 simulations should be close to the input value of $R$ in simulations for a good simulation model. The results show that this is affected by the importation rate (**imp_rate**) and to a lesser extent, the number of initial cases.

(the numbers at the bottom of Figure 5.26 indicate the number of indigenous and imported cases that appeared every day in the simulation). Visual inspection of the plot shows that the spatial method captures the changes in $R_t$ much more closely than the W&T method, which produced less variable estimates of $R_t$. We summarised these differences by calculating an RMSE using the difference in values of the true $R_t$ and the values of $R_t$ estimated using the spatial and W&T methods; RMSE was much lower using the spatial method than the W&T method (0.17 compared with 0.3).

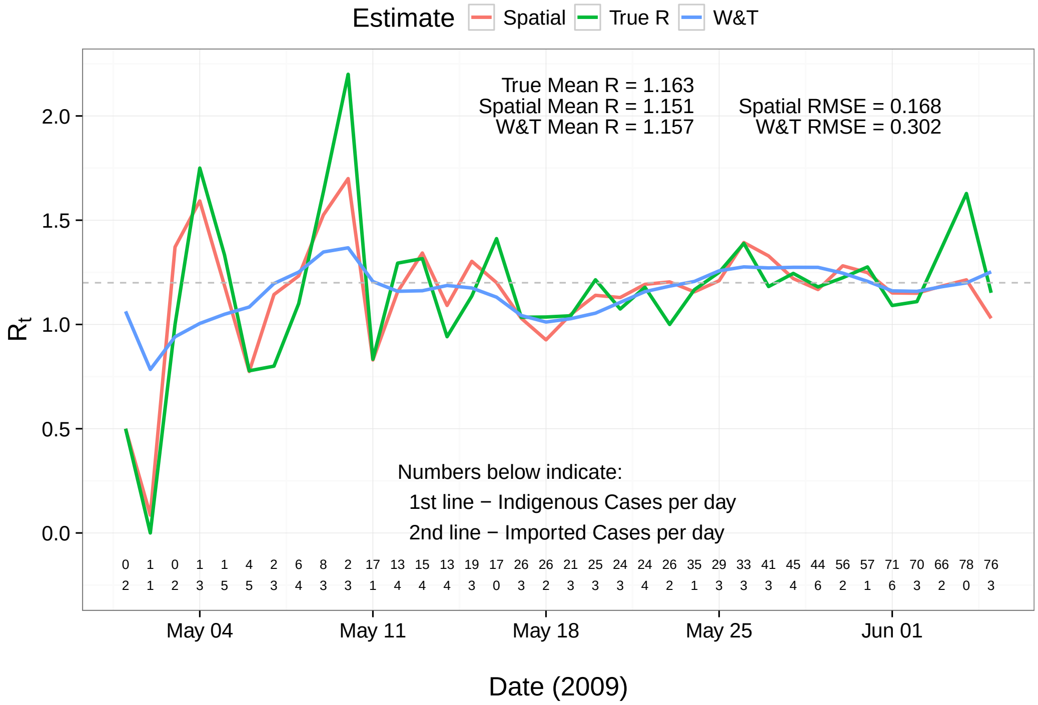


**Figure 5.26:** Comparison of the spatial and W&T methods of estimating $R_t$ for a simulated epidemic. RMSE was calculated from the difference between the values of the true $R_t$ and the values of $R_t$ estimated at every timepoint when there were cases using the spatial and W&T methods. In this simulation, $R$ was set at 1.2 (this value is marked by the grey dashed line). Estimations from the spatial method used the parameters $\lambda = 0.3$ and $\lambda_0 = 1$. The numbers at the bottom indicate the numbers of indigenous and imported cases per day.

To ascertain if these results would be consistent, we simulated 1,000 epidemics using the basic simulation model and 100 epidemics using the more complex simulation model, which is far more computationally expensive to simulate from. In all simulations, we used the same parameters as in the typical simulation we just described, with the exception of varying $R$ between 1.1, 1.2, 1.3, 1.4 and 1.5. In all simulations, estimations from the spatial method were made using $\lambda = 0.3$ and

$\lambda_0 = 1$. After each simulation, we recorded the mean $R$ from both methods, the RMSE for both methods, and the difference in the true mean $R$ over the period compared with the mean $R$ over the period estimated using the two methods which we call the "mean error". Finally, we calculated the mean over 1,000 epidemics for the basic model and the mean over 100 epidemics for the complex simulation models – the results are shown in Table 5.3.

| | **R = 1.1** | **R = 1.2** | **R = 1.3** | **R = 1.4** | **R = 1.5** |
|---|---|---|---|---|---|
| **Basic Model** | | | | | |
| *Spatial Method* | | | | | |
| Mean $R$ | 1.103 | 1.207 | 1.306 | 1.406 | 1.506 |
| Mean RMSE | 0.238 | 0.241 | 0.257 | 0.268 | 0.283 |
| Mean Error | 0.015 | 0.018 | 0.020 | 0.024 | 0.026 |
| *W&T Method* | | | | | |
| Mean $R$ | 1.101 | 1.205 | 1.303 | 1.404 | 1.504 |
| Mean RMSE | 0.285 | 0.290 | 0.309 | 0.324 | 0.341 |
| Mean Error | 0.018 | 0.020 | 0.024 | 0.026 | 0.031 |
| **Complicated Model** | | | | | |
| *Spatial Method* | | | | | |
| Mean $R$ | 1.102 | 1.210 | 1.314 | 1.420 | 1.512 |
| Mean RMSE | 0.248 | 0.242 | 0.274 | 0.276 | 0.283 |
| Mean Error | 0.016 | 0.021 | 0.022 | 0.025 | 0.029 |
| *W&T Method* | | | | | |
| Mean $R$ | 1.098 | 1.206 | 1.306 | 1.415 | 1.508 |
| Mean RMSE | 0.283 | 0.283 | 0.313 | 0.329 | 0.336 |
| Mean Error | 0.017 | 0.022 | 0.025 | 0.028 | 0.032 |

**Table 5.3:** Results from epidemic simulations to compare the spatial and W&T methods. When using the basic simulation model, 1,000 simulations were generated and 100 were generated when using the more complex simulation model. RMSE was calculated from the difference between the values of the true $R_t$ and the values of $R_t$ estimated from the two methods at every timepoint when there was at least one case. The "mean error" is the difference in true mean $R$ over the period and the mean $R$ over the period estimated using the two methods.

For all values of $R$ set in simulations, the mean $R$ over the period estimated using the two methods was slightly closer to the true value of $R$ when running simulations from the basic model compared with the more complex model but both simulation models performed well in this respect. Moreover, the spatial method always had lower mean RMSE compared with the W&T method and had lower mean error. Figure 5.27 displays the RMSE for both estimation methods for each simulation from the two models. We can observe that the RMSE for each simulation is nearly always larger from the W&T method than the spatial method. These results suggest that there is some merit in including a spatial factor in the calculation of $R_t$.

## 5.6 Discussion

During the early phase period of April to June 2009, the influenza epidemic was relatively mild in Scotland. After imputation for missing dates of symptom onset using the reporting delay distribution and correction for a period of under-reporting caused by a temporary halt in laboratory testing, $R$ was found to be around 1.36 using the exponential growth method [371] and around 1.26 using the time-dependent method (averaged over the whole period) [372]; without corrections to the data, $R$ was slightly lower as a result of the fewer cases reported. Time-dependent estimates of $R_t$ showed one major peak in transmission activity observed during the middle of May. This owed largely to a cluster of cases detected in Dunoon around that time. These cases all travelled together on the same bus.

The estimates of $R$ found here are fairly consistent with those found in surveillance reports from other countries such as early estimates from Mexico [75, 123], New Zealand [278], Peru [250] and Chile [288]. However higher estimates of $R$ being above two have also been observed in a number of studies including estimates found in Japan [266], Australia [237], the USA [382], and Mexico [37].

A notable point found in the reports from Japan and Australia is the higher rates of transmission found for minors (those aged under 20), particularly for transmission between minors to other minors. This emphasises the importance

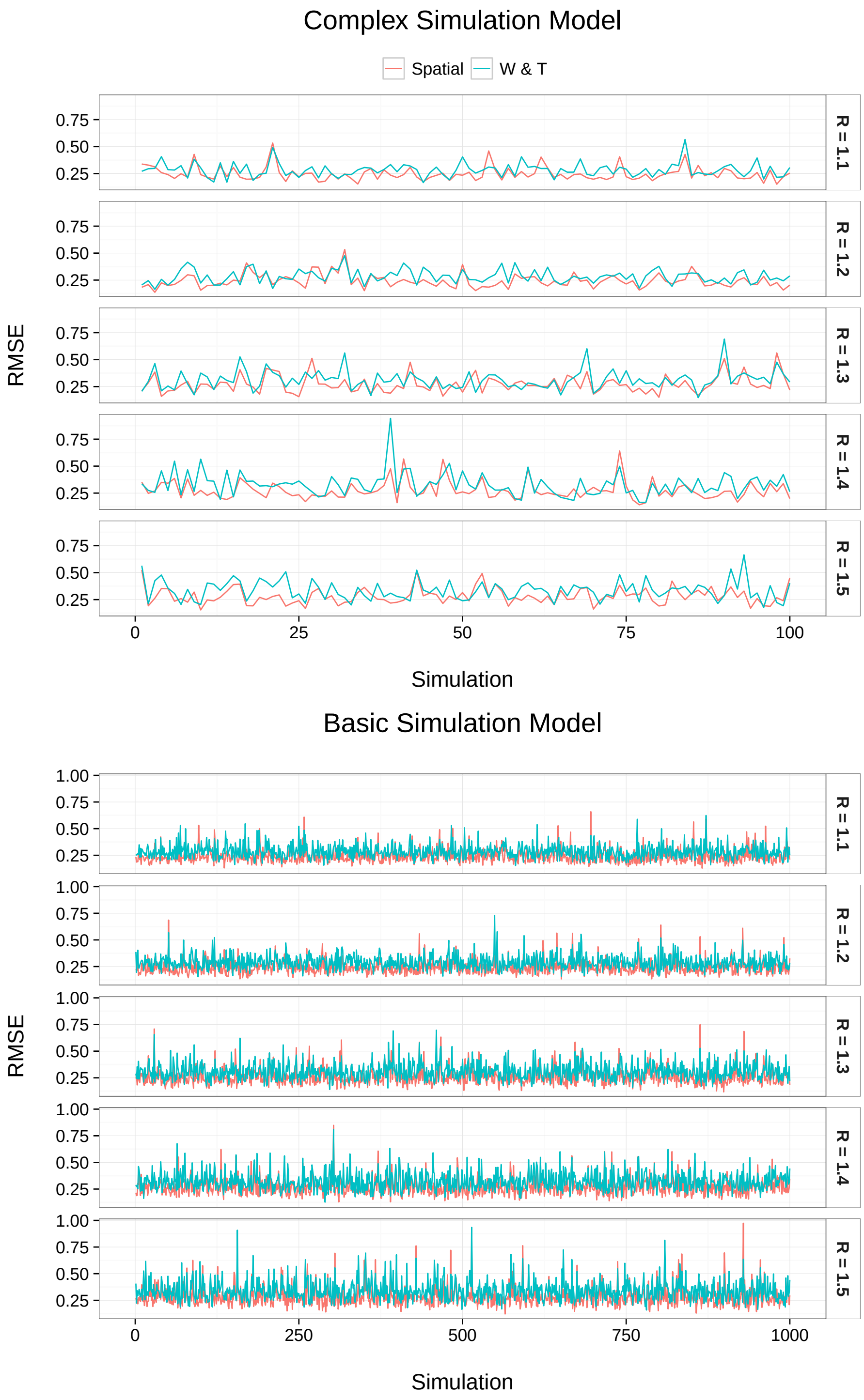


**Figure 5.27:** RMSE from the spatial and W&T methods for basic and complex simulations. 1,000 simulations were generated from the basic model and 100 simulations were generated from the more complicated model.

of social contact patterns when estimating $R$, and this underlines some of the limitations of our analysis which did not take into account these heterogeneities. Knowledge of social contact structures can aid greatly in efforts to contain the spread of disease. For example, if it is known that most of the transmission is coming from children, then the decision to close schools for a period of time may be an appropriate strategy. Additionally, more cost-effective vaccination campaigns would be possible with knowledge around what groups of individuals to target to achieve maximum reductions in transmission.

It is challenging to incorporate an accurate measure of the amount of contact that individuals of different age groups have with each other, but mixing patterns have been investigated before in European countries by Mossong et al. [248]. This offers a promising route to derive estimates of $R$ after accounting for differences in mixing patterns. However it must be noted that in those surveys, individuals were asked to record their own contact patterns in paper diaries, and so it is inevitable that there will be some level of error through for instance, imperfect memory recall.

The strengths and limitations of using the exponential growth method for estimating $R$ as we did has been discussed by Nishiura et al. [267]. Briefly, they highlight how the cut-off date can affect estimates of the exponential growth rate. This can be a problem when there are multiple peaks in case notifications; choosing to model the rate on an initially high peak in cases could potentially overestimate $R$ if that peak occurred as a result of biases in, for instance, case ascertainment. We demonstrated this issue when we used 7th June 2009 as a cut-off date for the exponential period. This cut-off date resulted in much higher estimates of $R \approx 1.7$ due to an overestimation of the growth rate.

Another limitation in the methods we used to estimate $R$ is that they perhaps, do not account appropriately for imported cases. As discussed in Roberts and Nishiura [309], $R$ can be overestimated by not dealing with importations in an adequate fashion. In our estimates of $R_t$, indigenous secondary cases have the same probability of having been infected by imported cases as by other indigenous cases (provided the time interval is the same). In reality, imported cases are likely to infect others in Scotland but perhaps at a different rate from indigenous cases. The factors that we should consider include the transmission dynamics in the country where imported cases were infected as well as an associated time-lag for importations. This happens as imported cases would spend a portion of their infectious period outside of Scotland, leaving only a fraction of time where they could cause secondary infections in Scotland.

We attempted to adjust the probabilities of infection estimated from the method of Wallinga and Teunis [372] via the inclusion of a spatial term. Our assumption was that increasing the distance between the households of a pair of cases would lower the probability of infection between this pair of cases. Specifically, we assumed that the probability decays exponentially with increasing distance. This assumption is likely to be fine for the most part, and particularly for cases living in more remote parts of Scotland where there may be less interaction between those individuals with others living far away. However, the assumption could be less likely to hold for people working in the major cities as many have to commute to work, and this creates many opportunities for transmission of infection to others during travel and at their place of occupation. In these circumstances, there is definite potential for transmissions to people that do not live close to the primary case.

An additional adjustment was that we increased the probability of infection for pairs of cases that we assumed to belong to the same household. We note that there could have been some inaccuracies with this as we did not have data specifying if cases belonged to the same household, but we inferred this information by assuming that if cases shared the same postcode, then they were from the same household. An example of where cases could share the same postcode but not belong to the same household would be for residents of different apartments within the same flat. Despite the potential inaccuracies in the way we derived the information on whether or not cases shared the same place of residence, we felt that it was important to try and account for the increased chance of disease transmission for individuals that live together.

Estimates of $R_t$ found by including the spatial term were more variable compared with the Wallinga and Teunis method; especially in April and the beginning of May where there were not so many case notifications. We also found a higher peak in $R_t$ with the spatial method due to the cluster of cases found in Dunoon who all lived relatively close to each other.

Initially estimates for the two additional parameters in the spatial model ($\lambda_0$ and $\lambda$) were chosen fairly arbitrarily but then we followed a more methodical approach to obtain improved estimates. We found the parameters by minimising the RMSE using an optimisation process, where the error was calculated by comparing the predicted daily numbers of new cases from the model against the data. We note that other scale-independent measures of predictive performance are available such as the mean absolute percentage error (MAPE) but in our case, this was not a concern since the scale is the same for the two estimation methods. Furthermore, we could also have used other methods for estimating the parameters such as Bayesian methods [129]. These methods assume that the parameters are random variables but this comes with the added complication of specifying suitable prior distributions for the parameters (although non-informative priors could be chosen to minimise their impact on estimations).

It is possible that the assumptions we made for the spatial method would not be entirely appropriate as distance between households of cases may not necessarily serve as an entirely accurate proxy for contact patterns. However the idea of adjusting estimates found from the Wallinga and Teunis method has the potential to offer better estimates of $R_t$ as they only consider generation intervals (or serial intervals). Whilst one of the strengths of their method is the relatively few assumptions they impose on transmission, this can also be a limitation.

Using our epidemic simulations, we showed that the spatial method could more closely capture the true value of $R_t$ compared with the Wallinga and Teunis method. However, we should bear in mind that in simulations, we had control over how distances were generated for secondary cases and so we did not estimate $\lambda$ and $\lambda_0$ when deriving reproductive numbers but used values based on the parameters set in simulations. Thus, this is likely to have led to more favourable results for our spatial method of estimating $R_t$. In a real influenza outbreak, we cannot be sure that the distances between the locations of primary and secondary cases would always mimic the patterns we generated.

Another consideration we make is that we could have incorporated our spatial term in other ways. One possibility is to only consider neighbouring datazones for secondary infections. This would mean that we would set the weight, $w(d_{ij}) = 1$ if another case is in the same or a neighbouring datazone and set $w(d_{ij}) = 0$ otherwise – this is similar to how a conditional autoregressive (CAR) model is implemented [33]. However, we decided not to use this as it is inherently unrealistic to impose these conditions on an influenza outbreak and the use of datazones as boundaries is also not ideal as they only cover very small areas of Scotland. One other way of including a spatial term is to give weights based on distance boundaries. For instance, we could have set three boundaries giving highest probability to those cases within 1km of a primary case, a moderate probability to cases within 1–5km and a low probability for any cases further than 5km away. This method would not exclude any cases on the basis of distance but would have added another decision in the form of choosing appropriate distance boundaries.

One of the questionable assumptions of both the spatial and Wallinga and Teunis methods is that of a constant generation interval distribution over time. Kenah et al. [202] have argued that the generation interval contracts as epidemics progress due to increased competition to infect remaining susceptibles; essentially as epidemics progress, there are a growing number of infectious individuals who would *race* to infect susceptibles. We investigated this idea with our subset of cases that reported a likely infector. While we found that cases with date of symptom onset in the latter part of the exponential growth phase had slightly shorter and less variable serial intervals compared with cases with date of symptom onset in the earlier part, our results did not indicate that there was any significant difference in the length of serial intervals. If there was a real contraction, more accurate estimates of $R$ would be achieved by using different serial interval distributions for different time periods. However, another issue with investigating this is that it would be difficult to know if any contraction is a real contraction or if the contraction is an artifact due to increases in surveillance as the epidemic progress.

All of our estimates of $R$ are indicative of a mild epidemic but as usual, the estimates are subject to a degree of uncertainty. The uncertainty stems from not only the methods used to estimate $R$ but also from the use of only laboratory-confirmed cases. In the early phase we assumed that everyone with symptoms was tested and so the 1,383 cases represent the totality of cases in the period April to June 2009. However, it is likely that certain types of individuals would be more likely to get tested, which could result in some biases.

Our attempts to include a spatial term in estimating $R_t$ could benefit from investigation into whether or not exponential decay is suitable for modelling distance. It is difficult to find suitable parameters for the rate of decay and also how much to increase the likelihood of infection by if a pair of cases is from the same household as there is not much data for which the chain of transmission is confidently known. Moreover, these parameters would also be different depending on the geography of the country being analysed. For example, we did not distinguish if cases are from high or low population density areas and it could be useful to add in this extra layer of complexity. As mentioned before, the method could also be improved by dealing better with imported cases. Furthermore, it would be interesting to explore the effects of adjusting by other important factors such as age.

To test the spatial method we developed a simulation model which could also be improved. Our simulations allowed cases to appear in any location within the bounds of datazones in Scotland. This, of course, would not happen in reality as cases should only be located in residential zones. Hence, a possible extension would be to sample from lists of postcodes once a datazone is chosen for cases; this would restrict cases to appearing in locations which have postcodes.

For surveillance purposes, it is vital that estimates of $R$ are obtained in as near real-time as possible during epidemics which is what these methods offer to an extent. However, we must keep in mind that the most accurate estimates of $R$ are probably only attainable retrospectively when more complete data is available on the population – for instance, large-scale serological data [373]. Therefore estimates made using the methods outlined in this chapter are useful for getting early estimates which allow decision-makers to react but these estimates should also be validated against other methods when and where possible.

Preventative measures can be used in conjunction with surveillance measures including those used here in order to monitor their impact. In the remaining chapters we consider interventions that can help prevent new cases from arising, focussing exclusively on vaccines. When vaccines are available against an infectious disease as a preventative measure, they offer one of the best routes to protecting the population. In the next chapter, we look at some different ways of determining how effective vaccines are as well as exploring various issues that are often encountered in studies of vaccine effectiveness.

## 5.A Appendix: Epidemic Simulation Algorithm

**Algorithm 1** Simulate influenza-like epidemic in Scotland

```
Require:
  Spatial polygon data for Scottish datazones (DZ)
  Population densities for DZs
Inputs:
  Reproductive number $R$
  Serial interval distribution
  Number of initial cases $n_{\text{init}}$
  Exponential decay rate $\lambda$
  Length of epidemic $T$
  Importation rate $r_{\text{imp}}$
  Probability of secondary case being from same household $p_{\text{SH}}$
Initialise with:
  Initialise with $n_{\text{init}}$ indigenous cases on day $t = 1$
  Sample DZs randomly with probabilities from population densities
  Obtain eastings and northings for DZs
```

**Following Initialisation:**

```
while day $t < T$ do
  if day $t \geq 2$ then
    Generate number of imported cases on day $t$ from Poisson with rate $r_{\text{imp}}$
    if Imported cases $\neq 0$ then
      Sample DZs randomly with probabilities from population densities
      Obtain eastings and northings for DZs
    end if
  end if
  for all New cases on day $t$ do
    Generate number of secondary cases from Poisson with mean $R$
  end for
  if Secondary cases $\neq 0$ then
    for all Secondary cases arising from primary cases on day $t$ do
      Generate serial intervals from serial interval distribution
      Randomly assign cases to be from same household with probability $p_{\text{SH}}$
      for all Secondary cases from same household do
        Assign eastings and northings to secondary cases from primary cases
      end for
      for all Secondary cases not from same household do
        Generate random distance $d_{ij}$ from negative exponential distribution
          with decay parameter $\lambda$
        Generate circle with radius $d_{ij}$ with centre as location of primary case
          {Secondary cases can only appear in DZs on the line of a circle}
        Sample from possible DZs randomly with probabilities from population densities
        Sample random easting and northing coordinates for secondary case
          from selected DZ
      end for
    end for
  end if
  Move to day $t + 1$
end while
return Dataframe containing date onsets, eastings and northings
```

# Chapter 6

# Studies of Influenza Vaccine Effectiveness

In this chapter, we move away from surveillance of disease outbreaks to the subject of control measures. Public health interventions can be crucial for reducing the numbers of individuals that experience severe outcomes from a disease, but the success of these interventions cannot be automatically assumed and hence, their impact must be assessed. The previous chapters focussed on the novel pandemic influenza outbreak in 2009. During that outbreak, public health bodies quickly recognised the important role that vaccines may have in reducing transmission of infection, and hurriedly started production on monovalent vaccines after outbreaks in the USA and Europe. Following this, the UK's national immunisation strategy group introduced two pandemic influenza vaccines (Pandemrix and Celvapan) to be used as a preventative measure.

In Scotland, the Chief Medical Officer worked alongside the Scottish government to establish a vaccination programme targeted initially at pregnant women, healthcare workers and vulnerable groups of individuals such as those with underlying health problems. The choice of groups of individuals to target for vaccination was supported by investigating mortality due to pandemic influenza [91]. Uptake of the vaccine began during the second wave of infection in late October 2009 and soon after, the effectiveness of the vaccines were assessed using various study designs and methodologies [151, 336].

It is not uncommon for the pandemic strains of influenza to garner the most attention among the public and media as they have the most potential to cause devastating consequences. As a result of this, the ongoing battle against seasonal influenza may sometimes be overlooked but it should not be as it is an ever-present disease which causes many deaths year-on-year, particularly among the frail and elderly [357]. With this in mind, our aim is to look at different ways of assessing the effectiveness of the seasonal influenza vaccine. Specifically, we have data on cohorts of patients taken from a sample of general practitioner's (GP) practices in Scotland for two post-pandemic seasons. This allows for estimation of influenza vaccine effectiveness (VE) by measuring it against a variety of clinical outcomes using a range of different methods and study designs. However, evaluating how effective influenza vaccines are can be anything but straightforward and there are various issues which must be considered when carrying out VE studies.

For this reason, we defer the analysis of influenza VE in Scotland until the following chapter, allowing this chapter to serve as an important prelude to our study. It does this by firstly discussing the typical types of study designs used for VE studies. We then give consideration to some of the issues which arise specifically when dealing with influenza VE studies. Finally we explore some of the methods which are used to deal with issues in VE studies – for instance, methods of detecting as well as reducing bias and confounding (both of these terms will be discussed in more detail later) which can often be present when using observational study designs. Although the main focus is on influenza VE, some of the methods are generic on VE and may not be suitable for looking at influenza VE but the majority are applicable to studies of influenza VE.

## 6.1 Introduction to Influenza Vaccination

For many diseases, vaccinations are highly effective and we only have to reminisce on the success of the smallpox vaccination campaign which led to the total eradication of the disease in 1977 to see how effective they can be [24, 42]. As well as this, the effect of vaccinations against some diseases are not expected to change much from year to year; a few examples include poliomyelitis [81], measles

[348] and hepatitis B [228]. However, producing an effective influenza vaccine has proved more troublesome. The main reason for this is that influenza is a continually evolving virus [342] and as the circulating influenza strains vary each year and by geographic regions, the vaccine must be created with concern to which strains are in dominant circulation. A consequence of this is that influenza VE calculated for one year will almost never be a good indicator of how effective the vaccine will be in following years. In spite of these issues, influenza vaccination remains the best way of preventing and controlling influenza and is widely believed to save many lives throughout the world [4, 292, 368]. This is especially true when there are severe influenza epidemics and the vaccine produced is a good match.

For these reasons, vaccination campaigns are widespread but as vaccinating the entire population is prohibitively expensive, a trade-off between cost and effectiveness in protecting individuals against adverse outcomes due to influenza has to be achieved. Hence, vaccination is usually targeted at individuals in risk groups such as the elderly, young children, healthcare workers (as they are more likely to have contact with vulnerable individuals) and those that have underlying health issues such as chronic disease. There are various reasons as to why certain groups of individuals may be targeted for vaccination – some are encouraged to get vaccinated as they are at an increased risk of developing serious complications following infection and others may be more likely to transmit influenza to other individuals (for instance, children who come into contact with many other children at schools).

To monitor how successful a vaccine has been in preventing disease, it is vital to make estimates of influenza VE. Moreover, estimates of influenza VE need to be constantly made as past estimations of VE cannot be guaranteed to be a good indicator for the VE in subsequent seasons. The accuracy of VE estimates are dependant upon many factors including the choice of clinical outcomes used to measure it against and the study design which may be influenced by the availability of data. The choice of statistical methodology used to estimate VE is also likely to depend on the study design used, which makes the choice of study design one of the more crucial aspects of conducting a study on influenza VE.

## 6.2 Study Designs

The gold standard for study designs to test the effectiveness of a treatment are randomised controlled trials (RCTs) [6] as they are supposed to allocate treatments to groups in a truly randomised manner. The main premise of an RCT is that the random allocation should ensure there are no systematic differences between the group of individuals used as controls and the group of individuals given the treatment. These systematic differences refer to both, known and unknown factors that affect the outcome of interest. However, this type of design is considered unethical for studies of influenza VE since individuals that need the vaccine for protection should not be denied it. This is particularly true when looking at elderly populations as they are most at risk of hospitalisation and mortality following influenza infection [263]. Thus, alternative study designs have to be considered. Observational studies offer feasible solutions to this problem as they can be used to assess the effects of healthcare interventions without affecting the care provided or the patients [34].

Cohort studies and various different versions of case-control studies are commonly used in studies of influenza VE. Another simple method which makes use of aggregated data on numbers of vaccinated and unvaccinated individuals to obtain rapid estimates of VE is the *screening method* [109]. Selecting which study design to use requires careful consideration of what data and resources are available. In addition to these, attention should be paid to the methodological constraints of a particular design as there are advantages and disadvantages in using each design [362]. While study designs differ, there are issues inherent in all observational studies of VE which will be discussed in more detail in what follows. These issues must be dealt with meticulously to obtain reliable VE estimates.

### 6.2.1 Cohort Studies

A cohort consists of a group of subjects that have at least one characteristic in common at a certain point in time. This could be a characteristic related to those individuals such as a common demographic (e.g. including only those aged 65+ at a specific timepoint) or it could be a shared experience (e.g. including only those registered with a specific group of medical practices at a specific timepoint). For studies of influenza VE, this could mean that only subjects eligible for influenza vaccination would be selected for a cohort. This may exclude individuals who are allergic to some of the ingredients used to create the influenza vaccine such as those with an egg allergy. However, we could even apply the least strict selection criteria where the cohort could simply be made up of individuals in the general population who are alive at a point in time. There are two different ways of conducting cohort studies – prospectively such as in studies by Castilla et al. [54], Eurich et al. [99] or retrospectively such as in studies by Herrador et al. [166], Kavanagh et al. [196], Simpson et al. [336].

In the prospective design, the study is preplanned and a cohort is formed by selecting individuals that meet some criteria at the present point in time. The cohort is then followed over a period of time to observe how many report outcomes associated with influenza among vaccinated and unvaccinated individuals. This period of time has to be long enough for a sufficient number of those outcomes (events) to occur so that meaningful statistical analysis can be conducted. The difference in the retrospective design is that the study is created at a time when the follow-up of a cohort has already finished. Consequently, the eligible individuals that are included in the cohort have to be identified and selected retrospectively. This cohort will be composed of individuals who meet certain criteria as at some past point in time. Then data on these individuals for a following period of time in the past is analysed and again, the aim is to observe how many report influenza-associated outcomes among vaccinated and unvaccinated individuals over that period.

In attempts to rank study designs based on a hierarchy of evidence, prospective cohort studies have been ranked higher than retrospective cohort studies [364]. One of the main reasons for this is that in a prospective study, the study can be planned and designed to answer specific questions of interest. For example, a researcher can preplan which variables to collect data on before the study has commenced. In contrast to this, in a retrospective study the researcher has to make use of existing data and so it is more likely that they may not have data on some variables that the researcher believes are important. Further discussion on the strengths and weaknesses of prospective and retrospective cohort study designs can be found in Euser et al. [102].

In some instances, a subset of the cohort can be used to conduct a *nested case-control study*. This may allow for additional estimates of VE based on more specific outcomes associated with influenza such as laboratory-confirmed influenza, which may only be available for some patients in the cohort [54, 184, 336]. An example of where this can happen is when some patients consulting with their general practitioner (GP) for symptoms associated with influenza have their diagnosis confirmed or disconfirmed through laboratory testing. Typically, the majority of patients are diagnosed with influenza based on symptoms alone rather than through laboratory testing as the test result would not affect their treatment. The subset of patients with laboratory test results can be considered as a *validation set* which can be used to validate estimates of VE that were derived using the whole cohort on less specific outcomes. If VE estimates found from the nested case-control study are consistent with those found from the whole cohort, this gives some reassurance about the reliability of estimates, particularly if subjects involved in the nested case-control study were sampled randomly [145].

Computerised databases containing data collected by GPs may allow for vaccinated and unvaccinated individuals to be identified and extracted. Moreover, common unique patient identifiers can be used to link different databases such as vaccination, hospital and GP medical records; in Scotland this unique identifier is the Community Health Index (CHI) number. This allows us to obtain vast amounts of information on patients, possibly also allowing influenza VE to be estimated using various outcomes [196]. This means that large populations can be involved when conducting cohort studies. Even though information for large numbers of people may be available, it is vital that efforts are made to assess if the cohort is representative of the general population. As well as this, it is important to assess if confounding is present and if there is good agreement between information provided in the various databases [362]. Person-time analyses which account for the amount of time each individual contributes to the study are normally used to estimate VE in cohort studies with rate ratios (RRs) calculated by comparing the incidence of influenza in vaccinated compared with unvaccinated individuals. Unadjusted VE estimates are calculated as $(1 - \mathrm{RR}) \times 100\%$ while adjusted VE estimates can be found by methods such as Poisson regression and Cox proportional hazards models.

### 6.2.2 Case-Control Studies

In case-control studies, cases are defined according to the outcome under consideration, and an example of this could be individuals who have laboratory-confirmed influenza. Since the outcome is already known in cases and information on cases has to be traced backwards in time to identify their exposure, case-control studies are always retrospective. In the simplest sense, controls just have to consist of individuals who have not yet experienced the outcome at the end of the study period. A case-control study where controls are selected in this fashion is sometimes called a *traditional* case-control study or a *case-noncase* study. However, often for a closer match between the characteristics of cases and controls, stricter criteria should be used for selecting controls. More specifically, controls should be similar to the cases in terms of their risk of developing the outcome. For studies of influenza VE, controls should come from the same source population as the cases and they should have similar risk of contracting influenza compared with cases. The case-control study would then compare the vaccination status of cases against the vaccination status of controls.

Issues which must be considered when conducting case-control studies include the representativeness of the cases and how to define what individuals should comprise the control group. In order to assess the representativeness of cases, information on potential confounding factors such as their access to healthcare must be available. One of the simplest options for choosing controls is to use the *test-negative* case-control design [84] and several studies of influenza VE have used this design [120, 199, 337, 339–341]. An example of how a test-negative case-control design could be carried out for a study on influenza VE study could be to use individuals consulting with their GP for influenza-like-illness (ILI) or acute respiratory illness (ARI) that also had a swab taken for laboratory testing. The individuals that subsequently tested negative would used as controls. Test-negative controls should have similar health-seeking behaviour to the source population and all patients with laboratory test results would be used in the study. A potential concern with the test negative design comes from the fact that some cases could be misclassified as controls and vice-versa but using a laboratory test with high sensitivity and specificity for influenza should minimise the chance of misclassification occurring.

Another simple method is to select controls randomly from the community. Again, it is important to ensure that selecting controls in this way would create a set of controls that are representative of the population that the cases came from. Furthermore, it is also important that the vaccination coverage for the community is a similar level to the coverage among the population from which cases were taken. For instance, if all cases were patients registered at GP practices, then we would be assuming that there is a similar vaccine coverage in the GP-registered population and the wider community. If the vaccination coverage in community controls is lower, then the VE may be underestimated and this was previously found in a European study of influenza VE [207].

If there are reasons to believe that randomly sampling controls from the community would not produce a representative sample, a more refined criteria may be to randomly sample controls only from GP practices. This may be a particularly good choice if the cases were all patients registered at GP practices. The simplest option here would be a *case-cohort design* where controls are selected from the list of all patients registered with GPs. An obvious disadvantage here is that controls are selected irrespective of whether or not they had the influenza outcome of interest and thus, influenza cases could sometimes be selected as controls. A more complicated option would be a *density* case-control design which is more in line with the traditional case-control study. With this design, controls are selected each time a case is reported, and the pool of patients to select controls from would include only those patients that had not yet reported the influenza outcome of interest at the time the case was reported.

In all case-control designs, a fundamental issue is that adequate sample sizes must be attained. While this is an issue for any study, it can more often be an issue for case-control studies compared with cohort studies as there are generally fewer subjects involved [219]. This is particularly crucial if subgroup analyses are to be conducted such as influenza VE estimates by age group or by influenza subtype. The required sample size depends on many factors such as the desired power of the study, the influenza incidence and the vaccination coverage among the source population [362]. In order to derive VE estimates in a case-control study, it is first necessary to calculate the odds ratio (OR) of being a case among vaccinated individuals compared to unvaccinated individuals. Influenza VE estimates are then found by $\text{VE} = (1 - \text{OR}) \times 100\%$ and adjusted VE estimates in case-control studies are commonly found by using logistic regression.

With so many study design options available for conducting studies of VE, it is no surprise that researchers have already attempted to investigate how accurate VE estimates are from different designs. A simulation study by Orenstein et al. [272] which looked at the accuracy of VE estimates calculated from cohort, traditional case-control and ILI test-negative case-control studies showed that all designs tended to underestimate VE but produced fairly similar results. In general, the traditional case-control design produced the highest and most accurate estimates followed by the test-negative design.

### Indirect Cohort Design

Although the indirect cohort method has the term “cohort” within it, it is actually a modified case-control design that uses non vaccine type cases as controls. This method was introduced by Broome et al. [44] who examined the effectiveness of the pneumococcal polysaccharide vaccine; this is why the indirect cohort method is also called the “Broome method”. Since its introduction, the method has been applied in numerous studies of VE for the pneumococcal vaccine [10, 83, 233, 318].

The idea of using non vaccine type cases as controls came around as they expected that the vaccine would be effective against serotypes contained in the vaccine but would be ineffective against serotypes not contained in the vaccine. This methodology has the advantage of usually providing well-matched controls as in this design, non vaccine type cases (controls) and vaccine type cases (cases) are likely to come from the same population. At the very least, cases and controls should have similar health-seeking behaviour since they all received vaccination. On the other hand, the design makes an assumption that the risk of non vaccine type infection is the same for both vaccinated and unvaccinated individuals and if that is not true, then VE estimates will be biased. As an example, if in reality, the chance of non vaccine type infection is higher in vaccinated individuals then the indirect cohort method would overestimate VE. It has been suggested that there is a lack of evidence to support this assumption [320] and there has also been evidence against the assumption [381]. However, a method was created to attempt to correct for potential biases caused by the violation of this assumption [10] which should allow to method to stay relevant.

The indirect cohort method has not really been applied outwith studies of pneumococcal vaccine. The only scenario where it could be applicable for studying influenza VE is when there are many different strains of influenza circulating in an area at once and the vaccine only includes some of those strains. In practice, sufficient numbers of reported and confirmed vaccine type and non vaccine type influenza cases will almost never be available in the course of an influenza season.

### 6.2.3 Screening Method Study Design

Farrington presented a simple method for estimating VE using aggregated data [109] which has been used in a number of VE studies [217, 303, 306, 386]. The primary concept behind this method is that VE is estimated by comparing vaccine coverage among cases against the vaccine coverage among the population where those cases were derived from. For this to work, we require a random sample of cases of disease, all arising over some given time period in a precisely defined population. From those cases, we have to be able to estimate the proportion of cases that have been vaccinated (PCV). Furthermore, we must also be able to reliably ascertain the proportion of the population vaccinated (PPV). Emphasis can be made on the word "reliably" as for the screening method, the value of PPV used is assumed to be the true value. To estimate VE in a screening method study design, we need the OR of being a vaccinated case relative to the OR of being vaccinated in the population where those cases came from.

$$\frac{\mathrm{OR}_{\mathrm{Cases}}}{\mathrm{OR}_{\mathrm{Population}}} = \frac{\mathrm{PCV}}{(1-\mathrm{PCV})} \Bigg/ \frac{\mathrm{PPV}}{(1-\mathrm{PPV})} \tag{6.1}$$

The VE is then

$$\mathrm{VE} = 1 - \frac{\mathrm{PCV}}{(1-\mathrm{PCV})} \times \frac{(1-\mathrm{PPV})}{\mathrm{PPV}} \tag{6.2}$$

One issue which must be taken into consideration when using the screening method is the potential of bias from confounding variables. Therefore stratification by possible confounding variables such as age and location are often necessary. Logistic regression is commonly used to estimate VE after adjusting for possible confounders. Here, the log-odds of vaccination in cases would be the response variable and the log-odds of vaccination in the population would be entered as an *offset* term in the model. A mathematical description of the logistic regression model helps to clarify the role of the offset term in the screening method.

$$\log\left[\frac{\mathrm{PCV}}{(1-\mathrm{PCV})}\right] = \log\left[\frac{\mathrm{PPV}}{(1-\mathrm{PPV})}\right] + \beta_0 + \beta_i x_1 + \ldots + \beta_k x_k \tag{6.3}$$

In Equation (6.3), we have $k$ covariates (or $k$ possible confounding factors under consideration), $\beta_0$ is the intercept term and the other $\beta_k$ are regression coefficients of the $x_k$ covariates. The offset term has no estimated coefficient and makes the odds of vaccination in the population become the denominator on the left hand side of Equation (6.3) (after exponentiating).

$$
\begin{aligned}
\log\left[\frac{\text{PCV}}{(1-\text{PCV})}\right] - \log\left[\frac{\text{PPV}}{(1-\text{PPV})}\right] &= \beta_0 + \beta_i x_1 + \ldots + \beta_k x_k \\
\log\left[\frac{\text{PCV}}{(1-\text{PCV})}\Big/\frac{\text{PPV}}{(1-\text{PPV})}\right] &= \beta_0 + \beta_i x_1 + \ldots + \beta_k x_k \qquad (6.4) \\
\frac{\text{PCV}}{(1-\text{PCV})}\Big/\frac{\text{PPV}}{(1-\text{PPV})} &= \exp(\beta_0 + \beta_i x_1 + \ldots + \beta_k x_k)
\end{aligned}
$$

An adjusted VE estimate can be found as VE $= 1 - \exp(\beta_0)$. Standard errors obtained from logistic regression can be used to compute confidence intervals for VE estimates.

The screening method has some major limitations. It assumes that the PPV derived is the true value and this is something that can only be known if comprehensive data on vaccinations are being collected. Accurate vaccine coverage rates are critical for the method to produce accurate VE estimates and this point has been emphasised before [306]. If cases and the reference group used to obtain vaccine coverage estimates are from different populations, then VE estimates would be biased. Another shortcoming is that detailed analysis of risk factors would require vaccine coverage statistics, stratified according to those risk factors and this is often not available. Despite these shortcomings, the screening method can be used as a rapid and cheap surveillance tool for routine monitoring to obtain crude VE estimates and this is it's primary purpose.

## 6.3 Issues in Studies of Influenza VE

As well as the choice of study design for conducting a study on influenza VE, there are many other issues that can have a large influence on VE estimates. These include how well the vaccine is matched to what influenza strains are circulating at a particular time in a particular location, the clinical outcomes used to measure VE against and confounding that is almost always present in observational studies.

### 6.3.1 Vaccine Match to Circulating Viruses

There are currently three types of influenza virus (A, B and C) which is an important consideration when making influenza vaccines. Influenza A strains are the most varied and have been found in many warm blooded animals including mammals, birds and humans. Furthermore, out of the three main categories, the A strains most often cause epidemics and can be broken down into subtypes based on the their hemagglutinin and neuraminidase description – 'H' (hemagglutinin) followed by the number of those proteins, and 'N' (neuraminidase) followed by the number of those proteins. For instance, 2009 pandemic influenza which was discussed in earlier chapters was an influenza A virus named as H1N1 in this way. Influenza types B and C are mainly found in humans (although influenza B has also been found in seals [277]). These types are not broken down into subtypes and evolve slower than type A influenza viruses. Compared to influenza types A and B, type C is rare and usually causes only minor reactions.

The degree of vaccine match to the currently circulating strains of influenza is an important factor in studies of influenza VE. Hence, the World Health Organisation (WHO) continually monitors and maps influenza viruses as they circulate around the world. As sufficient batches of the vaccine usually take around six months to produce, the WHO must make predictions of which strains will cause influenza outbreaks during the forthcoming winter season in the northern hemisphere well in advance. Since this is essentially a best guess by the WHO, there is always some uncertainty as to whether the vaccine will be effective. This level of vaccine match can be assessed during the influenza season by obtaining virological results from influenza cases. A good match between circulating virus strains and the vaccine used to protect the population is more likely to result in better vaccine effectiveness; this result has been seen in studies looking at antigenically drifted influenza [270].

A further difficulty is that in any period of influenza activity, there can be co-circulation of influenza viruses such as A and B type influenza viruses. The heterogeneity of co-circulating viral strains causes varying levels of cross-protection, possibly differing by population and outcome studied [263]. These problems make it challenging to ascertain the degree of vaccine match. Another possibility is that different viruses could be circulating in different geographical areas causing further difficulty in creating a well-matched vaccine. For this reason, it may be worthwhile to obtain strain-specific influenza VE estimates if they can be found from the available data [362].

### 6.3.2 Association of Outcomes to Influenza

Influenza VE should ideally be measured using influenza outcomes which are highly specific such as laboratory-confirmed influenza. However, attaining virological confirmation through collecting naso-pharyngeal swabs and performing reverse transcription polymerase chain reaction (RT-PCR) or culture analysis is expensive and time-consuming. In reality, various clinical outcomes are used to monitor influenza and the choice of outcome used for a study may be dictated by the availability of data in a country. Different outcomes can have varying levels of sensitivity, specificity and positive/negative predictive value which are important in gaining an accurate insight into VE.

In particular, the specificity of an outcome has been highlighted as being influential to VE estimates. Using outcomes with imperfect specificity tend to underestimate VE in cohort, traditional case-control and test-negative case-control studies [272]. Table 6.1 gives an overview of different clinical outcomes in relation to their levels of specificity against influenza. In the literature, the clinical outcomes which are most frequently reported include exhibiting symptoms consistent with acute respiratory infections (ARI) or influenza-like-illness (ILI), deaths or hospitalisations from all causes and laboratory-confirmed influenza. These outcomes can be loosely categorised into low, medium and high specificity with VE expected to be more accurate for the more specific outcomes and less accurate when using the less specific outcomes [224].

Even though laboratory-confirmed influenza is the most specific outcome, care must be taken to ensure that the selection of patients to swab is a systematic

| **Level of Specificity** | **Clinical Outcome** |
|---|---|
| Low ($\sim 5\%$) | All-cause mortality. All-cause hospitalisations |
| Medium | All influenza-like-illnesses (without laboratory confirmation). Pneumonia and influenza mortality and hospital admissions |
| High ($\sim 90\%$) | Laboratory-confirmed influenza |

**Table 6.1:** Clinical outcomes associated with influenza by level of specificity. The information in the table was taken from Simonsen et al. [334].

procedure and is consistent throughout the influenza season. An example of a systematic procedure for swabbing would be to sample the first patient at the beginning of the week. During an influenza season, consistency may not be possible if lower proportions of patients are swabbed during peak influenza periods due to limited laboratory capacity. Furthermore, clinicians may be less likely to swab patients who are known to be vaccinated which would cause an overestimation of VE.

The specificity of an outcome is also affected by the incidence of influenza compared with the incidence of other respiratory viruses which are in current-circulation. In seasons with low incidence of influenza in relation to other respiratory viruses, the specificity of an outcome such as ILI will be less specific causing further underestimation of VE [272]. Other factors which affect the specificity of outcomes are that the clinical case definition for identifying cases may vary by countries and could also change during the course of an outbreak. If using laboratory confirmation, an issue of some concern is that detection of influenza using swabs becomes more difficult with time after onset of illness. In general, detection of influenza is rare seven days after onset of illness. Adjusting VE estimates for swabbing delay may be necessary when using laboratory-confirmed endpoints [362].

### 6.3.3 Confounding in Observational Studies

The main idea behind the term "confounding" is that a perceived causal relationship between an independent and dependent variable can be partially (or even wholly) explained by an extraneous variable. This idea is illustrated in Figure 6.1 where $W$ is a confounding variable in the relationship between $X$ and $Y$. There, the variables $X$ and $Y$ both depend on $W$, which can be thought of as *exogenous* since it is determined by variables outside of the system depicted in the causal diagram. If the confounding variable $W$ is not accounted for when determining the effect of $X$ on $Y$, then biased estimates would be obtained. An explanation using less abstract concepts could come from looking at the relationship between alcohol and heart disease. Although there is a relationship between alcohol and heart disease [284], a confounding factor may be smoking as it is associated with alcohol consumption [21] and also increases the risk of heart disease [252]. Therefore, the effect of smoking would have to be accounted for when studying the effect of alcohol on heart disease.

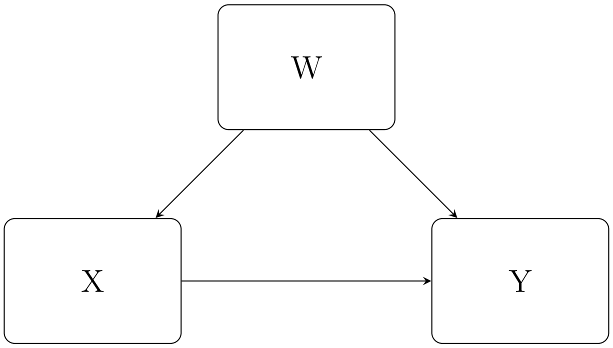


**Figure 6.1:** Causal diagram illustrating a simple case of confounding. In the diagram, $W$ is an extraneous confounding variable which affects both, an independent variable $X$ and and a dependent variable $Y$. As a consequence, $W$ distorts the relationship between $X$ and $Y$ and failure to account for it when determining the effect of $X$ on $Y$ will result in biased estimates.

When studying VE, a confounder would have to be a risk factor for influenza as well as being associated with the likelihood of getting vaccinated. As observational studies do not allocate vaccination randomly to subjects, it is almost certain that some level of confounding will be present. In other words, individuals with certain characteristics will be more or less likely to receive vaccination. It is crucial to attempt to identify these confounding factors as they will cause influenza VE estimates to be biased and adjustments have to be made to reduce this bias. However, finding important confounding factors is challenging as people in the vaccinated and unvaccinated groups differ in ways that can be difficult to measure [182]. There are two types of confounding which can plausibly occur in observational studies of influenza VE – *positive confounding* and *negative confounding*.

Positive confounding in studies of influenza VE can come in the form of the *healthy vaccinee effect* or *extreme frailty bias*. The healthy vaccinee effect is present when individuals who get vaccinated are healthier at the time of vaccination compared with those that do not get vaccinated. If the vaccinated population is over-represented by healthier people, it is likely that estimates of VE will be over-estimated because the vaccinated individuals, as a group, are less likely to report an influenza-related outcome compared with the group of unvaccinated individuals, regardless of vaccination. Positive confounding can also occur when frail individuals or individuals with functional status limitations have fewer opportunities to get vaccinated [183]. Vaccination may not be considered to be worthwhile for these patients as they may have end stage conditions and in this case, VE will again be overestimated if it is not appropriately accounted for. Negative confounding (corresponding to *confounding by indication*) is the opposite case. In the scenario where negative confounding is present, individuals who have more underlying health problems would be more likely to get vaccinated [144], and this would result in VE estimates that are underestimated.

## 6.4 Dealing with Issues in Studies of Influenza VE

Various methods have been designed to deal with confounding in VE studies. Some of these methods are designed to adjust for bias (for instance, propensity scores) while other methods such as modelling a hypothetical unobserved confounder are used only to detect bias.

### 6.4.1 Multivariable Models

When conducting non-randomised studies on influenza VE, it is essential to collect data on possible confounding factors for each individual. This allows for the calculation of adjusted VE estimates using covariates via methods such as Poisson regression or Cox proportional hazards models for cohort studies and logistic regression for case-control studies and the screening method design. However it should be noted that even after paying careful attention to potential confounders, *residual confounding* may still be present and can never be completely ruled out. Residual confounding can occur if there were extra confounding variables that were not adjusted for in models. Common reasons for this happening include data availability issues and a lack of understanding around the causal relationships that affect an outcome. In addition to this, residual confounding can also come from data errors in variables or a lack of measurement accuracy in variables. For example, there may be insufficient adjustment for age if it is grouped into twenty year age bands and there are differences in the risk of experiencing an outcome for individuals within those age categories.

Whilst there is no comprehensive list of all the important factors involved in studying influenza VE, a proposed minimum set of variables to collect has been documented [362], and this has been reproduced in Table 6.2. The proposed variables include smoking habits, underlying chronic diseases and functional status. These variables reflect characteristics of individuals that should affect their tendency to get vaccinated. Variables such as functional status, severity of chronic diseases and healthcare seeking behaviour are not easy to document. Possible proxy measures for functional status include requiring help for bathing and walking, needing to have a GP visit them in their own home, and the Barthel index [69] which is a valid measure of disability that has been used before in studies of influenza VE [353].

The severity of a chronic condition can be measured by examining the number of times an individual was admitted to hospital in the previous year for the condition, and also by examining the prescribing behaviour of individuals. One example of

| **Possible Confounding Variables** |
|---|
| Underlying chronic diseases |
| Severity of chronic diseases |
| Smoking history |
| Previous influenza vaccination |
| Functional status |
| Healthcare seeking behaviour |
| Antiviral usage |

**Table 6.2:** Minimum set of confounding factors to document in studies of influenza vaccine effect proposed by Valenciano et al. [362].

how to measure healthcare seeking behaviour is to look at the number of visits an individual has made to their GP within a previous period of time – for instance, the last couple of years. In addition to these variables, age is an obvious important factor due to the immune system becoming weaker as people grow older (immune senescence) and possible differences in health-seeking behaviour in different age groups. Time period can be influential when vaccination is not distributed to the population simultaneously.

### 6.4.2 Instrumental Variables

Instrumental variables (IV) analysis is a well-known analytical technique, traditionally used for dealing with apparent and hidden bias in observational studies. IV analysis can also be applied in RCT settings in the unwanted event of treatment contamination where patients do not always follow their assigned treatment protocol [351]. It has been well established in settings outside of healthcare such as econometrics [262] but can also be used in VE studies to adjust for unmeasured confounding – for example, in a study of influenza VE in community-dwelling elderly patients [389].

### Definition

IV analysis uses a variable which is strongly associated (correlated) with the factor under consideration but is not directly associated with the outcome variable or any potential confounders, whether observed or unobserved. These are the main assumptions for a valid IV analysis and it must be noted that it may never be possible to prove that the IV is not related to unmeasured variables; at best, this statement can only be strongly inferred. In VE studies, the IV will be strongly related to vaccination status but will not have an independent effect on outcome except through exposure to vaccination. In other words, the IV may have an effect on the likelihood of an individual being opportunistically vaccinated but should not have any impact on the risk of contracting influenza.

### Selection of IVs for Influenza VE studies

In principal, IV analysis makes perfect sense for providing unbiased estimates of influenza VE but finding appropriate IVs for use in studies has proved highly problematic. Analysis using gout as an IV has been suggested as a valid method to deal with confounding in influenza VE studies [398] but a later study rebutted this claim [141]. Thus far, no IVs have been established as being unequivocally appropriate for analyses of VE. Due to the uncertainty as to what a suitable IV would be, many IVs have been considered for VE studies.

One study which set out with the mission of finding IVs for influenza VE studies chose to consider four potential IVs based on classes of comorbidity, classes of medication and physician characteristics [141]. These potential IVs were a history of gout, a history of orthopaedic morbidity, a history of antacid medication and GP-specific vaccination rates. Linear regression analyses were used to find associations between IVs and influenza vaccination status and between IVs and observed confounders. The results showed that all of the IVs considered had some association with vaccination status. In addition to this, all potential IVs showed some apparent relation to mortality (the outcome) through observed confounding variables such as age, sex and comorbidity. These points clearly underline the difficulty in finding a *true* IV.

Despite these difficulties, a reasonable suggestion for an IV was made by Groenwold et al. [141]; they suggest that if patients are randomly allocated to either an encouragement program to get vaccinated or to routine care then it should be reasonable to assume that the encouragement program should only affect the outcome through vaccination. A study looking at influenza VE on community-dwelling elderly patients used census subdivision-specific (CSD) influenza vaccine coverage as an IV [389]. Their study showed that the measured risk factors were balanced across CSDs for all influenza seasons, providing reasonable assurance that unmeasured risk factors may also be balanced across CSDs. The conclusion drawn was that they had found a valid and strong IV. Furthermore, they examined post-influenza periods where a VE should not be noticeable due to the lack of influenza circulating during those times. The IV analysis appeared to produce estimates of VE more in-line with expectations compared with standard regression modelling and hence, IV analysis appeared to produce less biased estimates.

In a study protocol for seasonal influenza VE, variables suggested as candidates for an IV included previous antacid prescription, previous thyroxine prescription, gout and screening attendance [224]. In theory, these variables may indicate that a person is more likely to visit their GP, leading to more opportunities to be vaccinated but having these conditions should not affect the risk of contracting influenza. As previously mentioned, these potential IVs may not be entirely valid and the challenge of searching for appropriate IVs remains.

### 6.4.3 Modelling an Unobserved Confounder

To test how well confounders have been adjusted for, various sensitivity analyses can be performed. One such method is to model or simulate an unobserved confounder. After obtaining results from the main analysis, an attempt to quantify the error due to unmeasured confounding factors is made by repeating the same analysis with the inclusion of a new unobserved confounder. The sensitivity of influenza VE estimates is tested by examining how the estimates vary when values of the unmeasured confounding factor changes. The confounder is required to plausibly affect the outcome and if simulated, has to be based on realistic assumptions.

One study which used all-cause mortality as the outcome chose to use smoking status as a potential unobserved confounder [141]. This variable was chosen as smokers are known to be at an increased risk of mortality compared with non-smokers. As smoking status was not routinely reported in their medical database, different scenarios involving varying percentages of smokers in gout and non-gout groups (gout was included as an IV) were used to test the sensitivity of results. In this case, crude IV analysis showed a positive association between influenza vaccination and mortality risk and a negative association was observed after adjusting for observed confounders. However, after a final adjustment for smoking status, a positive association was again apparent. The study claims that the assumptions used to simulate smoking status were realistic, giving rise to a large possibility of unobserved confounding.

Another study looking at influenza VE in the elderly suggested that possible unmeasured confounders could include race, income and functional status [264]. They modelled an unobserved confounder to cause persons with the confounder to be half as likely to receive vaccination and 2–3 times more likely to be hospitalised or die compared with persons without the confounder. Furthermore, the prevalence of the confounder in the population varied between 20–60%. These sensitivity analyses demonstrated that their original estimates of VE could have been affected by residual confounding. Higher prevalences of the unmeasured confounder caused VE to be much lower (VE against hospitalisations decreased from 27% to as low as 7%), although the effect was still significant.

### 6.4.4 Propensity Scores

We firstly note here that although propensity scores are mostly discussed in the general terms of treatment and control groups, we will keep this discussion focussed on their application to studying VE and will thus, refer to the treatment and control groups as the vaccinated and unvaccinated groups throughout.

Using non-randomised observational study designs can often result in large differences on observed covariates between the vaccinated and unvaccinated groups (selection bias). This is a problem as this would mean that people getting vaccinated have different baseline characteristics from the people in the unvaccinated group, ultimately causing bias in the VE estimates. *Propensity scores* are a method of eliminating (or at least attenuating) bias resulting from these large differences which increases the precision of estimates. The formal definition of a propensity score is "the conditional probability of assignment to a particular treatment given a vector of observed covariates" [312]. In the context of VE, propensity scores are used to attempt to estimate the probability of individuals to get vaccinated given a set of background information related to those individuals (measured covariates) which may plausibly influence the decision to get vaccinated.

The propensity score model's main purpose is to adequately balance the covariates and the end result produced is a scalar summary of all the background information provided by covariates. After calculating propensity scores, groups of subjects with similar values of propensity score are expected to have similar characteristics – at least as far as measured by the covariates fed into the model. Using the probability that a subject would have been vaccinated via propensity scores essentially allows the creation of a *quasi-randomised* experiment. In other words, randomisation is artificially achieved to emulate a RCT and thus, two subjects with very similar propensity scores should be almost equally likely to be assigned to the vaccinated or unvaccinated groups.

### Mathematical Definition

For subjects $i$ ($i = 1, \ldots, N$), the conditional probability of being vaccinated ($Z_i = 1$) and unvaccinated ($Z_i = 0$), given a vector of observed covariates ($x_i$) is

$$e(x_i) = \Pr(Z_i = 1 | X_i = x_i) \tag{6.5}$$

where $e(x_i)$ is the propensity score for a subject $i$. It is assumed that the vector of covariates $\mathbf{X}$ contains all relevant information regarding getting vaccinated meaning that there are no *hidden* biases. Furthermore, given the observed covariates $\mathbf{X}$, it is assumed that the decisions to vaccinate for each subject $(Z_i)$ are independent giving

$$\Pr(Z_1 = z_1, \ldots, Z_N = z_n | X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{z_i} \left[1 - e(x_i)\right]^{1-z_i} \tag{6.6}$$

The propensity score creates a *balancing score* such that the conditional distribution of the covariates $\mathbf{X}$ is the same for vaccinated and unvaccinated units.

### Estimation of Propensity Scores

Two methods can be used to estimate propensity scores in the situation where covariates do not contain missing data; these are discriminant analysis and logistic regression [76]. Using discriminant analysis assumes that observed covariates have a multivariate normal distribution while the more commonly used method of logistic regression does not have this assumption. For the latter method, the vaccination decision is estimated by the inverse logit transformation of the linear predictor derived from the logistic regression model with vaccine status as the outcome. This is shown in Equation (6.7) where the $\beta_i$'s are the regression coefficients for each of the covariates $X_i$.

$$\begin{aligned} \ln\left[\frac{e(x_i)}{1 - e(x_i)}\right] &= \beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N \\ e(x_i) &= \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_N X_N)} \end{aligned} \tag{6.7}$$

For the propensity score model, it is vital to carefully consider including any variables which have some relationship to the decision to vaccinate. This is important to minimise the chance of hidden bias not explained by the measured covariates.

### Applying Propensity Scores

A two-stage process is used to apply propensity scores. Rather than simply fitting a large outcome model that includes all of the covariates in the propensity scores model, using a two-stage approach (fitting a model to obtain propensity scores and then using propensity scores in a final model) has advantages. For instance, parsimony is not required in the propensity score model as it's only goal is to model the likelihood of getting vaccinated. Moreover, this allows us to fit a smaller outcome model in the second stage, making it easier to assess it's validity.

Usually three main techniques are used in applying propensity scores – matching, stratification (subclassification) and regression adjustment. Another possibility is propensity score weighting but as this is less commonly used in the medical literature, we will not discuss it here; for more details on it, see Rosenbaum [315]. Regardless of the technique used, propensity scores are derived in the same manner and should produce unbiased estimates of the VE under the assumption that assignment to vaccination is *strongly ignorable*. This assumption suggests that the decision to get vaccinated can be determined by an independent Bernoulli random variable with some probability of success $\Pr(Z_i = 1|X_i = x_i)$.

After estimating propensity scores, the application of these scores can be applied in different ways. If adjustment for covariates is made prior to calculating VE, then matching and stratification can be used. On the other hand, if adjustment for covariates is made while calculating VE, then stratification and regression adjustment are possible. Here, we focus on explaining propensity score matching and propensity score stratification in more detail as they are more commonly used in the literature.

### Matching

Traditional matching techniques involve choosing unvaccinated subjects to match to vaccinated subjects based on certain background covariates that are believed to be important. The issue here is that it can be difficult to find suitable matches when matching on a large number of important covariates. This problem can be overcome by using propensity scores as matches can be made based on a scalar summary of the covariates rather than matching on a large number of covariates. The idea is that on average, there should be no systematic differences in baseline characteristics between propensity-matched pairs. It has been suggested previously by Austin et al. [17] that matching on propensity scores leads to the least biased treatment effect estimation. However, one consideration which must be taken into account when matching is that there could be a vast reduction in overall sample size. For instance, this could happen when there are many more unvaccinated subjects compared with vaccinated subjects and one-to-one matching is used (only one unvaccinated subject is matched to each vaccinated subject).

Rosenbaum and Rubin [314] give three different ways of matching based on the propensity score which are "nearest available matching on the estimated propensity score", "Mahalanobis metric matching" and "nearest available Mahalanobis metric matching within calipers defined by the propensity score". Here we will only discuss the third method as it is a combination of the other two methods and hence, an explanation of it should provide enough detail to understand the matching process for the simpler two methods. A further reason for choosing to discuss only this method is that they showed that it gave the best balance in covariates out of the three matching methods they tested.

For this method, there are two main steps and we would not only use the propensity scores but would also use a subset of covariates important in the decision to get vaccinated when matching vaccinated and unvaccinated subjects. Following the calculation of propensity scores for all subjects, the first step is to randomly order the vaccinated subjects and then one vaccinated subject is chosen for matching. The unvaccinated subject that is chosen to be matched has to be sufficiently similar to the vaccinated subject; formally, they need to have a propensity score within a *caliper* (preset range) of the vaccinated subject's propensity score. Note that often the logit of the propensity score is used instead of the raw propensity score as it is usually more normally distributed. Some researchers have examined what caliper sizes should be used for best results, with Rosenbaum and Rubin

[314] suggesting that a caliper size of a quarter of the standard deviation of the logit of the propensity scores should be used. More recently, Austin [16] conducted simulation studies and concluded that 0.2 times the standard deviation gave optimal results.

The next step is to make a choice of one unvaccinated subject to match to the chosen vaccinated subject. This has to be selected from only those unvaccinated subjects that have propensity score within the specified caliper of the vaccinated subject's propensity score. This is where Mahalanobis distances are used. These can give a measure for a multivariate effect size or a *distance* which in this scenario, is based on the subset of important covariates. The Mahalanobis distance is defined as

$$d_{ij} = (u_i - v_j)^T \, C^{-1} \, (u_i - v_j) \tag{6.8}$$

where $u$ denotes the values of the matching covariates for vaccinated subject $i$, while $v$ denotes the values of the matching covariates for unvaccinated subjects $j$, and $C$ is the covariance matrix of the matching covariates over the entire group of unvaccinated subjects. The unvaccinated subject that is finally chosen for matching is the one that fulfills $\min(d_{ij})$. This process is repeated until an unvaccinated subject is chosen for every vaccinated subject. If a vaccinated subject cannot be matched to a sufficiently similar unvaccinated subject, then they are excluded.

The process that we just described for matching relates to *greedy matching*. This is because each time a vaccinated subject is chosen (bearing in mind that vaccinated subjects have been ordered randomly), the *best matching* unvaccinated subject is chosen. This happens even if that particular unvaccinated subject may be a better match to another vaccinated subject. It is also possible to use another type of matching called *optimal matching*. Here, the goal in the matching process is to minimise the total within-pair difference of the propensity score which is more computationally expensive than greedy matching. A study by Gu and Rosenbaum [142] found that optimal matching did not perform any better than greedy matching for obtaining balance in matched samples.

### Stratification

Direct stratification of subjects on covariates is usually only practical when there are relatively few covariates. For this reason, propensity scores are useful when stratifying because they summarise numerous background characteristics into one scalar summary. This allows stratification based on just propensity scores rather than numerous covariates. Following the estimation of propensity scores using for example, logistic regression, cut-offs for different strata are typically chosen using the quintiles or deciles of the estimated propensity scores for both the vaccinated and unvaccinated groups combined. Quintiles are a popular choice as it has been stated before that using five strata based on the propensity score is usually sufficient to remove over 90% of the bias due to each of the covariates in the propensity score model [313]. Subjects in the vaccinated and unvaccinated groups subclassified into each of the strata should be similar in most of the characteristics measured by the covariates.

The second stage is to include the propensity score quintiles, possibly along with a subset of other important covariates, into a final model to estimate VE. As stated previously, this will result in a final model containing far fewer covariates than if a one-stage approach was used and this makes the process of model-checking simpler. This method of propensity score stratification was used in a study of influenza VE involving 1,066 adult patients from the Netherlands aged 18–64 with chronic obstructive pulmonary disease (COPD) or asthma [144]. VE was measured against the incidence of influenza associated complications, and a multivariable logistic regression model with age, underlying disease, number of GP visits, gender and health insurance included as covariates was used to calculate propensity scores of being vaccinated. Propensity scores were then categorised into quintiles to allow vaccinees and non-vaccinees with similar propensities of being vaccinated to be matched. The use of propensity scores had a marked effect on the VE estimate and also resulted in a smaller confidence interval. The estimate using multivariable logistic regression without propensity score stratification was $-27\%$

(95% CI: −94% to 16%) while the estimate using propensity score stratification was 14% (95% CI: −35% to 45%).

### Assessing Covariate Balance

When using propensity score methods, it is important to check whether or not the method has achieved its aim of creating covariate balance between the vaccinated and unvaccinated groups. To test the level of balance in baseline covariates, it is possible to use standard statistical methods such as $t$-tests for matching and $F$-tests for stratification. However, it has been argued before that significance testing is not appropriate for comparing balance [13]. This is because the statistic used to assess balance should do so with respect to the sample and not some hypothetical population, and also because the sample size should not affect the value of the statistic. The second issue is often a problem as the sample sizes are often much larger in the unvaccinated (control) group, and this can lead to a perceived increase in balance for matched samples, primarily caused by a reduction in sample size compared with an initial unmatched sample. Therefore much of the literature on propensity scores assesses balance by using the standardised difference, which is not affected by units of measurement or sample sizes.

$$d = \frac{100 \times |\bar{x}_V - \bar{x}_U|}{\sqrt{\frac{s_V^2 + s_U^2}{2}}} \tag{6.9}$$

In Equation (6.9), the subscripts $V$ and $U$ relate to vaccinated and unvaccinated groups, while $\bar{x}$ and $s$ relate to means and standard deviations; multiplication by 100 is used to derive estimates in terms of a percentage. In words, the standardised difference is simply the absolute difference in sample means divided by an estimate of the pooled standard deviation. Although there is no overwhelming consensus on the threshold for an important level of imbalance, a standardised difference of less than 10% is often used to indicate a negligible difference [15]. This particular derivation of the standardised difference only works for continuous numerical variables, but there is a corresponding measure for categorical variables

$$d = \frac{(\hat{p}_V - \hat{p}_U)}{\sqrt{\frac{\hat{p}_V(1-\hat{p}_V)+\hat{p}_U(1-\hat{p}_U)}{2}}} \tag{6.10}$$

where $\hat{p}_V$ and $\hat{p}_U$ denote the proportions in the vaccinated and unvaccinated groups, respectively. This, of course, only works for binary categorical variables; for categorical variables involving more than two categories, it is possible to treat each category as 'yes/no'. Another option proposed before is to use the multivariate Mahalanobis distance metric [397] as defined in Equation (6.8). Perfect balance following propensity score methods should never be expected for all baseline variables but a much greater degree of balance should be expected.

### 6.4.5 Negative Controls

*Negative controls* can be used to detect and resolve uncontrolled confounding and bias in observational studies. The use of negative controls comes with an increased cost of measuring additional variables or measuring variables over longer periods of time but it has been argued that they should be used as a routine precaution in observational studies [223]. This is achieved through checking for an effect that is hypothesised to be implausible. Properly selecting negative controls requires expert knowledge on the subject of interest but if chosen appropriately, this method can be used to investigate the credibility of a study. In studies of influenza VE, two different ways of employing negative controls have commonly been used to identify residual confounding. The first method is to analyse influenza VE in seasons where influenza is not circulating (non-influenza periods). A number of VE studies have taken this approach [99, 182]. The second approach is to measure influenza VE against non-influenza related outcomes and an example of where this has been employed is in the study by Baxter et al. [23].

#### Analysis of Non-Influenza Periods

Observational studies of influenza VE in elderly persons have been controversial due to the large effect sizes witnessed even after adjustment for several covariates related to health status. Some studies have reported approximately 50% mortality reduction in elderly populations [264, 367] and this led to suspicions of uncontrolled confounding. The results seem implausible as vaccination rates among the elderly have generally increased over the years but no proportional reduction in hospital admission rates or all-cause mortality have been evident. Furthermore, US ecologic studies examining excess deaths account for only about 5–10% of all winter-time deaths [334]. These excess death figures seem incompatible with VE figures suggesting 40–50% reductions in mortality.

A cohort study using only persons 65 years of age and older sought to provide evidence of bias using negative controls [182]. The central hypothesis is that a protective effect of the vaccine should not be observable in non-influenza periods (see Table 6.3 for a brief overview of the seasonal influenza periods as described in a protocol by the European Centre for Disease Prevention and Control [100]). Influenza VE should be measured before and after the influenza season so these estimates can be compared with estimates derived during influenza seasons.

| | |
|---|---|
| **Pre-influenza** | Beginning of September to start of influenza season |
| **Influenza season** | Time period with 70% of cases, first to last positive isolate |
| **Peak influenza** | Five week period that includes the two weeks before and after the week of peak influenza circulation |
| **Peri-influenza** | Winter weeks during the influenza period |
| **Post-influenza** | End of seasonal influenza until end of May |
| **Summer** | Beginning of June until end of August |

**Table 6.3:** Seasonal influenza periods as defined in a protocol by the European Centre for Disease Prevention and Control (ECDC) [100]. The exact dates for the influenza season varies every year which is why it cannot be defined more precisely.

If a clear vaccine effect is seen in periods when influenza is not circulating then this is indicative of bias. For a cohort study this means that the relative risk of an outcome for vaccinees compared with non-vaccinees should be approximately one and not be statistically significant in a non-influenza period. Hence, they chose to assess the risk of all-cause mortality and pneumonia or influenza related hospitalisation among vaccinated and unvaccinated individuals before, during and after influenza seasons. A Cox proportional hazards model with adjustment for health status covariates showed that the largest protective effects were observed before influenza seasons. It was suggested that this gave convincing evidence of residual confounding and that the risk reductions almost certainly could not be due to a true VE. Another study by Eurich et al. [99] looked specifically at mortality reduction due to influenza vaccination in non-influenza periods. They also found a roughly 50% reduced risk of mortality outside of the influenza season. After taking into account additional factors such as functional and socioeconomic status, the benefits seen were dramatically lower and more in accordance with the expected null result. This clearly highlights the point that false VEs may be observed, particularly if confounding factors are not properly adjusted for.

### Non-Influenza Related Outcomes

Another approach to employing negative controls in studies of influenza VE is to use outcomes that should not conceivably be linked to influenza. In other words, real protective effects of vaccination should only be observed in outcomes associated with influenza and findings which consistently suggest VE against control conditions give the indication that analysis did not sufficiently account for bias. Trauma-related hospitalisation is one example of a control outcome which has been used in studies of influenza VE as this outcome should not be caused by influenza [23, 182].

In the cohort study of seniors by Jackson et al. [182], injury or trauma related hospitalisation was used as a control outcome. However, influenza vaccination was found to reduce the risk of injury or trauma related hospitalisation even before influenza season, adding more fuel to the argument that confounding may have been inadequately controlled for. The opposite result has also been found before using the same control outcome. For a study on the effect of the 2009 pandemic influenza vaccine in Scotland, vaccinated patients were found to be more likely to be admitted to hospital for trauma, which can be interpreted as a negative VE against trauma [335].

### 6.4.6 Difference-in-Differences Approach

Using the notion that vaccination should have little or no effect in non-influenza periods, a method of assessing influenza VE by differentiating VEs from selection effects was created and coined as a *difference-in-differences* approach by Fireman et al. [119]. This method was employed on the back of the controversy surrounding large influenza VE estimates found in studies of seniors. The observed effects were believed to be due to selection bias not being adequately adjusted for. Specifically, the selection bias was probably due to the situation of mortality being higher in people unlikely to have received the vaccine which could happen, for instance, if the frailest individuals do not get vaccinated. As a consequence, this would give high VE estimates but in reality, would actually be more of a reflection of characteristic differences between vaccinees and non-vaccinees. It has been demonstrated that among seniors, decedents were less likely to get the vaccine than survivors outside of influenza season which gives a clear indication of selection bias [119].

Some researchers believe that it may not be possible to suitably adjust for confounding factors as selection bias is constantly changing with time. Thus, the difference-in-differences approach attempts to circumvent the issues of selection bias and unmeasured confounding. It posits that if an influenza vaccine really is effective in preventing outcomes associated with influenza, then a difference between two differences should be apparent. Firstly, the difference in the odds of prior vaccination between survivors and decedents (supposing the outcome under consideration is death) when influenza is circulating is noted. Then secondly, this is compared to the *expected* difference in the odds of prior vaccination between survivors and decedents on the same days if influenza was not circulating. In essence, the apparent effect of the vaccine outside influenza season is subtracted from the benefits seen during influenza season. In order to estimate VE using this method, a novel approach termed *case-centred logistic regression* is utilised. The method is termed as "case-centred" as the data set used only contains one entry for each case and hence, all the emphasis is placed on the cases.

### Expected Odds of Vaccination

For each decedent, it is necessary to calculate their expected odds of vaccination. Before this can be achieved, all potential *risk sets* (strata) have to be considered. The number of risk sets is equal to the number of days in the study multiplied by the number of age-sex groups. Therefore, if there were four age-sex groups and two years worth of data with 365 days in each year, there would be a total of 2,920 potential risk sets. For each risk set, the vaccine coverage on that day is derived empirically and hence, each risk set for each day contains vaccine coverage estimates for similar groups of people.

The complete set of results are stored in a look-up table so that each decedent's expected odds of vaccination can be found simply by looking up their age-sex group on their day of death. For example, supposing the start date of the study was 1st September 2012, and a man aged 67 died on 1st November 2012, then the expected odds of vaccination would come from the proportion of all males aged 65–69 including the decedent (assuming age groups are in five year age bands) who were vaccinated between 1st September and 1st November.

### Case-Centred Logistic Regression

The expected odds of vaccination for each decedent are entered as an offset term (a regression variable with a fixed coefficient of one) in the case-centred logistic regression model where the dependent variable is the decedent's vaccination status (whether or not an individual had received vaccination before dying). The simplest model with just an offset term and intercept term is

$$\operatorname{logit}(P_{\text{Obs}}) = \operatorname{logit}(P_{\exp}) + \beta_0 \tag{6.11}$$

where $P_{\text{Obs}}$ is the observed probability of vaccination and $P_{\exp}$ is the expected probability of vaccination. Entering the expected odds as an offset term essentially creates a denominator variable for the observed probability of vaccination (much like how an offset term was used in the screening method in Section 6.2.3). After exponentiating the whole equation, the intercept term $\exp(\beta_0)$ estimates the

OR of being vaccinated which can be found to equal the RR after some calculations.

$$\begin{aligned}\frac{\text{logit}(P_{\text{Obs}})}{\text{logit}(P_{\text{exp}})} &= \beta_0 \\ \exp(\beta_0) &= \frac{P(\text{Vac}=1|\text{Case}=1)}{P(\text{Vac}=0|\text{Case}=1)} \Bigg/ \frac{P(\text{Vac}=1)}{P(\text{Vac}=0)}\end{aligned} \tag{6.12}$$

From here, we can concentrate on the right side of Equation (6.12) to derive alternative expressions for $P(\text{Vac}=1|\text{Case}=1)$ and $P(\text{Vac}=0|\text{Case}=1)$. Using Bayes rule, we can begin with $P(\text{Case}=1|\text{Vac}=1)$ to eventually derive $P(\text{Vac}=1|\text{Case}=1)$.

$$P(\text{Case}=1|\text{Vac}=1) = \frac{P(\text{Vac}=1|\text{Case}=1)P(\text{Case}=1)}{P(\text{Vac}=1)} \tag{6.13}$$

This can be rearranged to find $P(\text{Vac}=1|\text{Case}=1)$.

$$P(\text{Vac}=1|\text{Case}=1) = \frac{P(\text{Case}=1|\text{Vac}=1)P(\text{Vac}=1)}{P(\text{Case}=1)} \tag{6.14}$$

We can use the same process to find $P(\text{Vac}=0|\text{Case}=1)$ – substituting what would be found for that along with Equation (6.14) into Equation (6.12) yields

$$\exp(\beta_0) = \frac{P(\text{Case}=1|\text{Vac}=1)}{P(\text{Case}=1|\text{Vac}=0)} = \text{RR} \tag{6.15}$$

In a study of influenza VE in the elderly, time of year was added to the model to measure selection effects over time as well as indicator variables for age group, sex and influenza year [119]. Finally an indicator for influenza season was added which is vital as the indicator for influenza season allows us to observe the difference it makes to the OR which is the goal of the method. VE is estimated by one minus the exponential of the difference that the influenza season indicator makes to the OR. In this way, the difference in the association between mortality and prior vaccination is compared inside and outside influenza seasons to give an estimate of VE.

This approach to estimating VE brings about the advantage of alleviating some concerns about confounding by time and selection effects but some issues still remain. It is particularly useful in the situation where there is much more data available on the time-vaccination association rather than the time-outcome association. However, before this method can be considered, it is required to have a large dataset spanning multiple years, preferably with influenza circulating on different dates in some years in order to increase statistical power. These requirements mean that it may only be possible to obtain estimates of average VE across years rather than year-specific VE estimates. It should also be noted that when using the difference-in-differences approach, the choice of what constitutes the period when influenza is circulating is an important parameter and some sensitivity analyses using different dates for influenza seasons may be beneficial [23]. Another concern is the choice of offset term in the model as it is critical this is correctly specified [301]. The risk sets used for finding the expected odds of vaccination are only created by age and sex, and possible confounding by other potential factors cannot be ruled out.

### 6.4.7 Dynamic Modelling

A relatively novel approach to estimating VE comes through the use of dynamic mathematical models similar to those discussed previously for estimating reproductive numbers in Chapter 4. Several different models were outlined by Gjini and Gomes [133] who aimed to show the use of these models in a conceptual and illustrative sense. To introduce their main ideas, we will give brief details on only the most basic model that they described. This is a susceptible-infectious (SI) model where there is only one pathogen type for a disease that is directly transmitted. Furthermore, vaccination is assumed to be continuously ongoing and while it has a protective effect, it is never impossible for vaccinated individuals to acquire disease. Note that in their method, VE is interpreted from a biological standpoint as it is defined as the reduction in the probability of pathogen acquisition per contact in vaccinated individuals relative to unvaccinated individuals.

Two sets of differential equations are required – one for vaccinated individuals and one for non-vaccinated individuals. Before showing these equations, it is best to clarify the notation involved: superscripts of $u$ and $v$ are used to show vaccinated/unvaccinated status; $\beta$ denotes the transmission rate; $\mu$ gives the birth rate and the equivalent death rate (a commonly used simplification for mathematical models of infectious disease); $\rho$ is the vaccination coverage; and finally, $w$ is a fraction ($0 \leq w \leq 1$) referring to the individual vaccine protection, from which VE is estimated as $1 - w$. For the non-vaccinated group, we have

$$\begin{aligned}\frac{dS^u}{dt} &= \mu(1-\rho) - \beta S^u(I^u + I^v) - \mu S^u \\ \frac{dI^u}{dt} &= \beta S^u(I^u + I^v) - \mu I^u\end{aligned} \tag{6.16}$$

and for the vaccinated group, the system of equations is

$$\begin{aligned}\frac{dS^v}{dt} &= \mu\rho - \beta w S^v(I^u + I^v) - \mu S^v \\ \frac{dI^v}{dt} &= \beta w S^v(I^u + I^v) - \mu I^v\end{aligned} \tag{6.17}$$

The main difficulty with this method comes from the challenge of parameter inference but a proof of principle numerical procedure is provided by Gjini and Gomes [133]. This procedure is based on ODE (ordinary differential equation) model fitting to cross-sectional data using a nonlinear least-squares optimisation algorithm. For more complicated models – which would be required for realistic epidemics – this challenge would be even greater. These would relax assumptions of homogeneous mixing and could include transmission between age groups and different spatial locations.

Another difficulty comes with trying to validate results from dynamic models as this would require complex spatiotemporal (or at least temporal) data which is often not available. However, if sufficient information can be obtained to accurately inform those parameters, it was argued that these models can account for crucial aspects of disease transmission that other statistical methods may not be able to deal with as effectively. This includes accounting for multiple effects simultaneously such as vaccine coverage, transmission rates, temporal effects of vaccination programmes such as changes in exposure over time, and competition between different strains of a disease. The use of dynamic models for estimating VE is clearly still nascent but with advances in data collection and methods for estimating parameters in systems of differential equations, they are likely to become more widely utilised.

## 6.5 Summary

In summary, we have provided in this chapter an overview of some of the various issues surrounding studies of influenza VE. In addition to this, we have outlined various different statistical methods used for measuring and assessing estimates of VE, which are summarised in Table 6.4. Ideally, we would be able to conduct controlled experiments using RCTs to estimate the effectiveness of influenza vaccines. However, as patients who stand to benefit from the vaccine cannot be denied access to it, it is never really possible to conduct an RCT for studying influenza VE and so we have to use observational studies instead. We have discussed some differences between the study designs commonly used to conduct VE studies including the cohort, case-control and screening method study designs. Furthermore, we have talked about some of the difficulties involved with producing influenza vaccines and how the end product may not always match up well to the presently circulating strains of influenza. This is likely to remain a problem, at least until better methods of producing the vaccine have been proven to be safe and can be scaled up to a level for mass production.

Another fundamental issue we considered is how different ways of diagnosing influenza can affect VE estimates. It is not practical or cost-effective to always conduct laboratory-testing for all cases and so other outcomes associated with influenza must be used such as diagnoses of influenza-like-illness based on symptoms. The choice of clinical outcome used is pivotal in the estimation of VE as they have varying levels of specificity and sensitivity. We outlined some different methods of detecting bias; for example, the use of instrumental variables and modelling the effect of unobserved confounding variables. Finally, some different methods to reduce bias and confounding were detailed. These included the commonly used multivariable regression models with the inclusion of appropriate covariates, propensity score methods as well as more recently applied methods to VE studies such as the difference-in-differences approach and dynamic modelling. We will be cognizant of these issues and apply a number of these methods in a study of VE for seasonal influenza using data from Scotland in post-pandemic seasons.

| Method | Main Reference | Details |
|---|---|---|
| Indirect Cohort Design | Broome et al. [44] | Measures VE by using non vaccine type cases as controls. Difficult to apply to influenza as non vaccine type cases are usually rare. |
| Screening Method | Farrington [109] | Uses aggregate data on outcomes and estimates of vaccine coverage. Works better if vaccine uptake in a population is mostly completed before the influenza season starts. |
| Multivariable Models | Valenciano et al. [362] | Adjusts bias against variables included in regression models. The reference suggested gives recommendations for variables which should be included for studies of influenza VE. |
| Instrumental Variables | Sussman and Hayward [351] | Detects bias by using a variable strongly correlated with vaccination but not directly associated with the outcome under consideration. |
| Modelling an Unobserved Confounder | Groenwold et al. [141] | Detects bias usually by simulating the effect of a confounder that could not be included in a model. If the unobserved confounder causes large changes to VE, it may indicate insufficient adjustment in a model. The reference suggested gives an example of where this approach was used. |
| Propensity Scores | D'Agostino [76] | Different methods of applying propensity scores to adjust for bias exist including matching and stratification. Scores are obtained by modelling vaccination probability for individuals and then those scores are used in a matched design or by stratifying patients into similar groups. |
| Negative Controls | Lipsitch et al. [223] | Detects bias by measuring influenza VE in periods where influenza is not circulating or by measuring VE against outcomes not associated with influenza. |
| Difference-in-Differences Approach | Fireman et al. [119] | Adjusts for bias by subtracting apparent effects of a vaccine found outside influenza season from the effect observed during influenza season. |
| Dynamic Modelling | Gjini and Gomes [133] | Systems of differential equations are used to estimate VE. May be able to simultaneously account for multiple effects such as vaccine coverage, transmission rates and temporal effects. However, parameter inference can be challenging and validation of results may require complex data. |

**Table 6.4:** Summary of the different methods of estimating vaccine effect outlined in this chapter, included in the order they appear. The main reference provided may be the original study where the method was proposed or may be a study that offers a useful application or discussion of the method.

# Chapter 7

# Influenza Vaccine Effectiveness in Scotland

In the previous chapter, we described a number of different methods for estimating vaccine effectiveness (VE) and discussed some advantages and disadvantages of using certain study designs and clinical outcomes to measure VE. Here we apply some of those methods to estimate influenza VE in Scotland where relevant data is available from surveillance systems employed there. This data allows us to obtain a range of VE estimates using a variety of methods and outcomes. Obtaining numerous estimates here is useful in order to gain a better understanding of how the methods and outcomes used can affect estimates. The main focus of this chapter is on a cohort of individuals registered with various National Health Service (NHS) General Practitioner's (GP) practices. This group of GP practices is considered to collectively contain patients that are broadly representative of the population of Scotland as a whole. In this chapter, we do not only present the various eventual VE results but also attempt to give details on the steps involved in obtaining those estimates.

## 7.1 Introduction

To meet the challenges of measuring influenza VE, the UK has established surveillance systems in place to closely monitor the annual effectiveness of vaccines [120]. These systems make use of routinely collected epidemiological data taken from individuals presenting in primary care. In Scotland, influenza VE has been estimated using routinely collected data since 2008 at Health Protection Scotland

(HPS). The present study seeks to provide estimates of seasonal influenza VE for seasons 2011–12 and 2012–13 in Scotland by making use of this data which comes from a sample of GP practices. Having data for two seasons gives us the added benefit of being able to compare and contrast our VE estimates for each season when using the same methods to estimate VE.

As discussed in Chapter 6, confounding and bias have to be considered when using observational study designs. We attempt to adjust for bias in this study by the use of covariate adjustment for possible confounding variables as well as propensity score (PS) methods [76]. It must be noted that these adjustments cannot guarantee that bias and confounding will be completely eliminated from the study. However, they should help in reducing bias and methods exist to detect residual bias; for instance, estimating VE using the same methods restricted to a non-influenza time period where a VE should not be found. Using a variety of methods provides a wider array of estimates from which we can draw upon for discussion. This is important as it has already been found before that estimates of VE can differ by choice of study design, the clinical outcomes considered as endpoints for influenza diagnosis and the statistical methods employed [196]. The analysis here can help us to further understand exactly how different methods can influence VE estimates.

With this in mind, we examined the effect of influenza vaccination against different types of GP consultations related to respiratory problems (which are to different extents, clearly indicative of a patient actually having influenza) and all-cause mortality for the entire cohort of patients as well as on a PS-matched sample of patients from the cohort when PS methods were used. We obtained more estimates of VE by using aggregated counts of GP consultations in conjunction with vaccine coverage estimates to calculate VE via the screening method study design originally proposed by Farrington [109]. Furthermore, laboratory confirmation of influenza was available for a subset of patients allowing us to obtain VE estimates using a nested case-control study design – specifically a *test-negative* design was employed where those that tested negative were used as controls in the study.

### 7.1.1 Vaccine Composition

When measuring VE, consideration has to be given to the influenza strains contained in the vaccine. An internationally accepted naming convention for influenza viruses was given by the WHO [391] and this is used to name the strains contained in the vaccine. The convention names the virus according to the type (A, B, C), the type of host for the strain (if not human), the geographical region of origin, strain/lineage number, year of isolation and if the virus is type A, the hemagglutinin and neuraminidase description is also provided.

In both seasons under consideration in this analysis (2011–12 and 2012–13), the vaccine included A and B type influenza viruses as well as the 2009 pandemic influenza strain. Specifically, in the northern hemisphere, the 2011–12 influenza seasonal trivalent vaccine was composed of the following strains [393]:

- A/California/07/2009 (H1N1)pdm09-like virus
- A/Perth/16/2009 (H3N2)-like virus
- B/Brisbane/60/2008-like virus

With the exception of the pandemic strain, the 2012–13 trivalent vaccine was composed differently [395]:

- A/California/07/2009 (H1N1)pdm09-like virus
- A/Victoria/361/2011 (H3N2)-like virus
- B/Wisconsin/1/2010-like virus

## 7.2 Methods

### 7.2.1 Cohort

The data on the cohorts come from the Pandemic Influenza Primary care Reporting (PIPeR) surveillance system – one of the systems used to monitor influenza in Scotland [159]. This system draws from GP practices that contribute to the Practice Team Information (PTI) network [179]. The GP practices on the PTI network have been selected so as to be broadly representative of the Scottish population in terms of age, gender, deprivation and urban/rural mix. For influenza season 2011–12, patient data was available from 22 of these GP practices and for the 2012–13 season, data was available from 25 GP practices. Note than one GP practice changed code between these seasons and this was the practice with code 46199 which changed to 46606. These cohorts make up our study populations for the two seasons. Many of the patients in the 2011–12 cohort were also in the following year's cohort, but there were a large number of new patients in the 2012–13 season due to the additional participating GP practices. Other variations between GP practices in the two seasons were caused by patients leaving GP practices and other new patients joining GP practices.

Every year, the seasonal influenza immunisation programme runs from 1st October to 31st March in the following year. Thus, the cohort of patients that we wished to follow up consisted of individuals who were registered with a GP practice one month before the start of the programme (1st September). In addition to this, patients known to have died or emigrated outside of Scotland before 1st September were removed from the cohort. In total, there were 157,026 patients to follow up for the study in season 2011–12 and 170,774 patients to follow up in season 2012–13, covering around 3% of the population of Scotland in each season. Patients were followed up until 31st March giving us study periods of seven months for each of the two seasons. Anybody that died during the study period was censored at their date of death.

### Risk Groups

In Scotland, the influenza vaccine is recommended and offered free for individuals considered to be at an increased risk of serious complications following influenza infection. Additionally, people that work closely with vulnerable individuals such as healthcare workers are also targeted for vaccination. We defined individuals in the *at-risk* group according to these conditions but restricted to variables for which data was available. Hence, at-risk individuals included all persons aged 65 or over at the beginning of the study period, pregnant women and anybody who had at least one underlying health condition out of the following list: chronic respiratory disease (including asthma), chronic heart disease, chronic liver disease, chronic renal disease, diabetes, immunosuppression and neurological disorders. Furthermore, the group of at-risk individuals also included people in long-stay residential care and people working as carers. Risk group status was assigned at the beginning of the study period and individuals in the at-risk group were considered to be at-risk for the whole study period. A binary variable for risk group status was created for use in covariate adjustments when examining those aged under 65.

### 7.2.2 Vaccine Uptake

In the two influenza seasons under consideration, vaccination exclusively for pandemic influenza had largely ceased but according to the data, some patients still received the pandemic vaccine (see Appendix Tables 7.18 and 7.19 for a breakdown of the types of influenza vaccinations received by patients over the study periods). These are likely to have resulted from coding or date entry errors. Due to this, it was first necessary to isolate the seasonal influenza vaccinations from all influenza vaccinations for the analysis. In addition to those vaccinated during the study period, we also considered anybody that received vaccination up to two months prior to the study start date (from 1st July 2011/1st July 2012) as vaccinated in the current study (see Figure 7.1) – this was to pick up a very small number of individuals who received vaccinations before the commencement of the immunisation programme in Scotland.

The end of season vaccine uptake rates were estimated for all patients in the cohort to get an overall uptake rate. Furthermore, uptake rates were estimated by age groups and for those belonging to risk groups. We also looked at cumulative vaccine uptake throughout the period of the study. This allowed for a comparison of whether uptake trends differed by several factors related to the characteristics

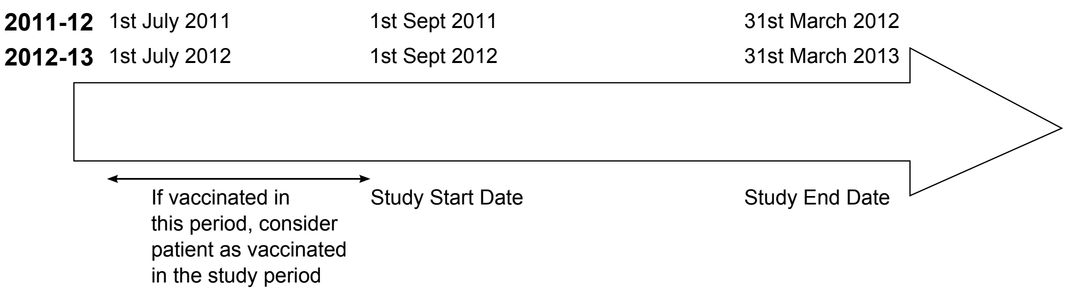


**Figure 7.1:** Timeline showing key dates in the studies.

of patients.

### 7.2.3 Vaccine Effectiveness

For the cohort study, end of season influenza VE against three types of GP consultations were considered as well as all-cause death. The GP consultations data came in the form of *Read codes* which was a system created to allow a standardised way for GPs to record relevant summary information from consultations [181]. Various Read codes were grouped together to form three broad types of consultations – influenza-like illness (ILI), acute respiratory infection (ARI) which includes some of the ILI consultations, and influenza-like illness or acute respiratory infection excluding asthma-related consultations (ILIARI). Note that ILIARI consultations was defined exactly to be a subset of ARI consultations and thus, any ILIARI consultations would also always have been ARI consultations. Details of the Read codes used for different consultation types can be seen in Appendix Table 7.20. The most specific of these consultations outcomes for influenza was ILI, followed by ILIARI and then ARI. All-cause death was the least specific outcome considered.

When measuring VE, a period of 14 days was allowed for vaccination to become *protective* as some time is required for the body to build up immunity to the virus [298]. Thus, if a consultation occurred within a two week period following vaccination, it was treated as a consultation for an unvaccinated individual rather than a vaccinated individual.

Firstly, VE was estimated using a Cox proportional hazards model with robust standard errors [355] to account for clustering of patients within GP practices, and adjustment for all available background covariates. Vaccination status was included as a time-dependent covariate in the Cox model to account for patients getting vaccinated at different times. Those vaccinated between 1st July and 1st September of the study year were treated as being vaccinated for the entire study period. In analyses using consultations outcomes, patients could have multiple events (restricted to one event per day) but for the death outcome, obviously only a single event is possible. Several potential confounding factors were considered and these were assigned at the beginning of the study period and remained constant for the duration of the study. These included gender, age, risk group status, whether or not patients received the seasonal influenza vaccination in the previous year, whether or not patients had ever previously received pandemic influenza vaccination, and the number of ILIARI consultations patients had within the previous year. The data zone (DZ) of residence for each patient was available and this was linked to deprivation as measured by the 2009 Scottish Index of Multiple Deprivation (SIMD) [328]. Furthermore, information on the urban/rural status (categorised as urban, small towns and rural) for patients was also found using the DZ of residence for patients.

By comparing the hazard ratio (HR) of those in the vaccinated group against those in the unvaccinated group, VE can be derived as $1 - \text{HR} \times 100\%$. VE estimates by age groups 0–64 and 65+ were obtained by including an interaction term between age group with vaccination status in the model. The first analysis used all available patients in the cohort in a Cox model and then further VE estimates were derived using a Cox model on a PS matched sample of patients, stratifying on pairs of matched patients. VE estimates were also found using other methods and datasets. These include using PS stratification with all patients in the cohort assigned as belonging to a PS decile; the screening method on aggregated consultations data; and by applying regression methods to the virology data. These are explained in more detail in the following sections.

### 7.2.4 Detection of Residual Bias

On the premise that a true VE should only be observed at times when influenza is circulating, in one analysis, we split the study period into a *pre-influenza* period (1st September to 31st October) and a *during influenza* period (1st November to 31st March), with an expectation of finding no VE in the pre-influenza period. The influenza season time period variable was included as a time-dependent covariate in the Cox model, and an interaction term between vaccination status and influenza period was used to obtain VE split by the pre/during influenza periods; VE estimates by those aged 0–64 and 65+ were found by additionally including age group in the interaction term.

If we assume that the level of bias found in VE estimates from the pre-influenza period remains consistent throughout the influenza period, then we can also derive a VE estimate during the influenza season adjusted for the amount of bias found in the pre-influenza period. This can be calculated by Equation (7.1) where $\mathrm{HR}_{\mathrm{during}}$ is the HR for vaccination in the during influenza period and $\mathrm{HR}_{\mathrm{pre}}$ is the HR for vaccination in the pre-influenza period.

$$\mathrm{VE} = 1 - \exp\left[\log(\mathrm{HR}_{\mathrm{during}}) - \log(\mathrm{HR}_{\mathrm{pre}})\right] \tag{7.1}$$

A study by Jackson et al. [182] found evidence of bias in VE estimates against mortality in seniors by using pre/during influenza season analysis. Therefore, we also investigated the robustness of VE estimates against all-cause death in those aged 65+ by modelling the effect of an unobserved confounder on those estimates in a similar fashion to Groenwold et al. [141]. Smoking is known to increase the risk of mortality but information on smoking was not available on patients in our study. More specifically, previous research by Kenfield et al. [203] found an odds ratio (OR) of approximately 2.8 for all-cause death among smokers compared with non-smokers. Hence, we simulated a variable to represent smoking status for individuals in the cohort, ensuring that the OR of all-cause death among smokers compared with non-smokers was around 2.8. For simulating the variable, we also accounted for varying prevalences of smoking among vaccinated and unvaccinated individuals as this would directly affect VE.

### 7.2.5 Propensity Scores

The PS in our scenario gives an estimate of each individual's probability of vaccination, given their background characteristics (further information can be found in Section 6.4.4), which essentially makes up risk factor profiles for every individual (for risk of vaccination in this case). Propensity scores are primarily used to try to create a quasi-randomised experiment and their sole purpose is to create more balance on the background covariates of individuals in the vaccinated and unvaccinated groups (or more generally, treatment/control groups). We attempted to attenuate the effects of selection bias on influenza VE estimates by using a PS model. It is not necessary to construct a parsimonious PS model as no outcome data is used in the model and hence, all potential confounders were considered as the model itself cannot cause bias in the estimate of influenza VE. The PS methods used in this analysis were PS matching and PS stratification.

#### Propensity Score Model

To create the PS model, we looked at patients that were vaccinated by an end date chosen as a time where vaccine uptake had largely levelled off. This allows us to create a binary variable indicating those who were and were not vaccinated by this time which acts as our response variable in the regression model. PS was modelled using a generalised additive mixed model (GAMM). This model allows the inclusion of nonparametric regression terms, while also allowing the inclusion of random effects as well as fixed effects in the predictor. The reason for using a mixed model here is that observations may not be independent due to the possible clustering of patients in different GP practices. Therefore, we allowed for the inclusion of an additional source of variation caused by the different GP practices. Although vaccination generally increases with increasing age, it is likely that this increase is non-linear and hence, using a smooth term for age in the GAMM should result in a better estimation of propensity scores.

Specifically, as we have a dichotomous dependent variable indicating if a patient was vaccinated ($Y = 1$) or not vaccinated ($Y = 0$) by a chosen end date, a generalised additive mixed-effects logistic regression model including the previously mentioned covariates and random intercepts for different GP practices is appropriate. The 2-level model is given by

$$\ln\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)}\right) = \beta_0 + f(\text{age}_{ij}) + \sum_{p=1}^{P} \beta_p x_{pij} + b_i + \epsilon_{ij} \qquad (7.2)$$

$$b_i \sim N(0, \sigma_b^2) \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad i = 1, 2, \ldots, I \quad j = 1, 2, \ldots, J$$

where the indice $ij$ refers to patient $j$ registered in GP practice $i$, $\beta_0$ is the intercept, $f(\text{age}_{ij})$ relates to the smooth function used for age, $\beta_p$ are the fixed-effect coefficients for each covariate $p$, $x_{ij} = x_{1ij}, \ldots, x_{pij}$ denote the fixed effect regressors, and $b_i$ is the random GP practice effect, which is assumed to follow a normal distribution with mean 0 and variance $\sigma_b^2$.

In the model, there were six possible categorical covariates (five in the model for those aged 65+ as risk-group cannot be included as all people aged 65+ are considered to be at risk) and one smooth term (age). The estimated coefficients $\beta_k$ give the expected change in the log odds of being vaccinated for an increase of one unit in the corresponding predictor covariate $k$ when holding all other covariates constant. Since everybody aged 65+ is targeted for vaccination, the characteristics affecting propensity to consult should differ quite considerably between those aged under 65 and those aged 65+. For this reason, we created separate PS models for those aged 0–64 and those aged 65+.

### Propensity Score Matching

The method employed for PS matching was to perform one-to-one matching (each vaccinated person is matched with one unvaccinated person) based on "nearest available Mahalanobis metric matching within calipers defined by the propensity score" [76]. This means that when searching for unvaccinated individuals to match to vaccinated individuals, only those within a given range (caliper) on certain variables can be considered. Theoretically, this should match vaccinated individuals to unvaccinated individuals who are *similar* to them in terms of the background covariates we include in the PS model. We based the matching on more than only the PS to aid the matching process. The other variables considered in matching were age and for those aged 0–64, we used risk group as well.

Using the logit of the estimated PS was preferred instead of using the raw PS as it is often approximately normally distributed (or at least more so than the raw PS). Austin [16] suggests that caliper size of 0.2 times the standard deviation of the logit PS be used and we followed this recommendation here. For matching those aged 0–64, matching was made exactly on risk group and for age, calipers were decided based on allowing matches to be made within a range of approximately ±3 years. These restrictions ensure that the importance of age and risk group are accounted for when choosing matches.

Once the PS matched sample was obtained, VE was estimated with a Cox model stratifying on pairs of matched patients. No further adjustment for background covariates was necessary on the PS matched sample as both groups should already be similar in all of their measured characteristics due to the matching process. Following matching, the level of balance in covariates was checked by looking at standardised differences [14] between the vaccinated and unvaccinated groups.

### Propensity Score Stratification

For PS stratification, we split individuals up according to the decile of PS that they belonged to. The PS deciles were subsequently entered into a Cox model with further covariate adjustment for age, risk group (risk group included for those aged 0–64 only) and vaccination status to estimate influenza VE. Having adjustment for these covariates keeps the methodology more comparable with our PS matching which used age and risk group for matching. Again, covariate balance was checked after dividing patients into PS deciles by using standardised differences.

### 7.2.6 Screening Method

The screening method as described by Farrington [109] uses aggregate data to measure VE. For this method, the focus was on only those patients with consultations for ILI, ARI or ILIARI. Moreover, we looked at numbers of consultations from vaccinated and unvaccinated individuals over a period of time. Vaccine coverage is required for the screening method and this was estimated at a time point close to the start of the period we looked at for consultations. VE estimates have previously been found to be sensitive to the input estimates used [67]. Hence, we estimated vaccination coverage at different time points for use in the screening method to investigate how much of an impact those estimates would have on VE estimates.

In the screening method analysis, confounding variables were adjusted for by using a mixed-effects logistic regression model with stratification by gender, age group (categorised as 0–64 and 65+), risk group, and a random effect for GP practice. To achieve this, we had to find the numbers of consultations from vaccinated and unvaccinated individuals broken down by gender, age group and risk group within GP practices. We also estimated vaccine coverage broken down by these groups as well.

### 7.2.7 Virology

A number of patients who had acute respiratory symptoms had a virology test for clinical reasons and data on the results of these tests were available from HPS. These swab tests were from all hospital laboratories in Scotland that submitted data to the Electronic Communication of Surveillance in Scotland (ECOSS) system [158] as well as the West of Scotland Regional Virus Lab where most of the testing is done on samples from GP practices. These were utilised in a test-negative case-control study design.

Prior to receiving the data for analysis, the virology data were linked to the vaccinations data using a common unique patient identifier to obtain information on whether or not patients had received an influenza vaccination. Following the linkage, the patient identifier was removed and an anonymised identifier was used for this study. For those that received vaccination, their date of vaccination was also known and with this, it was possible to discover whether or not patients were vaccinated before or after their virology test. If the test date was within 14 days after the date of the vaccination, then the patient was considered to have been tested before we could be reasonably certain that the vaccination had a chance to become protective – in this scenario, we would not treat it as a vaccine failure if the test result was positive. Conversely, if the test date was more than 14 days after the date of vaccination, then the patient was considered to be vaccinated before testing, and a positive test result for one of the influenza types contained in the vaccine would be indicative of vaccine failure.

Some patients had multiple tests and for these patients, it was necessary to extract only one test result and associated date to be used in analysis. There were different possibilities for these patients which differed by whether or not the patient received vaccination. The rules we followed for selecting a test result are displayed in Figure 7.2. For patients that did not receive vaccination and had a positive test result for influenza, we chose the date of their first positive test; if patients never received vaccination and only ever tested negative, then we used the date of their first test. However, for patients that received vaccination, the process was more complicated. For patients that were never tested post-vaccination (plus the two week period for immunity to build up), we followed the rules to choose dates as just described for unvaccinated patients. Finally for patients that were tested post-vaccination, we only looked at their tests post-vaccination and followed those rules for choosing dates.

A generalised additive logistic regression model was used to estimate influenza VE for the virology outcome. An additive model was chosen as it allows for a temporal trend to be modelled using the day of virology sample collection. This adjusts for the situation of individuals receiving vaccination at different times and

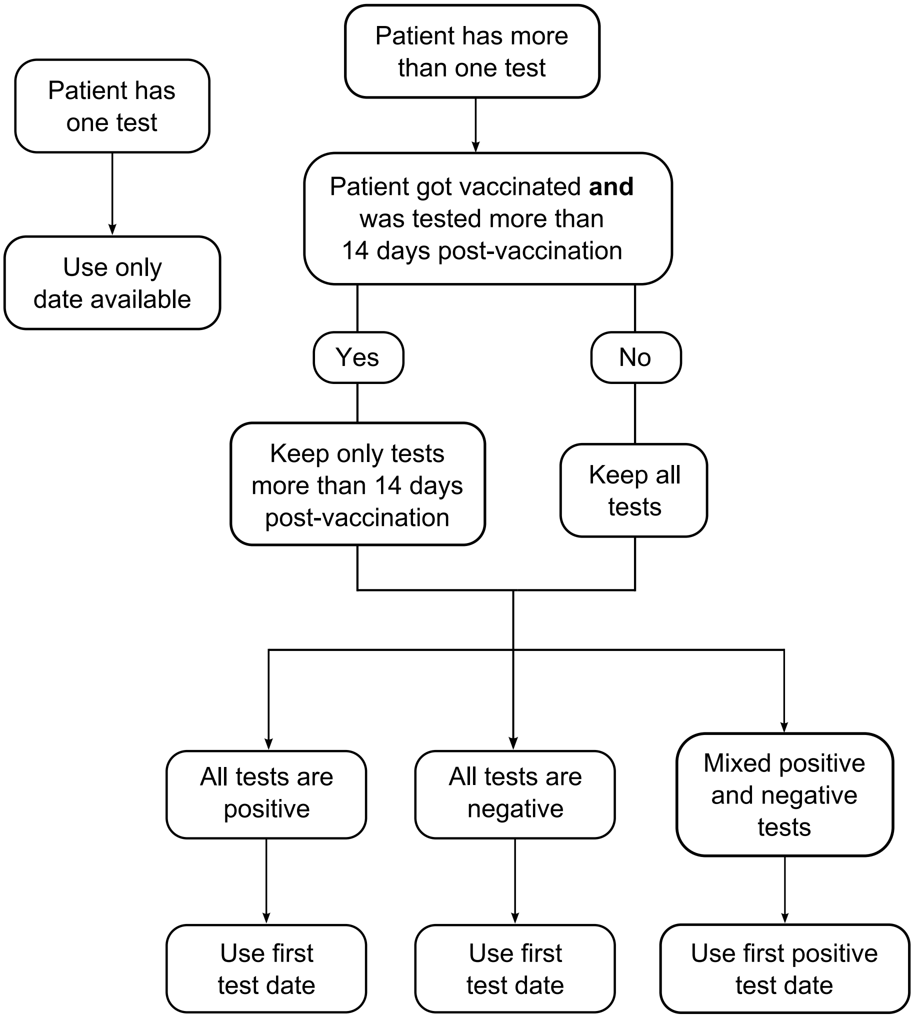

Patient has one test
Use only date available
Patient has more than one test
Patient got vaccinated **and** was tested more than 14 days post-vaccination
Yes
No
Keep only tests more than 14 days post-vaccination
Keep all tests
All tests are positive
All tests are negative
Mixed positive and negative tests
Use first test date
Use first test date
Use first positive test date

**Figure 7.2:** Flow diagram showing how date of virology test for patients was selected under different scenarios.

this is important as the background rate of disease in the community changes over time.

# 7.3 Results

## 7.3.1 Patient Cohort Characteristics

Table 7.1 shows that the gender split in the cohort was almost equal for both seasons and the majority of patients were in the working ages of between 16–64 (over 65% in both seasons). The gender and age split of the cohort matched up very closely to the wider Scottish population [255]. Patients were predominantly from urban areas as we would expect but almost the same number of patients came from small towns and rural areas. Compared with the general population, there were more people in the cohort from urban areas and less from rural areas [258]. As well as this, a larger proportion of the cohort lived in more deprived areas compared with less deprived areas. It should be noted that 1.2% of patients ($n$ = 1,939) in the 2011–12 season could not be matched to a deprivation quintile or urban/rural classification as they were missing information on their DZ; in the 2012–13 season, 2.4% of patients ($n$ = 4,173) did not have DZ information.

Approximately 14% of patients had consulted with their GP for ILIARI at least once within the previous year; 20% of patients had received the seasonal influenza vaccine within the previous year and just under 17% had previously been vaccinated for pandemic influenza. Those figures were consistent for both seasons.

There was considerable variation in the sizes of GP practices as one practice had as much as around 19,000 registered patients while others had as little as 2,000 patients. The amount of variability in GP practice sizes lends support to the notion that there could be some clustering effects within GP practices. These effects would have to be considered in our analyses.

| | | **2011–12** | **2012–13** |
|---|---|---|---|
| **Variable** | **Group** | ***n* (%)** | ***n* (%)** |
| Gender | Female | 79,212 (50.45) | 85,279 (49.94) |
| | Male | 77,814 (49.55) | 85,495 (50.06) |
| Age Group | 0–5 | 8,952 (5.70) | 9,921 (5.81) |
| | 6–15 | 17,244 (10.98) | 18,601 (10.89) |
| | 16–44 | 59,448 (37.86) | 63,922 (37.43) |
| | 45–64 | 44,087 (28.08) | 48,356 (28.32) |
| | 65–74 | 15,229 (9.70) | 16,991 (9.95) |
| | 75+ | 12,066 (7.68) | 12,983 (7.60) |
| Deprivation Quintile | Most Deprived | 33,746 (21.49) | 37,611 (22.02) |

***Continued on next page***

| | | 2011–12 | 2012–13 |
|---|---|---|---|
| **Variable** | **Group** | ***n* (%)** | ***n* (%)** |
| | 2 | 32,901 (20.95) | 35,126 (20.57) |
| | 3 | 32,859 (20.93) | 34,788 (20.37) |
| | 4 | 28,770 (18.32) | 31,484 (18.44) |
| | 5 | 26,811 (17.07) | 27,592 (16.16) |
| | Unknown | 1,939 (1.23) | 4,173 (2.44) |
| Urban/Rural | Urban | 115,178 (73.35) | 125,376 (73.42) |
| | Small Towns | 20,680 (13.17) | 21,275 (12.46) |
| | Rural | 19,229 (12.25) | 19,950 (11.68) |
| | Unknown | 1,939 (1.23) | 4,173 (2.44) |
| Prev Year Seasonal Vacc | No | 125,597 (79.98) | 136,469 (79.91) |
| | Yes | 31,429 (20.02) | 34,305 (20.09) |
| Any Pandemic Vacc | No | 130,985 (83.42) | 143,271 (83.90) |
| | Yes | 26,041 (16.58) | 27,503 (16.10) |
| Prev Year ILIARI Consultations | 0 | 135,507 (86.30) | 146,437 (85.75) |
| | 1 | 14,618 (9.31) | 16,406 (9.61) |
| | 2+ | 6,901 (4.39) | 7,931 (4.64) |
| Practice Code | 16193 | 0 (0.00) | 3,116 (1.82) |
| | 16244 | 2,039 (1.30) | 2,070 (1.21) |
| | 16259 | 4,992 (3.18) | 4,961 (2.91) |
| | 18057 | 14,089 (8.97) | 13,986 (8.19) |
| | 18413 | 8,244 (5.25) | 8,208 (4.81) |
| | 20409 | 8,191 (5.22) | 8,111 (4.75) |
| | 20471 | 8,902 (5.67) | 8,944 (5.24) |
| | 20752 | 19,249 (12.26) | 19,242 (11.27) |
| | 21755 | 9,563 (6.09) | 9,502 (5.56) |
| | 25046 | 12,559 (8.00) | 12,505 (7.32) |
| | 25173 | 5,369 (3.42) | 5,307 (3.11) |
| | 25205 | 8,551 (5.45) | 8,523 (4.99) |
| | 25436 | 2,928 (1.86) | 2,974 (1.74) |
| | 40046 | 9,917 (6.32) | 9,922 (5.81) |
| | 40493 | 2,606 (1.66) | 2,701 (1.58) |
| | 43538 | 2,306 (1.47) | 2,302 (1.35) |
| | 46199 | 3,455 (2.20) | 0 (0.00) |
| | 46606 | 0 (0.00) | 3,417 (2.00) |

***Continued on next page***

| | | 2011–12 | | 2012–13 | |
|---|---|---|---|---|---|
| **Variable** | **Group** | ***n*** **(%)** | | ***n*** **(%)** | |
| | 52382 | 2,925 | (1.86) | 3,093 | (1.81) |
| | 80753 | 13,090 | (8.34) | 13,066 | (7.65) |
| | 80772 | 3,047 | (1.94) | 2,979 | (1.74) |
| | 80895 | 7,292 | (4.64) | 7,442 | (4.36) |
| | 86162 | 0 | (0.00) | 5,749 | (3.37) |
| | 87023 | 3,467 | (2.21) | 3,483 | (2.04) |
| | 87339 | 0 | (0.00) | 5,071 | (2.97) |
| | 87343 | 4,245 | (2.70) | 4,100 | (2.40) |

**Table 7.1:** Counts and percentages for the cohorts of patients used in the influenza VE studies by various characteristics. Note that the GP practice with code 46199 in the 2011–12 season changed to code 46606 in the 2012–13 season.

### Risk Groups

Around 16% of patients aged 0–64 in the 2011–12 cohort were considered to be at-risk from influenza ($n$ = 21,131) and this increased to around 18% for the 2012–13 cohort ($n$ = 25,546). Table 7.2 shows that the most common underlying illnesses for patients aged 0–64 in risk groups were chronic respiratory disease, diabetes and chronic heart disease. Note that information on underlying illnesses was poorly recorded in the data for those aged 65+, but all people in that age category are targeted for influenza vaccination in Scotland so we do not require information on underlying illnesses to assign them into a risk group. Adding all patients aged 65+ to those in a risk group aged 0–64 gave us an overall figure of just under a third of patients being in a risk group in both seasons.

## 7.3.2 Vaccine Uptake

In both seasons, around 22% of patients got the seasonal influenza vaccination ($n$ = 33,876 in 2011–12 and $n$ = 36,764 in 2012–13). Of these vaccinations, 58 were received in the two months prior to September in 2011 and 80 people were vaccinated in the lead up to September in 2012. People that got vaccinated in the two month lead up period were treated as vaccinated for the season under

| | **2011–12** | **2012–13** |
|---|---|---|
| **Risk Factor** | ***n*** **(%)** | ***n*** **(%)** |
| Carer | 1,284 (0.99) | 1,724 (1.22) |
| Heart Disease | 3,364 (2.59) | 4,098 (2.91) |
| Liver Disease | 981 (0.76) | 1,380 (0.98) |
| Renal Disease | 812 (0.63) | 1,004 (0.71) |
| Respiratory Disease | 8,120 (6.26) | 10,317 (7.33) |
| Diabetic | 4,444 (3.43) | 4,734 (3.36) |
| Immunosuppression | 2,146 (1.65) | 2,625 (1.86) |
| Neurological Disease | 2,790 (2.15) | 3,453 (2.45) |
| Pregnant | 1,531 (1.18) | 1,768 (1.26) |
| Care Home Resident | 71 (0.05) | 91 (0.06) |
| Clinical Risk Group | 21,131 (16.29) | 25,546 (18.14) |

**Table 7.2:** Numbers of patients in the cohort aged 0–64 with different risk factors. Percentages are out of the 129,731 patients aged 0–64 in the cohort for the 2011–12 season and out of the 140,800 patients for the 2012–13 season.

consideration.

The vaccine uptake rates for all individuals and split by age groups over the two study periods are shown in Table 7.3. The percentages vaccinated were very similar in the two seasons. Overall vaccine uptake averaged over all age groups was around 20% but there was a clear pattern of higher uptake rates with increasing age. The vaccine uptake rate was only 10% for those aged under 65 but was over 70% for those aged 65+.

The various plots from Figure 7.3 illustrate that most vaccinations were given out in the period of September to the end of November. Since the vaccine uptake rates over time were very similar for the 2012–13 season, we have not included plots for that season here. Essentially, any points made on vaccine uptake here regarding the 2011–12 season can be echoed for the following season. From December, the cumulative vaccine uptake rates were fairly constant showing that very few people received vaccination after December. From Figure 7.3a, it is again evident that vaccine uptake rates increased with increasing age and was particularly high for

| Age Group | Vaccinated | Patients | Uptake (%) | 95% CI |
|---|---|---|---|---|
| *2011–12* | | | | |
| 0–5 | 161 | 8,952 | 1.80 | 1.54 – 2.09 |
| 6–15 | 731 | 17,244 | 4.24 | 3.74 – 4.80 |
| 16–44 | 4,041 | 59,448 | 6.80 | 6.18 – 7.48 |
| 45–64 | 8,277 | 44,087 | 18.77 | 17.52 – 20.10 |
| 65–74 | 10,882 | 15,229 | 71.46 | 68.80 – 73.97 |
| 75+ | 9,409 | 12,066 | 77.98 | 75.56 – 80.22 |
| **Total** | 33,501 | 157,026 | 21.33 | 20.46 – 22.23 |
| *2012–13* | | | | |
| 0–5 | 173 | 9,921 | 1.74 | 1.45 – 2.09 |
| 6–15 | 745 | 18,601 | 4.01 | 3.46 – 4.63 |
| 16–44 | 4,362 | 63,922 | 6.82 | 6.21 – 7.50 |
| 45–64 | 8,752 | 48,356 | 18.10 | 16.93 – 19.33 |
| 65–74 | 12,196 | 16,991 | 71.78 | 67.91 – 75.35 |
| 75+ | 10,124 | 12,983 | 77.98 | 75.86 – 79.96 |
| **Total** | 36,352 | 170,774 | 21.29 | 20.20 – 22.41 |

**Table 7.3:** Seasonal influenza vaccine uptake by age groups. The number of patients vaccinated includes anybody that received vaccination in the period from 1st July up to 31st March in the following year.

those aged 65+. Although vaccine uptake rates were lowest for those living in areas belonging to the most deprived quintile, there was no clear indication that levels of deprivation had a large influence on vaccination rates as individuals in deprivation quintile 2 (more deprived) had the highest vaccination rates. Those with unknown deprivation had very low vaccine uptake. Part of the reason for this was due to some of these individuals leaving their GP practice early in the study period and if they got vaccinated after leaving their practice, their data would not have been linked.

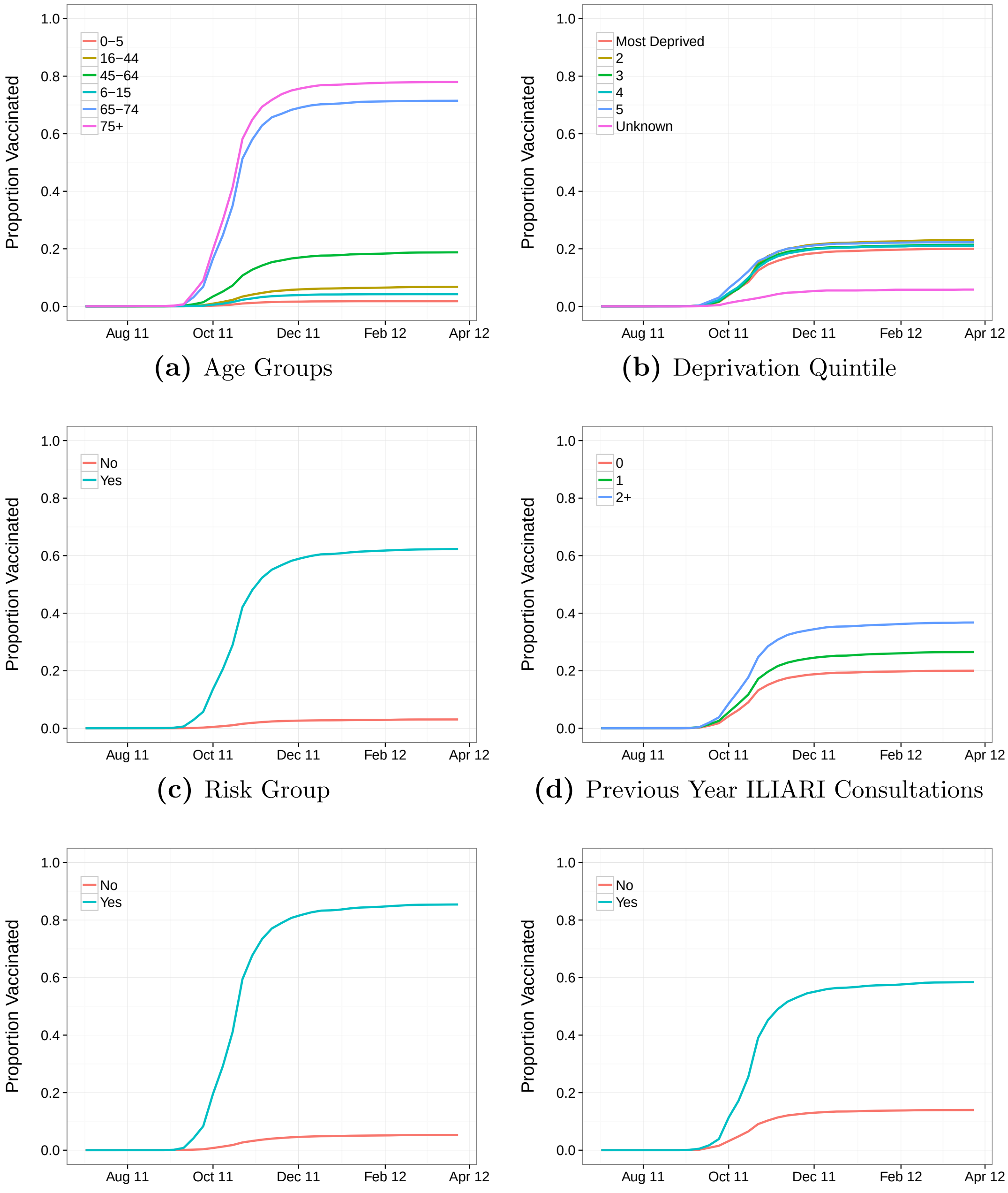
Patients in a risk group had much higher vaccination uptake compared to those not belonging to a risk group as we would expect (Figure 7.3c). As well as this, those that consulted with their GP for ILIARI more often within the previous year appeared to be more likely to get vaccinated as depicted in Figure 7.3d.

Furthermore, Figures 7.3e and 7.3f show that individuals who had previously been vaccinated for pandemic influenza and those who received the seasonal influenza vaccination within the previous year were much more likely to have been vaccinated for seasonal influenza in the current season. A reason for this is due to the targeted vaccination for those in a risk group – many of the individuals that were vaccinated in the previous year or previously vaccinated for pandemic influenza would have been in a risk group at that time and would remain in a risk group for the 2011–12 season as well. Vaccine uptake over the period was very similar for males and females and by urban/rural classification (plots not shown).

### Vaccine Uptake for Individuals At-Risk

Out of the group of at-risk individuals (which includes all patients aged 65+), around 65% received the seasonal vaccine in each season ($n$ = 30,167 in 2011–12 and $n$ = 34,604 in 2012–13). Within the 0–64 age group, uptake varied between around 30–55% for patients at-risk and was highest for those aged 44–64 (Table 7.4).

Figure 7.4a shows that vaccine uptake for at-risk patients aged under 65 was almost always lower than the uptake for those aged 65+ and the difference in uptake rates began to accelerate rapidly from around mid-October. By the end of the season, uptake was about 25–30% lower for those aged 0–64 at-risk than for those aged 65+. However, it is clear that vaccine uptake was much higher for those aged 0–64 at-risk than for all patients aged 0–64. Figure 7.4b illustrates a more vivid pattern of higher uptake for individuals residing in less deprived areas when only considering those at-risk compared with when looking at the whole cohort. Similar patterns to what was observed in the analysis of vaccine uptake for all patients emerged when examining vaccine uptake for only those at-risk split by number of GP consultations for ILIARI within the previous year, any previous pandemic influenza vaccination and having seasonal influenza vaccination within the previous year (plots not shown).



**(a)** Age Groups

**(b)** Deprivation Quintile

**(c)** Risk Group

**(d)** Previous Year ILIARI Consultations

**(e)** Previous Year Seasonal Vaccine

**(f)** Any Pandemic Vaccine

**Figure 7.3:** Vaccine uptake over time for the study period in the 2011–12 season (including the 2 months prior to the starting date of the study) split by various characteristics. Vaccine uptake patterns in the 2012–13 season were very similar.

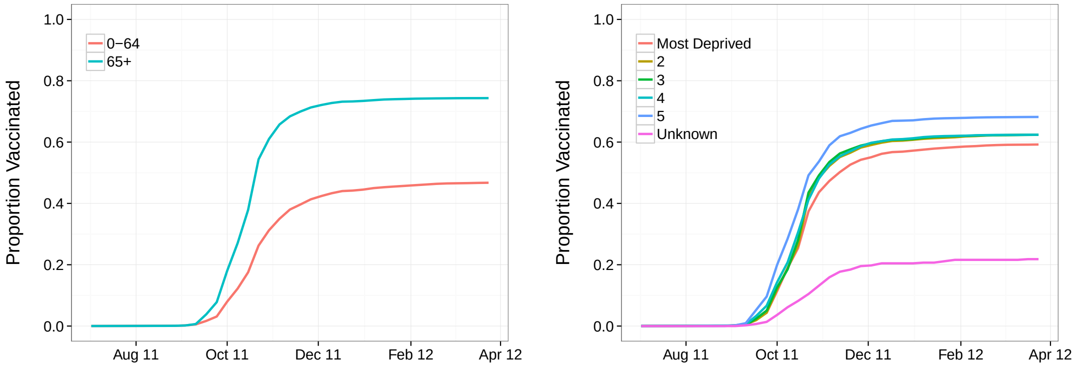
| Age Group | Vaccinated | Patients At-Risk | Uptake (%) | 95% CI |
|---|---|---|---|---|
| *2011–12* | | | | |
| 0–5 | 100 | 337 | 29.67 | 24.05 – 35.99 |
| 6–15 | 561 | 1,424 | 39.40 | 34.93 – 44.05 |
| 16–44 | 2,937 | 7,911 | 37.13 | 33.88 – 40.49 |
| 45–64 | 6,278 | 11,459 | 54.79 | 51.73 – 57.81 |
| 65–74 | 10,882 | 15,229 | 71.46 | 68.80 – 73.97 |
| 75+ | 9,409 | 12,066 | 77.98 | 75.56 – 80.22 |
| **Total** | 30,167 | 48,426 | 62.30 | 60.04 – 64.49 |
| *2012–13* | | | | |
| 0–5 | 125 | 431 | 29.00 | 23.02 – 35.82 |
| 6–15 | 640 | 1,712 | 37.38 | 31.84 – 43.27 |
| 16–44 | 3,691 | 9,358 | 39.44 | 36.15 – 42.84 |
| 45–64 | 7,828 | 14,045 | 55.74 | 52.73 – 58.70 |
| 65–74 | 12,196 | 16,991 | 71.78 | 67.91 – 75.35 |
| 75+ | 10,124 | 12,983 | 77.98 | 75.86 – 79.96 |
| **Total** | 34,604 | 55,520 | 62.33 | 59.79 – 64.80 |

**Table 7.4:** Seasonal influenza vaccine uptake by age groups for patients in risk groups. The number of patients vaccinated includes anybody that received vaccination in the period from 1st July up to 31st March in the following year. Those aged 0–64 considered to be at-risk include care workers, those that are pregnant and those that have at least one underlying illness. All individuals aged 65+ were considered to be at-risk.

### 7.3.3 Consultations and Deaths

The percentages of patients from the cohort within different characteristic groupings that consulted with their GP for ARI, ILIARI or ILI in each season are illustrated in Figure 7.5. Similar patterns on which types of patients consulted were seen in both seasons but the percentages were higher in 2012–13.

Although the gender split in the cohort was almost even, a higher percentage of females consulted. Those aged 0–5 were most likely to have a consultation, especially in 2012–2013 where over 30% of the cohort aged 0–5 consulted. This result can be expected since respiratory infections are among the most common reasons for



**(a)** At-Risk by Age Group (0–64, 65+) **(b)** At-Risk by Deprivation Quintile

**Figure 7.4:** Vaccine uptake in 2011–12 over time for those in risk groups split by age group (under 65 and 65+) and by deprivation quintile. Vaccine uptake patterns in the 2012–13 season were very similar.

young children to consult [49]. Outwith infants, older patients generally consulted more. The best indicator of patients having a consultation was if they consulted within the previous year for ILIARI. Additionally, those in risk groups or with previous influenza vaccination were also more likely to have a consultation. Slightly higher percentages of those from more deprived areas consulted but percentages were similar by urban/rural classification.

Weekly numbers of consultations for ILI, ARI and ILIARI by whether or not the individual was vaccinated or unvaccinated at the time of consulting are presented in Figure 7.6; weekly deaths by the same criteria are also shown there. Note that only data relating to complete weeks in the study periods were used for the plots. Therefore, two days worth of data at the end of the 2011–12 study period was omitted from the plot while six days worth of data at the end of the 2012–13 study period was omitted.

There were clearly not many ILI consultations in the 2011–12 season (58 for unvaccinated patients and 21 for vaccinated patients by the end of the study period). This indicates that there was not very much influenza circulating in Scotland during that season. Moreover, this gives an early indication that VE estimates should be fairly low for 2011–12. Since there were so few consultations for ILI, we excluded

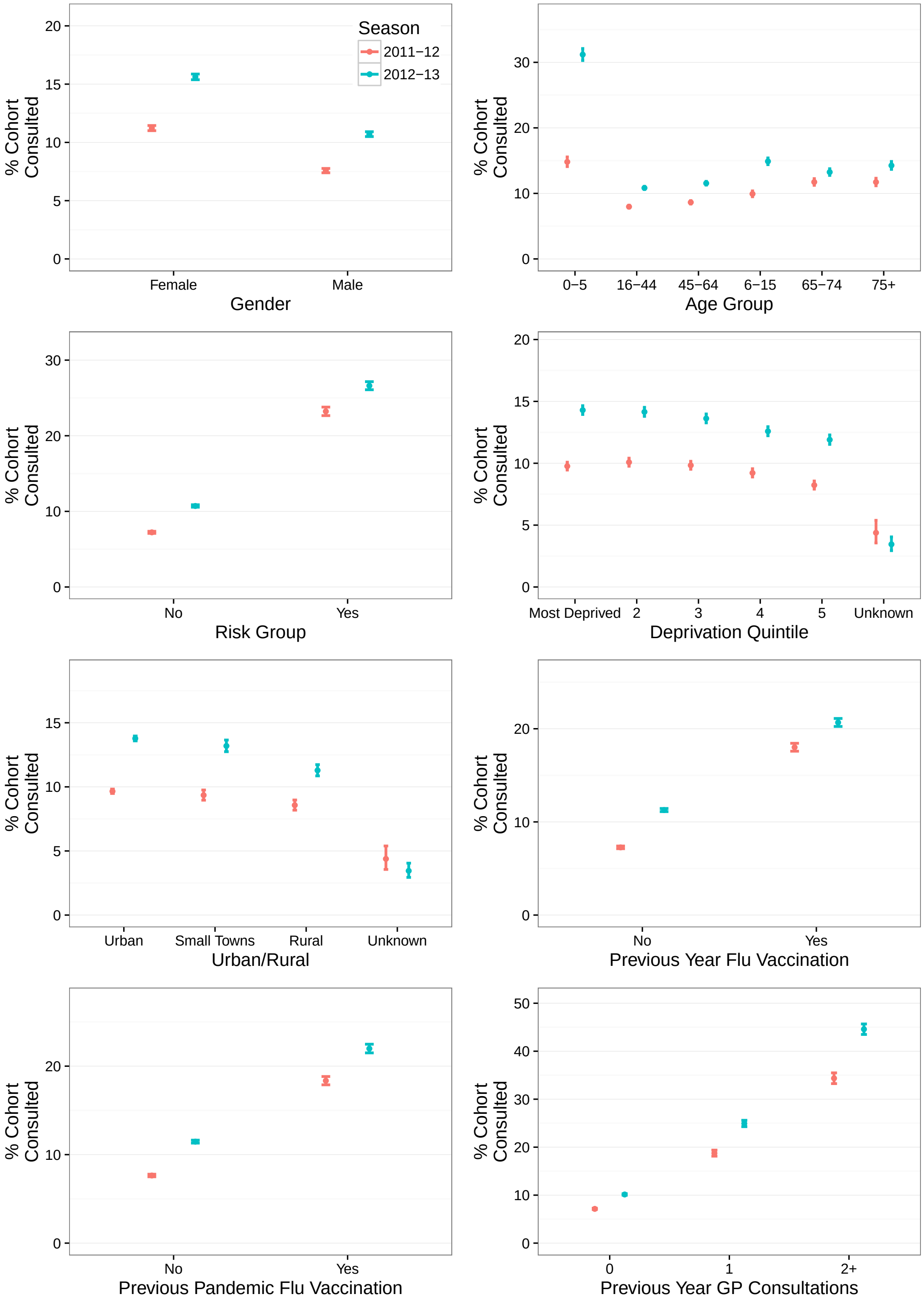


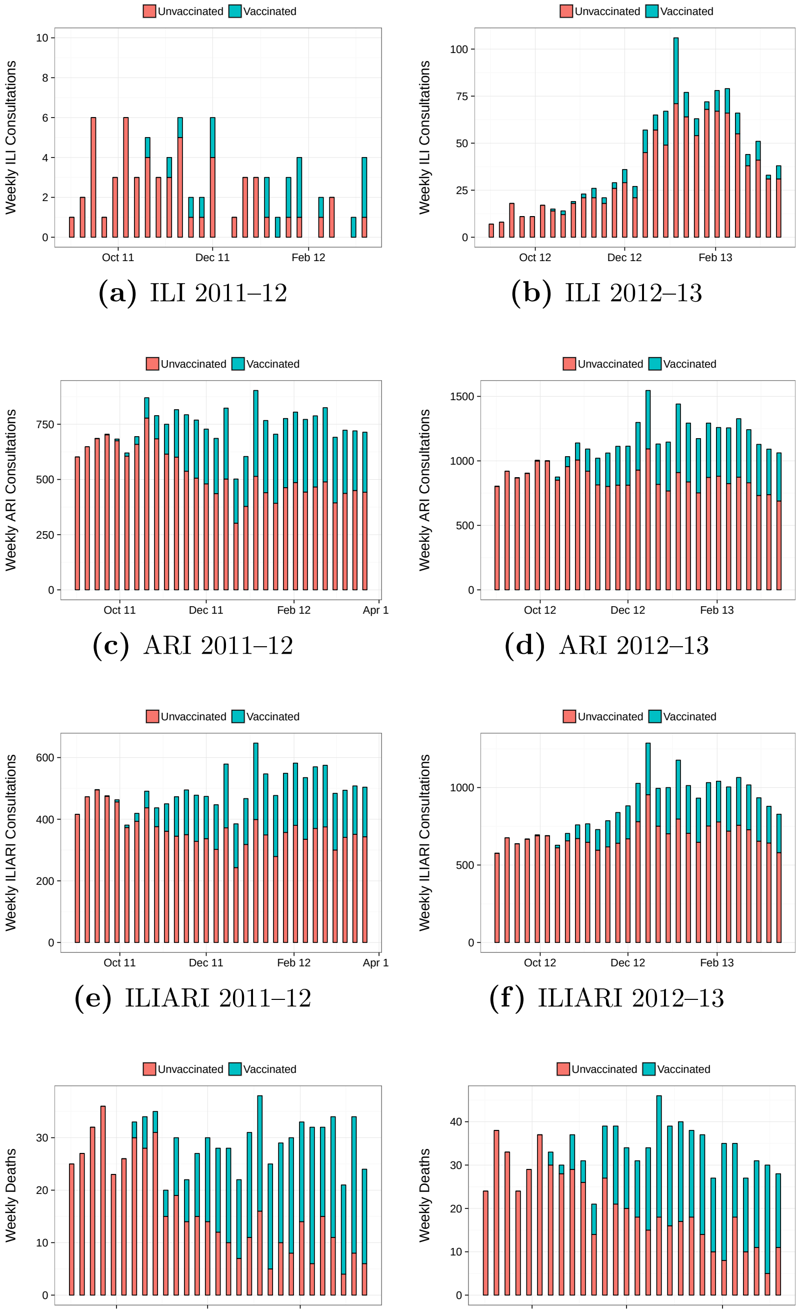
**Figure 7.5:** Percentages of cohort patients within different groups that consulted. Confidence intervals were calculated using Wilson's method [3]. Note that the $y$-axis scale varies by plot.

using them as a clinical outcome in later analyses on VE for 2011–12. However, as there were over 1,000 ILI consultations for the 2012–13 season, we could use them to estimate VE for that season. Furthermore in 2012–13, there was a clear peak in ILI consultations around January 2013 (Figure 7.6b). Weekly numbers of ARI and ILIARI consultations were fairly consistent over time in 2011–12 except around the holiday period (end of December and beginning of January) where more variability can be expected. In 2012–13, there was a noticeable increase in ARI and ILIARI consultations from around December to February, particularly for ILIARI.

Over the seven month study period for 2011–12, there were 885 deaths, giving us a crude death rate of 9.7 per 1,000 population per year (95% CI: 9.1–10.3). In 2012–13, there were 1,005 deaths giving a crude death rate of 10.1 per 1,000 population per year (95% CI: 9.5–10.7). In both seasons, the crude death rates were a little lower than the overall death rate in Scotland of 10.3 per 1,000 population [130]. Figures 7.6g and 7.6h illustrate that from around mid-December, there was a general trend of more weekly deaths from vaccinated individuals than from unvaccinated individuals – this was particularly true for the 2011–12 season. This pattern occurs because of the higher rates of vaccination in older individuals who are also more at risk of death, particularly in winter.

### 7.3.4 Vaccine Effectiveness Estimates

The adjusted HRs of having ARI and ILIARI consultations in the 2011–12 season are given in Table 7.5, while the adjusted HRs of having ARI, ILIARI and ILI consultations in the 2012–13 season are shown in Table 7.6. To obtain the HRs, two models were fitted – firstly a model without interaction between vaccination and age group was used to derive HRs needed to obtain VE for all ages, and then secondly a model including the interaction term was included to add in the HRs required for age-specific VE estimates. Note that since the patients that had unknown deprivation status were the same patients who had unknown urban/rural category, we chose not to include both variables in the model as then estimates would not be obtained for all coefficients due to the issue of perfect multicollinearity. Therefore, only deprivation was used as this generally had a larger influence on



**(a)** ILI 2011–12 **(b)** ILI 2012–13

**(c)** ARI 2011–12 **(d)** ARI 2012–13

**(e)** ILIARI 2011–12 **(f)** ILIARI 2012–13

**(g)** All-Cause Deaths 2011–12 **(h)** All-Cause Deaths 2012–13

**Figure 7.6:** Weekly numbers of GP consultations for ILI, ARI and ILIARI as well as weekly numbers of all-cause deaths over time by whether or not patients had received seasonal influenza vaccination prior to their consultation or death. Only data for complete weeks in the study periods were used – two days of data at the end of the 2011–12 period was omitted and six days of data at the end of the 2012–13 period was omitted. Note that the $y$-axis scale varies considerably by plot.

results. The VE can be calculated from the HR comparing the vaccinated group against the unvaccinated group, and the VE for those aged 0–64 and 65+ can be derived from the HRs related to the interaction between vaccination and age.

For ARI consultations in 2011–12, the VE was estimated to be 4% (95% CI: −4% to 11%); VE estimates against ARI were similar when splitting between those aged 0–64 and 65+. For the ILIARI consultation outcome, VE was slightly lower. VE against ILIARI for all ages was −5% (−12% to 1%), and again, VE did not differ much when looking separately at under 65s and 65+. Hence, for the two consultation types with sufficient amounts of data for analysis in 2011–12, this method produced very low VE estimates and none of the VEs were significant.

In the 2012–13 season, VE estimates were higher. VE against ARI consultations for all ages was 11% (4% to 17%) and against ILIARI consultations, VE was 4% (−2% to 10%). VE for these two consultation types was also similar when looking at those aged 0–64 and 65+ separately. The highest VE estimates were obtained when measured against ILI consultations where it was 39% (23% to 51%). For ILI, estimates differed between those aged 0–64 and those aged 65+, with higher VE estimates for those aged 0–64. VE for those aged 0–64 was 42% (28% to 54%) and for those aged 65+, VE was 31% (−8% to 56%).

Besides VE, the HRs in Tables 7.5 and 7.6 also tell us the effect of the different variables included in the model on the risk of having consultations. The effects of variables were very similar for the two seasons for ARI and ILIARI so we can focus on 2012–13 which has results for ILI consultations as well. There were similarities for all three consultation types – males (compared with females) had a significantly lower risk of having consultations; being in a risk group and having ILIARI consultations within the previous year significantly increased the risk of having consultations; deprivation had no major impact. However, there were also some differences between ARI/ILIARI compared with ILI. Those aged 0–5 had a significantly higher risk of having ARI/ILIARI consultations compared with all other age groups, but there were no significant differences by age group for

| | ARI | | ILIARI | |
|---|---|---|---|---|
| | **HR** | **95% CI** | **HR** | **95% CI** |
| *Vaccinated* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 0.96 | 0.89 – 1.04 | 1.05 | 0.99 – 1.12 |
| *Interactions* | | | | |
| Age 0–64, Vacc | 0.93 | 0.83 – 1.04 | 1.03 | 0.95 – 1.11 |
| Age 65+, Vacc | 1.03 | 0.96 – 1.10 | 1.08 | 0.99 – 1.17 |
| *Age Group* | | | | |
| 0–5 | 1.00 | | 1.00 | |
| 6–15 | 0.72 | 0.67 – 0.79 | 0.60 | 0.54 – 0.67 |
| 16–44 | 0.52 | 0.48 – 0.57 | 0.51 | 0.47 – 0.55 |
| 45–64 | 0.43 | 0.39 – 0.47 | 0.50 | 0.46 – 0.56 |
| 65–74 | 0.86 | 0.75 – 0.99 | 0.77 | 0.66 – 0.89 |
| 75+ | 0.75 | 0.65 – 0.87 | 0.80 | 0.68 – 0.94 |
| *Gender* | | | | |
| Female | 1.00 | | 1.00 | |
| Male | 0.78 | 0.74 – 0.82 | 0.74 | 0.70 – 0.80 |
| *Risk Group* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 3.58 | 3.26 – 3.94 | 1.58 | 1.46 – 1.71 |
| *Dep Quintile* | | | | |
| Most Deprived | 1.00 | | 1.00 | |
| 2 | 1.04 | 0.94 – 1.16 | 1.03 | 0.92 – 1.15 |
| 3 | 1.10 | 0.97 – 1.25 | 1.11 | 0.96 – 1.27 |
| 4 | 1.07 | 0.96 – 1.19 | 1.03 | 0.92 – 1.15 |
| 5 | 1.00 | 0.76 – 1.32 | 0.92 | 0.73 – 1.17 |
| Unknown | 0.71 | 0.47 – 1.05 | 0.64 | 0.39 – 1.04 |
| *PY Seasonal Vacc* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 1.61 | 1.41 – 1.85 | 1.30 | 1.21 – 1.38 |
| *Prev Pandemic Vacc* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 1.25 | 1.10 – 1.44 | 1.17 | 1.07 – 1.27 |
| *PY ILIARI Cons* | | | | |
| 0 | 1.00 | | 1.00 | |
| 1 | 2.29 | 2.09 – 2.51 | 2.92 | 2.61 – 3.28 |
| 2+ | 4.65 | 4.06 – 5.33 | 6.89 | 5.72 – 8.30 |

**Table 7.5:** Hazard ratios of having ARI and ILIARI consultations found from a Cox proportional hazards regression model for season 2011–12. Vaccine effect can be derived from the hazard ratio for the vaccinated group, and vaccine effect for those aged 0–64 and 65+ can be found from the interaction terms. Note that ILI consultations were not considered as there were not enough ILI consultations recorded for the 2011–12 season.

| | **ARI** | | **ILIARI** | | **ILI** | |
|---|---|---|---|---|---|---|
| | **HR** | **95% CI** | **HR** | **95% CI** | **HR** | **95% CI** |
| *Vaccinated* | | | | | | |
| No | 1.00 | | 1.00 | | | |
| Yes | 0.89 | 0.83 – 0.96 | 0.96 | 0.90 – 1.02 | 0.61 | 0.49 – 0.77 |
| *Interactions* | | | | | | |
| Age 0–64, Vacc | 0.87 | 0.80 – 0.95 | 0.99 | 0.93 – 1.05 | 0.58 | 0.46 – 0.72 |
| Age 65+, Vacc | 0.92 | 0.84 – 1.01 | 0.93 | 0.84 – 1.03 | 0.69 | 0.44 – 1.08 |
| *Age Group* | | | | | | |
| 0–5 | 1.00 | | 1.00 | | | |
| 6–15 | 0.49 | 0.46 – 0.53 | 0.44 | 0.41 – 0.48 | 0.97 | 0.45 – 2.09 |
| 16–44 | 0.35 | 0.32 – 0.38 | 0.34 | 0.32 – 0.37 | 1.82 | 0.80 – 4.15 |
| 45–64 | 0.30 | 0.28 – 0.33 | 0.33 | 0.30 – 0.37 | 2.21 | 0.92 – 5.30 |
| 65–74 | 0.48 | 0.42 – 0.55 | 0.44 | 0.38 – 0.50 | 1.56 | 0.82 – 2.95 |
| 75+ | 0.48 | 0.42 – 0.54 | 0.49 | 0.43 – 0.56 | 1.27 | 0.60 – 2.69 |
| *Gender* | | | | | | |
| Female | 1.00 | | 1.00 | | | |
| Male | 0.75 | 0.72 – 0.78 | 0.73 | 0.70 – 0.76 | 0.79 | 0.69 – 0.90 |
| *Risk Group* | | | | | | |
| No | 1.00 | | 1.00 | | | |
| Yes | 2.87 | 2.64 – 3.13 | 1.52 | 1.39 – 1.66 | 1.40 | 1.15 – 1.71 |
| *Dep Quintile* | | | | | | |
| Most Deprived | 1.00 | | 1.00 | | | |
| 2 | 0.99 | 0.93 – 1.04 | 0.98 | 0.92 – 1.04 | 0.97 | 0.71 – 1.33 |
| 3 | 0.99 | 0.92 – 1.06 | 0.96 | 0.88 – 1.04 | 0.82 | 0.55 – 1.21 |
| 4 | 0.95 | 0.87 – 1.04 | 0.90 | 0.81 – 1.01 | 0.92 | 0.61 – 1.39 |
| 5 | 0.91 | 0.80 – 1.03 | 0.84 | 0.77 – 0.93 | 0.89 | 0.54 – 1.46 |
| Unknown | 0.40 | 0.19 – 0.82 | 0.35 | 0.17 – 0.75 | 0.14 | 0.04 – 0.51 |
| *PY Seasonal Vacc* | | | | | | |
| No | 1.00 | | 1.00 | | | |
| Yes | 1.49 | 1.35 – 1.64 | 1.28 | 1.18 – 1.39 | 1.13 | 0.83 – 1.55 |
| *Prev Pandemic Vacc* | | | | | | |
| No | 1.00 | | 1.00 | | | |
| Yes | 1.19 | 1.08 – 1.30 | 1.11 | 1.04 – 1.18 | 1.17 | 0.94 – 1.46 |
| *PY ILIARI Cons* | | | | | | |
| 0 | 1.00 | | 1.00 | | | |
| 1 | 2.19 | 2.07 – 2.30 | 2.50 | 2.36 – 2.65 | 1.68 | 1.45 – 1.95 |
| 2+ | 4.14 | 3.84 – 4.48 | 5.30 | 4.87 – 5.77 | 2.89 | 2.18 – 3.84 |

**Table 7.6:** Hazard ratios of having ARI, ILIARI and ILI consultations found from a Cox proportional hazards regression model for season 2012–13. Vaccine effect can be derived from the hazard ratio for the vaccinated group, and vaccine effect for those aged 0–64 and 65+ can be found from the interaction terms.

ILI consultations. Moreover, those that received the seasonal vaccination within the previous year and those that had previous pandemic influenza vaccination were more at risk of having ARI/ILIARI consultations but the effects were not significant for ILI.

### All-Cause Deaths

HRs of all-cause death along with 95% confidence intervals, estimated from a Cox model for both seasons are given in Table 7.7. Note that for the deaths outcome, we used age groupings of 0–64 and 65+ in the model rather than the the smaller age bands as there were no deaths in some of the smaller age groupings. A further note is that in 2011–12, risk group was not included in the model as nobody aged 0–64 in a risk group died in that season.

In 2011–12, we found a positive VE for all ages of 24% (5% to 39%), and a slightly higher VE of 34% (16% to 48%) when looking at those aged 65+. However, vaccinated patients aged 0–64 had a higher HR of death with negative VE of −206% (−277% to −149%). This result is likely due to the fact that we examined VE against deaths from any cause and perhaps, those aged under 65 who got vaccinated could have been more likely to have underlying health issues (negative confounding) that we have not accounted for in the model. In the 2012–13 season, the VE results followed the same pattern as the previous season, but VE estimates were higher throughout: 46% (32% to 57%) for all ages; 52% (40% to 61%) for age 65+; and −14% (−77% to 26%) for age 0–64.

The other HRs in Table 7.7 show that age (divided into 0–64 and 65+) and being in a risk group had by far the largest effects on risk of all-cause death. Those that had the the seasonal vaccine in the previous year and those that had ILIARI consultations within the previous year also had a significant increase in risk of death. Additionally, patients living in areas of higher deprivation had a significant increase in risk of death, with differences found when comparing those from the most deprived quintile against those from deprivation quintiles 3–5. The effect of having previous pandemic vaccination differed in the two seasons – in 2012–13,

| | **2011–12** | | **2012–13** | |
|---|---|---|---|---|
| | **HR** | **95% CI** | **HR** | **95% CI** |
| *Vaccinated* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 0.76 | 0.61 – 0.95 | 0.54 | 0.43 – 0.68 |
| *Interactions* | | | | |
| Age 0–64 and Vacc | 3.06 | 2.49 – 3.77 | 1.14 | 0.74 – 1.77 |
| Age 65+ and Vacc | 0.66 | 0.52 – 0.84 | 0.48 | 0.39 – 0.60 |
| *Age Group* | | | | |
| 0–64 | 1.00 | | 1.00 | |
| 65+ | 21.20 | 16.64 – 27.00 | 36.13 | 31.68 – 41.20 |
| *Gender* | | | | |
| Female | 1.00 | | 1.00 | |
| Male | 1.10 | 0.98 – 1.25 | 1.12 | 0.98 – 1.28 |
| *Risk Group* | | | | |
| No | | | 1.00 | |
| Yes | | | 5.71 | 4.79 – 6.81 |
| *Dep Quintile* | | | | |
| Most Deprived | 1.00 | | 1.00 | |
| 2 | 1.05 | 0.83 – 1.33 | 0.98 | 0.77 – 1.24 |
| 3 | 0.74 | 0.58 – 0.94 | 0.74 | 0.62 – 0.89 |
| 4 | 0.62 | 0.49 – 0.78 | 0.69 | 0.57 – 0.84 |
| 5 | 0.65 | 0.53 – 0.79 | 0.77 | 0.61 – 0.96 |
| Unknown | 0.75 | 0.21 – 2.68 | 0.50 | 0.10 – 2.51 |
| *PY Seasonal Vacc* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 1.48 | 1.20 – 1.82 | 1.53 | 1.21 – 1.94 |
| *Prev Pandemic Vacc* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 1.15 | 0.96 – 1.37 | 1.22 | 1.06 – 1.40 |
| *PY ILIARI Cons* | | | | |
| 0 | 1.00 | | 1.00 | |
| 1 | 1.38 | 1.12 – 1.69 | 1.32 | 1.12 – 1.56 |
| 2+ | 2.32 | 1.77 – 3.04 | 1.91 | 1.65 – 2.21 |

**Table 7.7:** Hazard ratios of all-cause death from a Cox proportional hazards regression model. Vaccine effect can be derived from the hazard ratio for the vaccinated group, and vaccine effect for those aged 0–64 and 65+ can be found from the interaction terms.

those that had the pandemic vaccine were significantly more at risk of death but the effect was not significant in 2011–12. Gender had no significant effect on the risk of death.

### 7.3.5 Pre/During Influenza Season Analysis

With seemingly little influenza activity in 2011–12, we performed the pre/during influenza season analysis only on the 2012–13 season. This analysis added a time period factor (pre-influenza: 1st September to 31st October 2012; during influenza: 1st November 2012 to 31st March 2013) into the Cox model which was included as an interaction with vaccination and the results are shown in Table 7.8. The results show that there was no VE (low or negative VE with 95% CIs spanning zero) against any of the consultation types in the pre-influenza period and this is the expected result when little-to-no influenza was circulating. However, we found a positive VE against all-cause death for those aged 65+ in the pre-influenza period (VE: 24.2%; −15.5% to 50.3%), which is indicative of residual bias. VE estimates found for the during influenza period were very similar to those found using the whole study period.

We also obtained VE estimates adjusted for the amount of bias found in the pre-influenza period by assuming that the level of bias found in the pre-influenza period remained consistent during the influenza season for all outcomes. This produced positive VE estimates for all outcomes – against ARI, ILIARI and all-cause death, VE was around 20–40% (except ARI for those aged 65+ where VE was low); against ILI, VE was high with estimates above 70% for all ages and for those aged 65+, while VE was just under 50% for those aged 0–64. A notable point is that the adjustment increased VE for all estimates except against ARI for those aged 65+ and against all-cause death for all ages and in particular, those aged 65+.

#### Modelling an Unobserved Confounder

As the results from our pre/during influenza season analysis indicated that there was a VE against all-cause mortality in those aged 65+ in the pre-influenza period, we decided to model the effect of an unobserved confounder (smoking – yes/no) for

| | **Pre 2012–13** | | **During 2012–13** | | **Adjusted 2012–13** | |
|---|---|---|---|---|---|---|
| | **VE (%)** | **95% CI** | **VE (%)** | **95% CI** | **VE (%)** | **95% CI** |
| *ARI* | | | | | | |
| All Ages | −9.3 | −29.0 to 7.4 | 11.4 | 4.6 to 17.8 | 19.0 | 11.2 to 26.1 |
| 0–64 | −24.0 | −56.3 to 1.6 | 13.5 | 5.7 to 20.6 | 30.2 | 19.3 to 39.6 |
| 65+ | 1.1 | −21.8 to 19.6 | 8.1 | −0.5 to 16.0 | 7.1 | −4.6 to 17.5 |
| *ILIARI* | | | | | | |
| All Ages | −25.7 | −46.0 to −8.2 | 4.6 | −1.5 to 10.3 | 24.1 | 17.2 to 30.5 |
| 0–64 | −41.7 | −83.8 to −9.3 | 2.0 | −4.1 to 7.7 | 30.8 | 15.5 to 43.4 |
| 65+ | −16.4 | −49.0 to 9.1 | 7.7 | −1.8 to 16.2 | 20.7 | 7.8 to 31.7 |
| *ILI* | | | | | | |
| All Ages | −106.7 | −756.3 to 50.1 | 39.9 | 23.9 to 52.5 | 70.9 | 4.8 to 91.1 |
| 0–64 | −8.6 | −812.6 to 87.1 | 42.6 | 28.1 to 54.3 | 47.2 | −253.8 to 92.1 |
| 65+ | −210.2 | −1613.7 to 43.8 | 33.3 | −5.7 to 57.9 | 78.5 | 25.0 to 93.8 |
| *All-Cause Death* | | | | | | |
| All Ages | 18.9 | −29.7 to 49.4 | 47.3 | 32.8 to 58.7 | 35.0 | 18.4 to 48.2 |
| 0–64 | −54.5 | −1014.6 to 78.6 | −12.0 | −73.3 to 27.6 | 27.5 | −238.3 to 84.4 |
| 65+ | 24.2 | −15.5 to 50.3 | 53.2 | 40.8 to 62.9 | 38.2 | 25.4 to 48.8 |

**Table 7.8:** Vaccine effect estimates from the pre/during influenza season analysis for 2012–13. The pre-influenza period was defined as 1st September 2012 to 31st October 2012 and the during influenza period was the remainder of the study period. Vaccine effect estimates adjusted for the level of bias found in the pre-influenza period were also derived. This assumes that the level of bias in the pre-influenza period persists throughout the influenza season.

that outcome in that age group. Figure 7.7 illustrates that additionally including the effect of smoking (under the assumption that smoking increases the risk of all-cause mortality by 2.8 times) would reduce the observed VE estimate. For the pre-influenza estimates, we could theoretically say that this would remove some of the residual bias as we expect VE close to zero. In particular, increasing the prevalence of smoking among unvaccinated compared with vaccinated individuals reduces VE. However, even with a 95% prevalence of smoking among unvaccinated individuals and 20% prevalence of smoking among vaccinated individuals, a positive VE was still observed in the pre-influenza period, which suggests that further adjustment would still be needed to remove bias from the VE estimate. Our data lacked information on long-term health conditions for those aged 65+, and it is likely that details on those would help to remove more bias from the VE estimates against all-cause mortality.

### 7.3.6 Propensity Scores

To keep the analysis more concise, we only used the 2011–12 data to report the results related to the development of the PS model and covariate balance checks. However, we report the PS-adjusted VE estimates for both seasons. This does not lead to much loss of information as most of the results in the intermediate steps leading to PS-adjusted VE estimates in the 2012–13 season were almost identical to the 2011–12 season.

In order to model the propensity for individuals to get vaccinated, it was first necessary to choose an appropriate cut-off date for influenza vaccination. As was previously mentioned in Section 7.3.2, vaccine uptake largely ceased after the end of November and for this reason, we chose 30th November 2011 as the cut-off date for modelling each person's probability of getting vaccinated. Only 1.7% of patients ($n$ = 2,746) were vaccinated after the cut-off date. Hence, using this date for the model is appropriate as we want to capture the characteristics of those individuals that get vaccinated reasonably early rather than capturing the characteristics of those that get vaccinated late. The vaccine uptake pattern over time was similar in the 2012–13 season, and so we used 30th November 2012 as the

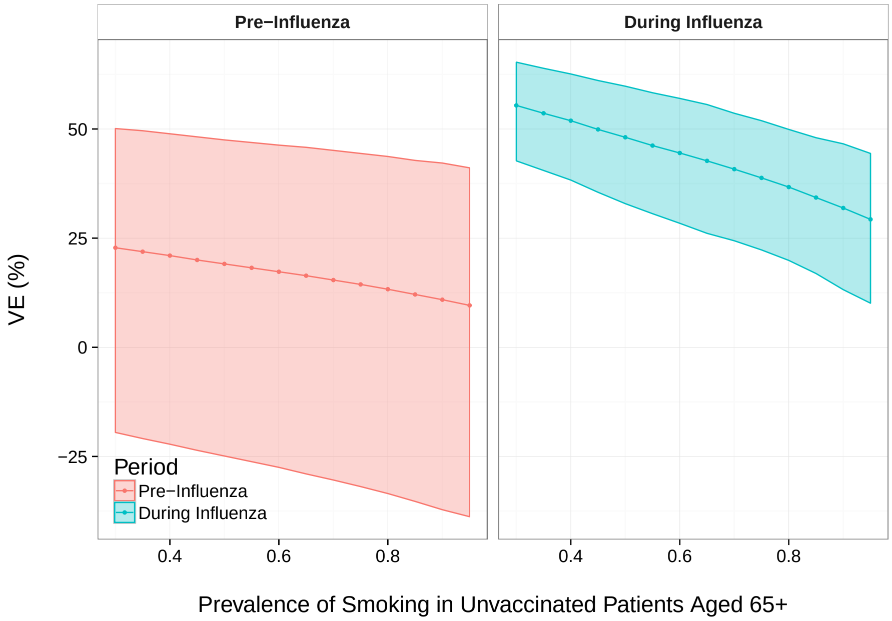


**Figure 7.7:** Vaccine effect estimates against all-cause death for those aged 65+ in the pre/during influenza season analysis after adjusting for an additional simulated unobserved confounder in a Cox proportional hazards model. The shaded areas give the 95% confidence interval around vaccine effect estimates. Previous research found an odds ratio of around 2.8 for all-cause death among smokers compared with non-smokers [203] and so the simulated confounder here mimicked this. The smoking prevalence among vaccinated individuals was fixed at 20% but the prevalence among unvaccinated patients was allowed to vary from 30% to 95% in 5% intervals.

cut-off date to construct the PS model for that season.

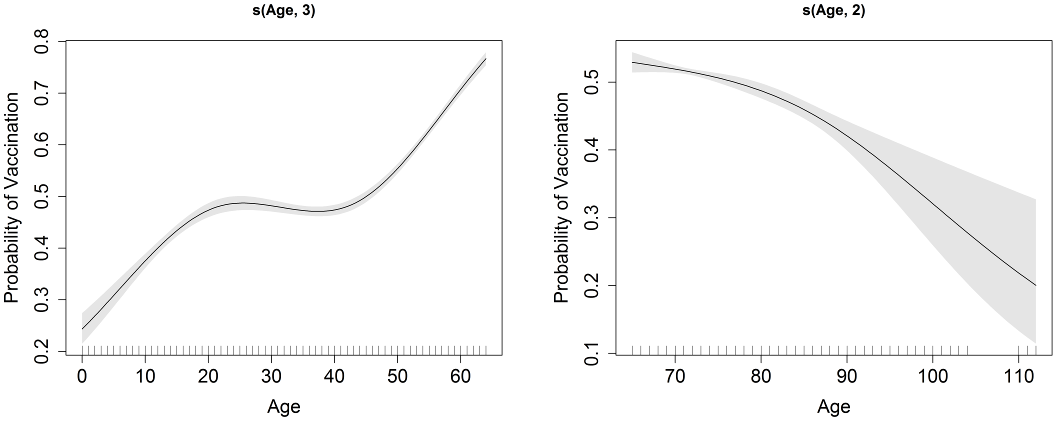
Table 7.9 gives the estimates from the GAMMs used to model propensity to vaccinate for patients aged 0–64 and 65+. To give us further reassurance that the cut-off date chosen was appropriate, we also ran the model using the end of the influenza season (31st March 2012) as the cut-off date for vaccination and found that this had only minimal impact on the estimated coefficients (results not shown).

The factors which appeared to be influential in getting vaccinated differed between the two age groups. For those aged under 65, individuals who had seasonal influenza vaccination within the previous year, those that had any previous pandemic influenza vaccination, and as expected, those that belonged to a risk group

| | **Age 0–64** | | **Age 65+** | |
|---|---|---|---|---|
| | **OR** | **95% CI** | **OR** | **95% CI** |
| (Intercept) | 0.01 | 0.01 – 0.02 | 0.27 | 0.23 – 0.31 |
| *Gender* | | | | |
| Female | 1.00 | | 1.00 | |
| Male | 0.88 | 0.84 – 0.92 | 1.01 | 0.94 – 1.08 |
| *Risk Group* | | | | |
| No | 1.00 | | | |
| Yes | 8.16 | 7.77 – 8.58 | | |
| *Dep Quintile* | | | | |
| Most Deprived | 1.00 | | 1.00 | |
| 2 | 1.09 | 1.01 – 1.18 | 1.10 | 0.98 – 1.23 |
| 3 | 1.05 | 0.97 – 1.14 | 1.13 | 1.00 – 1.28 |
| 4 | 0.94 | 0.86 – 1.02 | 1.20 | 1.06 – 1.37 |
| 5 | 1.07 | 0.98 – 1.17 | 1.40 | 1.21 – 1.60 |
| Unknown | 0.60 | 0.44 – 0.82 | 0.20 | 0.13 – 0.29 |
| *PY Seasonal Vacc* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 13.78 | 13.08 – 14.53 | 26.16 | 24.17 – 28.31 |
| *Prev Pandemic Vacc* | | | | |
| No | 1.00 | | 1.00 | |
| Yes | 2.77 | 2.62 – 2.93 | 1.77 | 1.61 – 1.95 |
| *PY ILIARI Cons* | | | | |
| 0 | 1.00 | | 1.00 | |
| 1 | 1.35 | 1.25 – 1.46 | 1.13 | 0.99 – 1.28 |
| 2+ | 1.63 | 1.48 – 1.80 | 0.95 | 0.81 – 1.12 |
| **Random Effects** | **SD** | | **SD** | |
| GP Practice | 0.37 | 0.32 – 0.44 | 0.24 | 0.15 – 0.36 |
| Residual | 0.87 | 0.87 – 0.88 | 1.00 | 0.99 – 1.01 |

**Table 7.9:** Estimates from the GAMM used to model propensity scores for vaccination. A binary response variable was created by considering those that did and did not receive seasonal influenza vaccination by 30th November 2011. Random intercepts were used for GP practice.

were much more likely to be vaccinated. In addition to this, those consulting with their GP for ILIARI at least once within the previous year were also more likely to be vaccinated while males were slightly less likely to get vaccinated than females. Deprivation had no significant impact on the likelihood of getting vaccinated; the only conclusion we could draw was that those with unknown deprivation were much less likely to get vaccinated which we found previously. Figure 7.8a illustrates the propensity to vaccinate by age for those aged 0–64, adjusted for the other factors included in the model. The propensity to vaccinate generally increased with age but in a non-linear fashion. In particular, the probability of vaccination was lowest for children and infants, but increased sharply for individuals after about 50 years of age.



**(a)** Ages 0–64

**(b)** Ages 65+

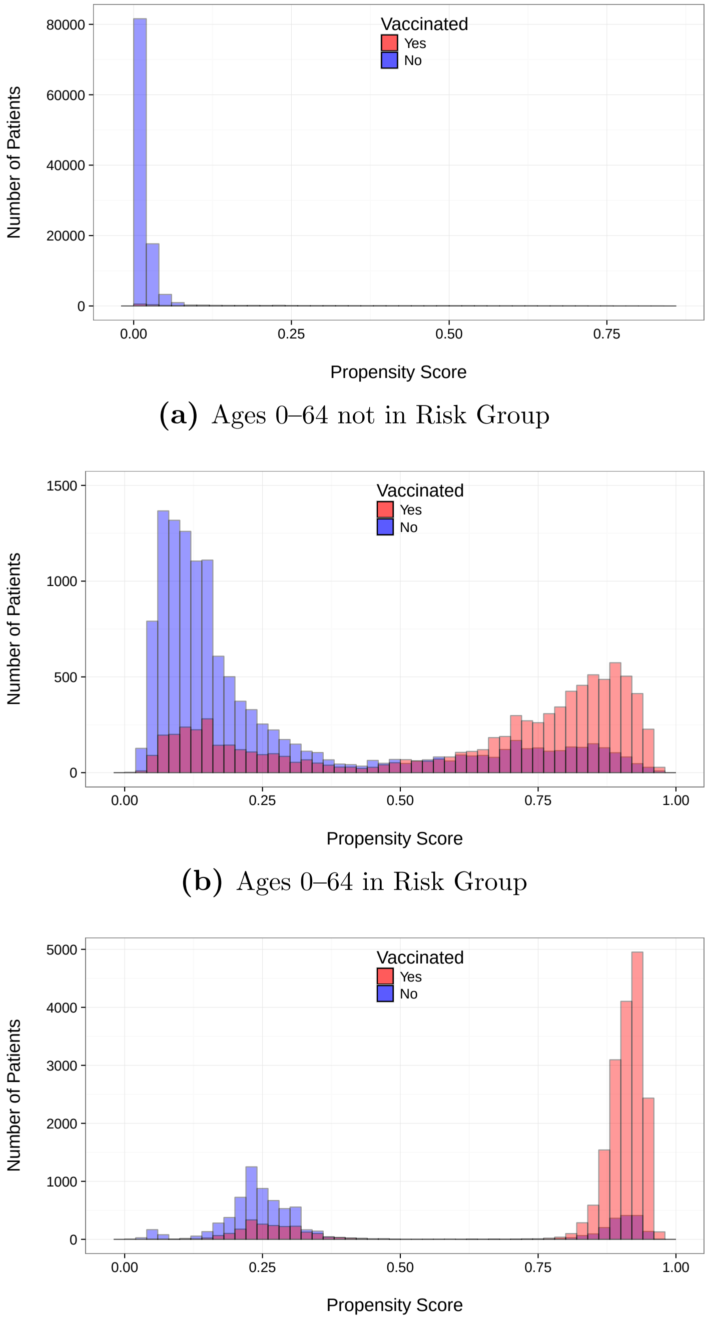
**Figure 7.8:** Plots of the smooth functions used for age in the GAMMs used to model propensity scores for vaccination. The shaded areas show the 95% confidence intervals (Bayesian credible intervals) and the rug plots at the bottom give an idea about the number of data points at each $x$-value.

For those aged 65+, gender and having consulted with a GP for ILIARI within the previous year did not play a significant role in getting vaccinated (see Table 7.9). Similarly to the model for those aged 0–64, having been vaccinated within the previous year for seasonal influenza and having previous vaccination for pandemic influenza were highly significant. A difference for this age group was that deprivation appeared to have an effect on vaccination, with individuals from deprivation quintiles 4–5 being more likely to get vaccinated compared with those from the most deprived quintile. In fact, for patients aged 65+, more deprivation generally lowered the probability of vaccination. The probability of vaccination decreased sharply beyond about the age of 85 (see Figure 7.8b). The explanation for this could be that the oldest and most frail individuals are less likely to get vaccinated as they may have serious health issues which would reduce the beneficial effect of the vaccination. From the standard deviation estimate for the random effect of GP practice (bottom of Table 7.9), we can see that there was more variation by GP practice in the model for those aged 0–64 than for those aged 65+, which is to be expected as all individuals aged 65+ are targeted for vaccination.

### Distributions of Propensity Scores

The histograms in Figure 7.9 illustrate the differing distributions of propensity scores between patients aged 0–64 not in a risk group, patients aged 0–64 in a risk group and those aged 65+. To create pairs of matched vaccinated-unvaccinated individuals, the first criteria is that they must have similar PS. Thus, the histograms for vaccinated and unvaccinated individuals have been overlaid on top of each other so that it is easy to discern PS ranges that are covered by vaccinees and non-vaccinees. These areas of overlap signal PS ranges where matches can be made based on PS. For those aged 0–64 not in a risk group, the overwhelming majority had extremely low propensity scores and very few got vaccinated (Figure 7.9a). Hence most of the unvaccinated individuals aged 0–64 will not be matched.

For those aged 0–64 in a risk group, there were also many people who got vaccinated despite having reasonably low propensity scores for vaccination which can be seen in Figure 7.9b. This most likely comprises individuals who require vaccination for occupational reasons while their demographic and health characteristics would suggest that they would not have a high propensity for vaccination. This includes, for example, carers and healthcare workers who require vaccination as they may need to have close contact with vulnerable patients but are otherwise, not at any increased risk of adverse outcomes from influenza.



**(a)** Ages 0–64 not in Risk Group

**(b)** Ages 0–64 in Risk Group

**(c)** Ages 65+

**Figure 7.9:** Histograms of propensity scores for vaccinated and unvaccinated individuals split by age groups 0–64 and 65+. The propensity scores give the probability of being vaccinated based on the observed background variables related to patients. The histograms for vaccinated and unvaccinated individuals have been overlaid on top of each other. The parts of bars that overlap are filled in a blue-red (violet) colour and this shows propensity score ranges where there are vaccinated and unvaccinated individuals with similar scores.

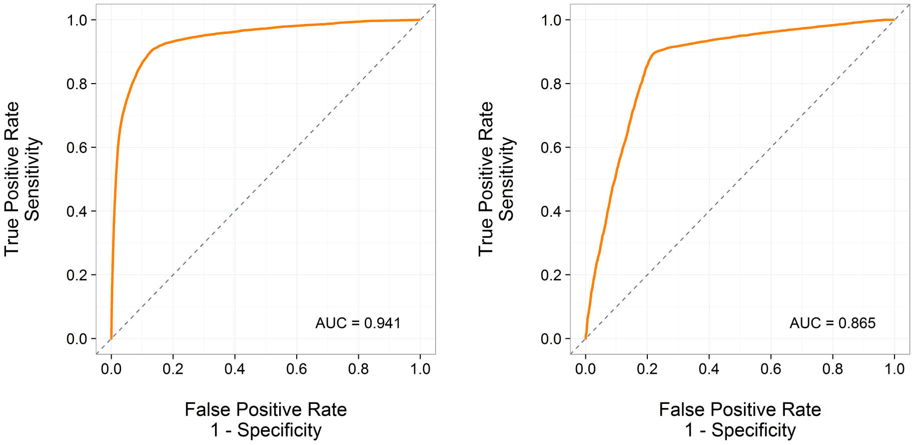
The PS model for patients aged 65+ did not distinguish between vaccinees and non-vaccinees as clearly as the model for those aged under 65. In particular, a group of individuals with a high propensity to get vaccinated did not actually get the vaccine (Figure 7.9c). These people may be a group of old and frail individuals who are less likely to get vaccinated (as we mentioned previously) but their background characteristics were similar to the majority of patients that have high propensity to get vaccinated.

#### ROC Curves

Figure 7.10 shows the receiver operating characteristic (ROC) curves for the propensity score models. The ROC curves reinforce some of the conclusions drawn earlier from the histograms in Figure 7.9. These are that the model excellently discriminated between vaccinees and non-vaccinees for the 0–64 age group. The area under the curve (AUC) which gives a measure of the discriminatory ability of the model reinforces this point with a very high value of 0.941. The high level of discrimination is largely due to the separation achieved by using risk group and variables associated with risk group status such as numbers of consultations in the previous year. The model for those aged 65+ also discriminated very well but not quite as clearly as for the 0–64 age group (AUC: 0.865). The slightly poorer performance could be due to not having a variable to split those aged 65+ in terms of risk severity like we could for those aged 0–64.

### 7.3.7 Propensity Score Matching

The histograms of propensity scores displayed previously in Figure 7.9 showed that there were many instances where it would not be possible to match vaccinated individuals to unvaccinated individuals based on propensity score. This is due to a lack of unvaccinated individuals with suitably similar scores within some ranges of propensity score. We can observe that there were not enough unvaccinated individuals aged 0–64 with propensity scores above approximately 0.5 to match to vaccinated individuals with similar scores. Similarly, there were not enough unvaccinated individuals to match against vaccinated individuals aged 65+ with propensity scores above around 0.7. The matched sample contained 13,948 indi-



**(a)** Ages 0–64 **(b)** Ages 65+

**Figure 7.10:** ROC curves for the propensity score models split by age groups 0–64 and 65+. The area under the curve (AUC) measures the discriminatory ability of the model. This tests how well the model correctly classifies those who get vaccinated and those who do not get vaccinated.

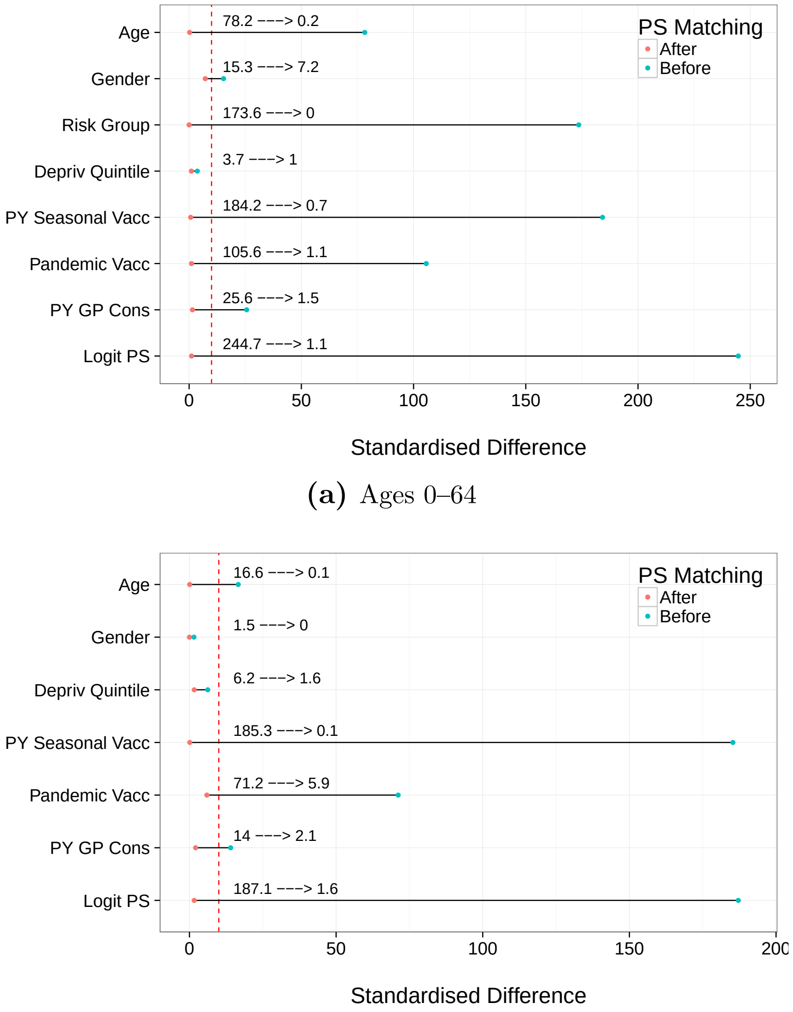
viduals aged 0–64 (10.8% of the 0–64 cohort; 6,974 pairs of matches), and 7,544 individuals aged 65+ (27.6% of the 65+ cohort; 3,772 pairs of matches).

### Covariate Balance After Matching

Figure 7.11 depicts the level of balance achieved in the covariates between vaccinated and unvaccinated individuals before and after matching by propensity scores. The standardised differences for all variables were below 10% after PS matching for those aged 0–64 which suggests that a good level of balance had been achieved in all of the background covariates (risk group was 0% after matching as this was a necessary condition in matching). In age group 65+ (Figure 7.11b), standardised differences for all variables were also well below 10% after matching.

### Vaccine Effect After Matching

The VE estimates for the two seasons after PS matching are given in Table 7.10. On the whole, VE estimates were lower than what was found before when using all patients in a Cox model, and this was generally true for both seasons. In addition to this, there were larger confidence intervals due to the large reductions in sam-



**(a)** Ages 0–64

**(b)** Ages 65+

**Figure 7.11:** Standardised differences for background covariates between vaccinated and unvaccinated individuals before and after propensity score matching. The red dashed line shows a standardised difference of 10% which can be considered as a cut-off between a negligible and non-negligible difference between the two groups of patients. The numbers above the horizontal lines for covariates show the values of standardised difference before (left side number) and after matching (right side number) for each covariate. For instance, for those aged 0-64, the standardised difference in age between vaccinated and unvaccinated individuals before matching was 78.2 and after matching, it was 0.2.
**Depriv**: Deprivation; **PY**: Previous year; **Vacc**: Vaccination;
**Cons**: Consultations; **PS**: Propensity score

ple size. For the ARI consultations outcome, all VE estimates were negative and ranged from around −9.6% to −13.1%. The ILIARI estimates were slightly higher than the ARI estimates, with the exception of when looking at only those aged 65+ where VE was −10.7% (95% CI: −30.1% to 5.8%). For the death outcome, we found a positive VE when looking at all ages (35.2%; 14.3% to 51%), but a large negative VE for those aged 0–64 (−242.9%; −695.7% to −47.7%). The highest VE was found for those aged 65+ when measured against all-cause mortality (51.7%; 33.7% to 64.8%).

| | **2011–12** | | **2012–13** | |
|---|---|---|---|---|
| | **VE (%)** | **95% CI** | **VE (%)** | **95% CI** |
| *ARI* | | | | |
| All Ages | −12.4 | −20.2 to −5.1 | −13.8 | −20.8 to −7.3 |
| Age 0–64 | −13.1 | −22.0 to −4.9 | −15.1 | −23.1 to −7.6 |
| Age 65+ | −9.6 | −26.8 to 5.3 | −9.5 | −24.3 to 3.4 |
| *ILIARI* | | | | |
| All Ages | −9.8 | −20.3 to −0.1 | −13.3 | −22.1 to −5.2 |
| Age 0–64 | −9.3 | −22.2 to 2.2 | −16.2 | −27.1 to −6.3 |
| Age 65+ | −10.7 | −30.1 to 5.8 | −6.8 | −22.4 to 6.9 |
| *ILI* | | | | |
| All Ages | | | 20.0 | −11.1 to 42.4 |
| Age 0–64 | | | 25.4 | −7.6 to 48.3 |
| Age 65+ | | | −7.7 | −129.1 to 49.4 |
| *All-Cause Death* | | | | |
| All Ages | 35.2 | 14.3 to 51.0 | 60.6 | 48.1 to 70.0 |
| Age 0–64 | −242.9 | −695.7 to −47.7 | 0.0 | −95.9 to 48.9 |
| Age 65+ | 51.7 | 33.7 to 64.8 | 66.9 | 54.9 to 75.6 |

**Table 7.10:** Vaccine effect estimates after propensity score matching. The matching method used was *nearest available Mahalanobis metric matching within calipers defined by the propensity score.* Matching was based on the logit propensity score, age and risk group (only for those aged 0–64).

In the 2012–13 season, the PS-matched VE estimates against the ARI and ILIARI outcomes were again all negative, but we obtained positive VE estimates for the ILI outcome except for those aged 65+. Against ILI consultations, VE for all ages was 20% (−11.1% to 42.4%); for those aged 0–64, VE was 25.4% (−7.6% to 48.3%); and for those aged 65+, VE was −7.7% (−129.1% to 49.4%). We also found positive VE estimates against the death outcome for all ages and those aged 65+; for those aged 0–64 there was no VE (0%). This can be contrasted to all other VE estimates against all-cause death found so far for those aged 0–64 where we had always estimated a large negative VE.

## 7.3.8 Propensity Score Stratification

Having created a model to estimate the propensity to vaccinate for each individual, systematic differences between individuals in the vaccinated and unvaccinated groups can also be controlled for by stratification on the PS. In an effort to remove some bias, the combined group of all vaccinated and unvaccinated individuals were used to create ten strata (deciles) with the cut-offs being determined by using the deciles of the PS. Essentially, covariates within these deciles should be more balanced between vaccinees and non-vaccinees than without stratification.

### Covariate Balance Within Strata

To check covariate balance, we again used standardised differences to compare vaccinated and unvaccinated individuals on covariates within strata. Note that some comparisons could not be made within strata as there were no people in the vaccinated group within some strata-variable combinations. Compared with PS matching, the level of covariate balance achieved is expected to be less here since all patients are kept in PS stratification. This is what we found as some strata still had fairly large standardised differences for a number of covariates. For instance, in age group 0–64, there were many gender imbalances within PS deciles and there were also imbalances for the deprivation and previous seasonal and pandemic vaccination variables (Table 7.11). Nevertheless, stratification of individuals into deciles using the PS achieved some degree of balance as compared with before (see Figure 7.11 for standardised difference values before any PS adjustment). Furthermore, adjustment for age group and risk group (risk group was included in the model for those aged 0–64 only) in the PS stratification model should help alleviate some concerns with regards to levels of covariate balance.

| | **Age** | **Gender** | **Dep Quint** | **Risk Group** | **PY Seas Vacc** | **Any Pand Vacc** | **PY ILIARI Cons** |
|---|---|---|---|---|---|---|---|
| *Age 0–64* | | | | | | | |
| Decile 1 | 18.2 | 2.0 | 35.9 | | | 5.4 | 23.4 |
| Decile 2 | 10.6 | 47.0 | 28.6 | | | 12.9 | 19.9 |
| Decile 3 | 5.4 | 44.3 | 31.7 | | | 15.4 | 18.9 |
| Decile 4 | 23.8 | 41.1 | 18.6 | | | 16.9 | 11.9 |
| Decile 5 | 11.8 | 31.2 | 17.8 | | | 3.3 | 23.1 |
| Decile 6 | 7.1 | 34.6 | 20.7 | | | 3.9 | 13.6 |
| Decile 7 | 14.0 | 18.6 | 25.4 | 1.8 | | 25.2 | 25.1 |
| Decile 8 | 21.6 | 2.2 | 33.8 | 30.6 | 2.9 | 32.9 | 9.9 |
| Decile 9 | 16.3 | 1.1 | 0.9 | 4.0 | 20.3 | 7.4 | 4.4 |
| Decile 10 | 19.1 | 12.1 | 3.9 | 11.6 | 58.6 | 24.9 | 2.3 |
| *Age 65+* | | | | | | | |
| Decile 1 | 1.7 | 5.0 | 54.1 | | | 4.2 | 20.9 |
| Decile 2 | 9.5 | 1.3 | | | | 6.8 | 12.9 |
| Decile 3 | 2.7 | 7.5 | 23.2 | | 3.3 | 42.2 | 6.1 |
| Decile 4 | 8.0 | 3.5 | 11.8 | | 20.7 | 13.1 | 11.6 |
| Decile 5 | 10.8 | 5.1 | | | | 11.7 | 12.9 |
| Decile 6 | 3.7 | 4.1 | | | | 23.5 | 4.1 |
| Decile 7 | 6.7 | 1.9 | | | | 17.9 | 11.6 |
| Decile 8 | 15.7 | 5.7 | | | | 1.9 | 18.9 |
| Decile 9 | 14.2 | 5.0 | | | | 15.7 | 10.9 |
| Decile 10 | 17.0 | 8.5 | | | | 9.4 | 8.7 |

**Table 7.11:** Covariate balance checks after propensity score stratification. All numbers are standardised differences (in %s) comparing the characteristics of vaccinated and unvaccinated patients within strata (deciles). Some comparisons could not be made due to there being zero counts.

### Vaccine Effect After Stratification

A logistic regression model with random intercepts for GP practice and vaccination status, age group and risk group along with PS deciles as covariates was used to estimate VE. The estimates are summarised in Table 7.12. Note that estimates for all ages were not obtained as we used separate PS models for age groups 0–64 and 65+, and unlike with PS matching, we cannot stratify within matched pairs.

VE estimates against all-cause death followed the same patterns as found previously with other methods in both seasons. Moreover, in the 2011–12 season, the VE estimates against ARI and ILIARI consultations were also similar but slightly lower to what was found previously. For the 2012–13 season the results were a little different to what we got from PS matching. Here, we found a positive VE estimate against ARI consultations when looking at those aged 65+ (5.8%, −6.2% to 16.4%).

| | **2011–12** | | **2012–13** | |
|---|---|---|---|---|
| | **VE (%)** | **95% CI** | **VE (%)** | **95% CI** |
| *ARI* | | | | |
| Age 0–64 | −20.2 | −30.2 to −11.0 | −11.3 | −19.2 to −3.9 |
| Age 65+ | −9.5 | −23.3 to 2.8 | 5.8 | −6.2 to 16.4 |
| *ILIARI* | | | | |
| Age 0–64 | −27.7 | −37.0 to −18.9 | −22.7 | −30.9 to −15.0 |
| Age 65+ | −18.6 | −39.3 to −1.0 | 2.3 | −13.0 to 15.5 |
| *ILI* | | | | |
| Age 0–64 | | | 37.7 | 22.2 to 50.1 |
| Age 65+ | | | 14.7 | −62.6 to 55.3 |
| *All-Cause Death* | | | | |
| Age 0–64 | −189.2 | −360.6 to −81.6 | −27.7 | −97.4 to 17.4 |
| Age 65+ | 34.1 | 17.5 to 47.4 | 52.3 | 41.3 to 61.2 |

**Table 7.12:** Vaccine effect estimates after propensity score stratification. Patients were split into strata according to the propensity score deciles they belonged to. The model used to derive vaccine effect adjusted for propensity score decile, age group and risk group (risk group was included in the model for those aged 0–64 only).

### Summary of VE Estimates Derived from Cox Models

At this point, we have used Cox proportional hazards models to derive numerous estimates of VE against consultations and all-cause mortality for all ages and split by those aged 0–64 and 65+. As we already have so many VE estimates and the remaining analyses in this chapter uses other types of models to estimate VE, it is useful to summarise the main results so far. There were insufficient ILI consultations to derive VE estimates against that outcome in 2011–12 and in general, VE was difficult to measure for that season as there was not much influenza in circulation. Hence, in Table 7.13, we have summarised VE estimates against the most specific consultation type (ILI) and all-cause mortality for the 2012–13 season only. This allows us to easily compare the estimates from the different methods that all used the Cox model in eventuality.

As we would expect, VE was similar for the Cox-adjusted method taken for the whole season and for the during influenza season analysis as the latter analysis was identical, with the exception of removing September and October 2012. Adjusting for levels of bias found in the pre-influenza season in VE estimates resulted in higher estimates except against all-cause death for those aged 65+ where the adjustment reduced VE. Using PS methods resulted in lower VE estimates against consultations for ILI, particularly using the PS matching method. Since PS matching considerably reduced the number of patients and changed the methodology to comparing effects within matched pairs, it is not surprising for this method to produce estimates that differed from the other methods which used the whole cohort. No VE was found against ILI in the pre-influenza period. However, as positive VE estimates were obtained against all-cause mortality for those aged 65+ in the pre-influenza period, this strongly indicates that there was insufficient adjustment for confounding factors in that age group.

## 7.3.9 Screening Method

For the screening method analysis, we examined consultations over the two month period with most influenza activity in the 2012–2013 season (according to numbers of ILI consultations), which were January and February 2013. This time period

| | 2012–13 | | | | | |
|---|---|---|---|---|---|---|
| | **All Ages** | | **0–64** | | **65+** | |
| **Method** | **VE (%)** | **95% CI** | **VE (%)** | **95% CI** | **VE (%)** | **95% CI** |
| *ILI* | | | | | | |
| Cox Adjusted | 39 | 23 to 51 | 42 | 28 to 54 | 31 | −8 to 56 |
| Cox Adjusted: Pre-Influenza | −107 | −756 to 50 | −8.6 | −813 to 87 | −210 | −1614 to 44 |
| Cox Adjusted: During Influenza | 40 | 24 to 53 | 43 | 28 to 54 | 33 | −6 to 58 |
| Cox Adjusted: During Minus Pre | 71 | 5 to 91 | 47 | −254 to 92 | 79 | 25 to 94 |
| Propensity Score Matching | 20 | −11 to 42 | 25 | −8 to 48 | −8 | −129 to 49 |
| Propensity Score Stratification | | | 38 | 22 to 50 | 15 | −63 to 55 |
| *All-Cause Death* | | | | | | |
| Cox Adjusted | 46 | 32 to 57 | −14 | −77 to 26 | 52 | 40 to 61 |
| Cox Adjusted: Pre-Influenza | 19 | −30 to 49 | −55 | −1015 to 79 | 24 | −16 to 50 |
| Cox Adjusted: During Influenza | 47 | 33 to 59 | −12 | −73 to 28 | 53 | 41 to 63 |
| Cox Adjusted: During Minus Pre | 35 | 18 to 48 | 28 | −238 to 84 | 38 | 25 to 49 |
| Propensity Score Matching | 61 | 48 to 70 | 0 | −96 to 49 | 67 | 55 to 76 |
| Propensity Score Stratification | | | −28 | −97 to 17 | 52 | 41 to 61 |

**Table 7.13:** Summary of vaccine effect results against ILI and all-cause death derived from Cox models for the 2012–13 season. Note that numbers have been rounded to the nearest percentage. **During Minus Pre** refers to the vaccine effect found after adjusting for the level of bias found in pre-influenza analyses.

for consultations was also chosen as we know from results in Section 7.3.2 that vaccine coverage did not change much after the end of November. Hence, vaccine coverage was estimated as at 31st December 2012 and as at 30th November 2012. Using vaccine coverage as at the end of November for examining consultations from January/February 2013 allows us to observe how sensitive VE estimates from the screening method are to underestimates of vaccine coverage.

The VE estimates obtained from the screening method are given in Table 7.14. The estimates were lower compared with those found previously in Cox models (especially for ARI and ILIARI consultations) and positive VE estimates were only produced from ILI consultations. Furthermore, VE was highly sensitive to slight changes in vaccine coverage – for instance, VE from ILI with vaccine coverage taken at the end of December was 16.3% (−224.3% to 78.4%) while if vaccine coverage was taken at the end of November, VE decreased to 7.5% (−235.7% to 74.5%).

| **Cons Type** | **Cons** | **Vacc Cons** | **% Vacc Cons** | **Vacc Cov** | **VE (%)** | **95% CI** |
|---|---|---|---|---|---|---|
| ARI | 10,826 | 3,689 | 34.1 | Dec 31 | −112.4 | −164.6 to −70.4 |
| ARI | 10,826 | 3,689 | 34.1 | Nov 30 | −133.1 | −189.9 to −87.5 |
| ILIARI | 8,771 | 2,552 | 29.1 | Dec 31 | −76.1 | −130.5 to −34.5 |
| ILIARI | 8,771 | 2,552 | 29.1 | Nov 30 | −93.5 | −152.0 to −48.5 |
| ILI | 623 | 115 | 18.5 | Dec 31 | 16.3 | −224.3 to 78.4 |
| ILI | 623 | 115 | 18.5 | Nov 30 | 7.5 | −235.7 to 74.5 |

**Table 7.14:** Vaccine effect estimates for 2012–13 from the screening method. The consultations were taken over the two months with most influenza activity (January and February 2013). Total consultations over the period for each type are shown on the table but for the screening method, these were broken down by GP practice, gender, age group and risk group. Vaccine coverage (broken down by the same groups) was estimated at the end of December 2012 and at the end of November 2012.
**Cons**: Consultations
**Vacc Cons**: Consultations from vaccinated patients
**Vacc Cov**: Vaccine coverage

### 7.3.10 Virology

With little influenza circulating in 2011–12, virology data was available for only 327 patients, but a much larger amount of virology data was available for the 2012–13 season ($n$ = 846). Table 7.15 summarises the demographic characteristics of patients that were tested and also gives the number of patients that tested positive for specific types of influenza as well as the number that tested negative.

More females were tested than males (54% vs 46% in 2011–12; 59% vs 41% in 2012–13) and nearly 90% of tests were on those aged under 65 in both seasons. A lot of tests were on infants aged 0–5 (almost 30% in 2011–12 and around 20% in 2012–13). These results fit with the patterns seen previously when examining those that consulted in the cohort (Figure 7.5). However, the percentage of tests which were on those aged 0–5 in 2011–12 is still very high which may indicate some bias in testing. There were also more tests from people living in urban and more deprived areas. In both seasons, around one fifth of people tested had the seasonal vaccine within the previous year, and around the same amount had previously been vaccinated for pandemic influenza. Almost three quarters of those tested in 2011–12 did not have any ILIARI consultations within the previous year; in 2012–13, the figure was slightly lower at around 70%.

The vast majority of virology tests were negative in 2011–12 and therefore, there was a low positivity rate of 6.7% ($n$ = 22). However, the positivity rate was much higher in 2012–13 (22.3%, $n$ = 189). Influenza A and AH3 were more common in 2011–12 but influenza B was most commonly detected in 2012–13.

#### Virology VE Estimates

Table 7.16 displays the virology test results by whether or not individuals were vaccinated at the time when they were tested (specifically vaccinated at least two weeks prior to testing to allow for the period of time where immunity builds). At the time of testing, most patients were unvaccinated (81%, $n$ = 265 in 2011–12; 80.5%, $n$ = 681 in 2012–13).

| | | **2011–12** | **2012–13** |
|---|---|---|---|
| **Variable** | **Group** | ***n* (%)** | ***n* (%)** |
| Gender | Female | 175 (53.52) | 495 (58.51) |
| | Male | 152 (46.48) | 351 (41.49) |
| Age Group | 0–5 | 98 (29.97) | 165 (19.50) |
| | 6–15 | 26 ( 7.95) | 72 ( 8.51) |
| | 16–44 | 98 (29.97) | 311 (36.76) |
| | 45–64 | 65 (19.88) | 207 (24.47) |
| | 65–74 | 25 ( 7.65) | 58 ( 6.86) |
| | 75+ | 15 ( 4.59) | 33 ( 3.90) |
| Risk Group | No | 246 (75.23) | 609 (71.99) |
| | Yes | 81 (24.77) | 237 (28.01) |
| Deprivation Quintile | Most Deprived | 84 (25.69) | 295 (34.87) |
| | 2 | 80 (24.46) | 227 (26.83) |
| | 3 | 61 (18.65) | 164 (19.39) |
| | 4 | 53 (16.21) | 96 (11.35) |
| | 5 | 47 (14.37) | 58 ( 6.86) |
| | Unknown | 2 ( 0.61) | 6 ( 0.71) |
| Urban/Rural | Urban | 232 (70.95) | 744 (87.94) |
| | Small Towns | 47 (14.37) | 59 ( 6.97) |
| | Rural | 46 (14.07) | 37 ( 4.37) |
| | Unknown | 2 ( 0.61) | 6 ( 0.71) |
| Prev Year Seasonal Vacc | No | 256 (78.29) | 679 (80.26) |
| | Yes | 71 (21.71) | 167 (19.74) |
| Any Pandemic Vacc | No | 256 (78.29) | 676 (79.91) |
| | Yes | 71 (21.71) | 170 (20.09) |
| Prev Year ILIARI Consultations | 0 | 243 (74.31) | 594 (70.21) |
| | 1 | 40 (12.23) | 137 (16.19) |
| | 2+ | 44 (13.46) | 115 (13.59) |
| Influenza Test Result | A | 9 ( 2.75) | 27 ( 3.19) |
| | AH3 | 10 ( 3.06) | 77 ( 9.10) |
| | B | 3 ( 0.92) | 85 (10.05) |
| | Negative | 305 (93.27) | 657 (77.66) |

**Table 7.15:** Counts and proportions for patients that had a virology test by various characteristics. The test results for patients are also shown and for patients that tested positive, the type of influenza they tested positive for is given.

| | Vaccinated | Unvaccinated | **Total** |
|---|---|---|---|
| *2011–12* | | | |
| Positive | 4 | 18 | 22 |
| Negative | 58 | 247 | 305 |
| **Total** | 62 | 265 | 327 |
| *2012–13* | | | |
| Positive | 31 | 158 | 189 |
| Negative | 134 | 523 | 657 |
| **Total** | 165 | 681 | 846 |

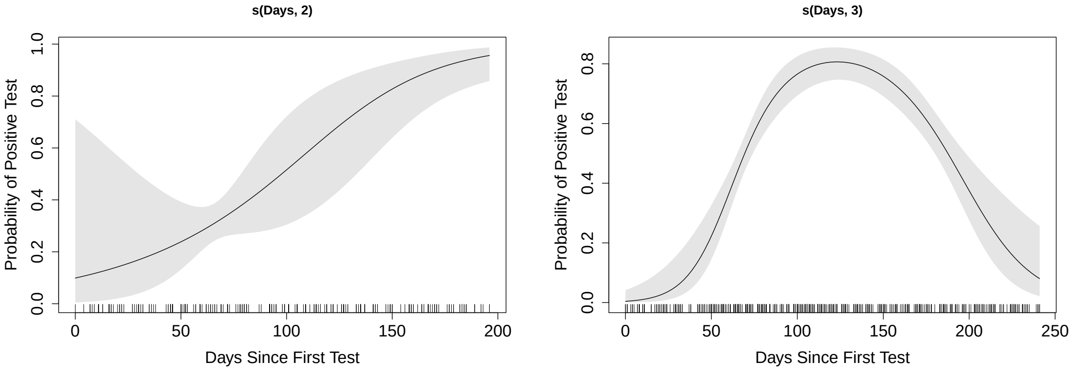
**Table 7.16:** Cross tabulation of virology test result against whether or not a patient was vaccinated before testing. The definition of a patient being vaccinated before testing used here meant that the patient was tested a minimum of two weeks after the date they got vaccinated.

The unadjusted ORs and corresponding 95% CIs were obtained by considering the counts in Table 7.16 as two separate 2×2 tables (one for each season). In 2011–12, the OR was 0.95 with 95% CI: 0.31 to 2.9. This corresponds to an unadjusted VE of 5% (−190% to 69%) and hence, that vaccinated and unvaccinated individuals were almost equally likely to have a positive test for influenza. Furthermore, the CI was wide due to the small quantity of virology data available for that season. The virology figures for 2012–13 suggest that the vaccine was more effective in that season. The OR was 0.77 with 95% CI: 0.5 to 1.19, corresponding to a VE of 23% (−19% to 50%).

### Generalised Additive Logistic Model

An adjusted VE estimate was obtained by using a GAM with days since first virology test modelled by a cubic spline function to adjust for changes in swab positivity over time. Initially, the model included seven covariates: days since first virology test, vaccination status, age group, gender, any GP consultations within the previous year, risk group status and deprivation quintile. However we wanted the most parsimonious model that includes age (retained since testing rates varied considerably by age group), the time factor and vaccination status. Only these three factors were kept in the final model as all other covariates were found to be not statistically significant – the results from the final model can be seen in Table 7.17.

The time factor was significant in both seasons and from Figure 7.12, we can see how the pattern of swab positivity changed over time. In 2011–12, the probability of a positive test generally increased over the time period. However, the pattern in 2012–13 was much different – the peak in probability of testing positive was reached around the middle of the period (coinciding with the time when there were most ILI consultations) and then fell steadily towards the end of the period. From the GAM, the VE estimate in 2011–12 was −88.2% (−769% to 59.3%), and in 2012–13, VE was 23.4% (−30.2% to 54.9%). In particular, the virology VE estimate for 2012–13 is in line with VE estimates found earlier against ILI consultations.



**(a)** 2011–12

**(b)** 2012–13

**Figure 7.12:** Plots of the smooth functions used to model days since first test in the GAM virology models. The shaded area shows the 95% confidence intervals (Bayesian credible intervals). A rug plot which gives an idea about the numbers of data points at each $x$-value is displayed at the bottom.

## 7.4 Discussion

Seasonal influenza vaccine effect estimates for seasons 2011–12 and 2012–13 in Scotland have been provided in this study. The main goals were not only to provide VE estimates for Scotland but also to better understand how estimates

| | 2011–12 | | |
|---|---|---|---|
| | **OR** | **95% CI** | ***p*** |
| (Intercept) | 0.03 | 0.01 – 0.10 | $< 0.001$ |
| *Vaccinated at Test* | | | |
| No | 1.00 | | |
| Yes | 1.88 | 0.41 – 8.69 | 0.418 |
| *Age Group* | | | |
| 0–5 | 1.00 | | |
| 6–15 | 3.44 | 0.84 – 14.09 | 0.086 |
| 16–44 | 0.85 | 0.27 – 2.70 | 0.781 |
| 45–64 | 0.14 | 0.01 – 1.43 | 0.097 |
| 65+ | 0.20 | 0.02 – 2.45 | 0.207 |
| | **edf (Ref df)** | $\chi^2$ | ***p*** |
| s(Days) | 2 (2) | 23.35 | $< 0.001$ |

| | 2012–13 | | |
|---|---|---|---|
| | **OR** | **95% CI** | ***p*** |
| (Intercept) | 0.14 | 0.09 – 0.23 | $< 0.001$ |
| *Vaccinated at Test* | | | |
| No | 1.00 | | |
| Yes | 0.77 | 0.45 – 1.30 | 0.325 |
| *Age Group* | | | |
| 0–5 | 1.00 | | |
| 6–15 | 4.36 | 2.17 – 8.76 | $< 0.001$ |
| 16–44 | 1.14 | 0.67 – 1.95 | 0.632 |
| 45–64 | 1.35 | 0.77 – 2.39 | 0.296 |
| 65+ | 1.08 | 0.44 – 2.66 | 0.864 |
| | **edf (Ref df)** | $\chi^2$ | ***p*** |
| s(Days) | 3 (3) | 70.66 | $< 0.001$ |

**Table 7.17:** Odds ratios from the GAM using virology data. Days since first virology test was modelled using a cubic spline function. The time trend, age group and vaccination status have been kept in the model but all other covariates tested were not statistically significant so were dropped.

of VE vary by methodologies applied, study designs employed, and clinical outcomes used to measure VE against. Estimations were made against five different clinical outcomes – ARI, ILIARI and ILI consultations, all-cause mortality and virology test results obtained from a subset of patients in the cohort who were swabbed. ARI consultations contained within them some of the clinical codes for ILI consultations and ILIARI was identical to ARI but excluded asthma-related consultations.

It is important to use a range of outcomes to measure VE as each has their advantages and disadvantages. In particular, these outcomes vary fairly substantially in their levels of specificity for measuring influenza VE. Certain types of consultations such as ARI/ILIARI are plentiful due to the wide variety of respiratory illnesses they cover. However these consultation types have low specificity for influenza and many individuals consulting with their GP for symptoms related to ARI/ILIARI will not actually have influenza, but will instead have other respiratory illnesses. On the other hand, ILI consultations are rarer as they are much more specific to influenza rather than other respiratory infections.

Another issue to be aware of is that correct recording of consultation types may vary by GP practice in Scotland, especially ILI consultations. It is conceivable that ILI consultations were under-recorded in some practices as those with influenza but presenting to their GP with milder symptoms may have been diagnosed as having ARI/ILIARI instead. The virology data provides the outcome with highest specificity and we very briefly explored the possibility of misclassified ILI consultations by using the virology data. We found that around 7% of patients in 2012–13 with laboratory-confirmed influenza had a consultation for ARI/ILIARI rather than ILI close to the time of testing. Although these consultations may not necessarily have been misclassified as the diagnoses would have been made based on symptoms, the positive test results give some indication that classifying those consultations as ILI may have been more appropriate.

It is seen as essential in a study of influenza VE to have at least some data on a highly specific outcome (preferably laboratory-confirmed influenza) [362] as VE results measured against that outcome can then be used to validate VE results from other outcomes. The drawback of virology data is that it is the most difficult type of data to obtain and consequently, VE estimates made using this outcome can often have wider confidence intervals due to the decreased availability of virology data. Out of the outcomes we considered, ILI consultations provided the best balance between specificity for influenza and having an outcome with sufficient data (at least in 2012–13) to obtain reasonably precise estimates of VE. Moreover, it must also be noted that to obtain fairly precise VE estimates specific to subgroups such as age or risk group, we also require sufficient outcome data within each of those subgroups as well as accurate recording of influenza vaccinations.

The severity of outcomes is also an important factor that has to be taken into account in discussions of VE. Having a death outcome is useful for giving some indication as to how effective the vaccine may be in reducing mortality. It is crucial to explore the effects of the vaccine against a severe outcome such as death or hospitalisation as one of the fundamental aims of influenza vaccination strategies is to protect high-risk individuals against severe outcomes [101]. However, we only had data on all-cause mortality which is a very non-specific outcome and it would have been preferable to have, for example, pneumonia and influenza related mortality instead, but this data was not available. In addition to this, it would have been interesting to measure VE against influenza related hospitalisations as this would likely have provided an outcome with more data points than all-cause mortality but this was also not available.

A larger quantity of data was available on all outcomes in 2012–13 compared with 2011–12. This was partly due to having more participating GP practices in the 2012–13 season. However the main reason for having less outcome data in 2011–12 was that very little influenza circulated in Scotland in that season [156]. As a consequence, one of the major issues with the 2011–12 season was that there were far too few ILI consultations to conduct analyses using that outcome. More generally, a lack of influenza circulating makes it difficult to ascertain VE using any of the methodologies and study designs we considered. Thus, more focus was given to examining VE in 2012–13.

We used a variety of different methods to estimate VE. When measuring

VE against consultations outcomes and all-cause mortality, these methods included Cox proportional hazards models with adjustment for confounding variables, propensity score adjustments (PS matching and PS stratification) and the screening method. For the virology outcome, we used a generalised additive logistic regression model to estimate VE. The estimates varied fairly substantially between some methods but in general, the choice of outcome affected estimates more. This finding is consistent with what has been found previously in a study of VE in Scotland for the 2009–10 pandemic influenza season [196]. While similar methods and outcomes were used in their study, the work here builds on what they found as we also include PS methods in our comparisons of VE.

The most severe outcome is all-cause death and it is of major interest to ascertain whether or not the vaccine is effective in preventing influenza-related deaths. Our results suggested this to be the case as we obtained positive VE estimates when looking at all ages, and estimates were larger when examining just those aged 65+. However, we obtained large negative VE estimates when looking at those aged under 65 for this outcome. This pattern of VE estimates was consistent for all methods used in both seasons. However, we showed caution against drawing the conclusion that the vaccine was highly effective at preventing death, especially in seniors. This was due to the findings of other researchers [182].

A study in the US found that only about 5% of excess deaths could be attributable to influenza [334]. However, our lowest estimate of VE against all-cause mortality for all ages in 2012–13 was 42% (95% CI: 32% to 57%) which was obtained from the Cox model with covariate adjustment. Hence, with consideration to the results from the US study, the positive VE estimates we got were much larger than we might expect for all-cause mortality. To further investigate the high VE estimate against all-cause death, we split the study period into a pre-influenza (September and October 2012) and during influenza period (November 2012 to March 2013) based on the argument that a VE should not be detected when little-to-no influenza is circulating. However, we still found a VE of 24% (95% CI: −16% to 50%) in the pre-influenza period and our belief was that this result was indicative of residual bias.

One of the major issues with the data for those aged 65+ was that we did not have risk information (underlying health issues) on them like we did for those aged 0–64. Hence a possible explanation was that the inflated estimates could be due to frailty selection where a group of non-vaccinated, very frail elderly people contributed to a large proportion of the deaths examined. To investigate the possible impact of an unobserved confounder on VE, we simulated data on smoking (either smokes or does not smoke), which is one of the variables contained in the minimum set of variables to have in influenza VE studies proposed by Valenciano et al. [362]. We based the simulated smoking data on previous research which showed that the OR of all-cause mortality in smokers compared with non-smokers was around 2.8 [203]. Furthermore, we varied the prevalence of smoking among unvaccinated individuals aged 65+ while keeping the prevalence among vaccinated individuals at 20% (close to the prevalence of smoking in the general Scottish population [329]). The results showed that even with a 95% prevalence of smoking in unvaccinated individuals, there was still a positive VE. Therefore, this gives further evidence on the necessity of having risk information and good frailty indicators on those aged 65+ to produce less biased VE estimates.

We also found negative VE estimates for those aged 0–64 against all-cause mortality which may suggest that there was some bias possibly caused by negative confounding in that age range. Specifically, a reason could be that those aged under 65 that got vaccinated were more likely to have underlying health issues and we have not fully accounted for this in our models. In a brief exploration of this issue (not covered in the main results), we calculated VE within risk group for those aged 0–64 in 2012–13. This produced a large negative VE for those not in a risk group while producing VE close to zero for those in a risk group. These results suggest that negative confounding was present, particularly for those not in a risk group and thus, there is also a need for variables that further adjust for health levels in that age group if measuring VE against all-cause death.

As part of the pre/during influenza season analysis, we calculated VE estimates adjusted for amounts of bias found against outcomes in the pre-influenza period. For this, the assumption was that levels of bias from the pre-influenza period would persist and remain consistent throughout the influenza season. This assumption suggests that the types of individuals that get vaccinated early cause bias in VE estimates in ways that are not accounted for by our covariates, and this *hidden bias* lingers on for the remainder of the study period. Making this adjustment generally produced much higher VE estimates against all outcomes except all-cause death where it reduced VE for those aged 65+; this exception is because of the positive VE estimates found against all-cause death in the pre-influenza period. The strength of this adjustment is that it allows VE to be adjusted for factors that are possibly difficult to measure. However, the assumption of a constant bias being able to be measured in this way and then being propagated throughout the influenza season may be questionable and is difficult to assess.

From the consultations outcomes, the VE estimates were always lowest when measured against ARI and ILIARI, regardless of the method employed. Excluding VE estimates found when adjusting for levels of bias in the pre-influenza season, VE estimates against ILIARI were either negative or very low in either season. The highest estimate was 7% (−3% to 16%) for those aged 65+ in 2012–13. Similar results were obtained when using the ARI outcome, but these were mostly slightly higher than the ILIARI estimates.

On the whole, we would expect the ILIARI and ARI estimates to be reasonably similar as they differ only on their inclusion of asthma-related consultations. Since influenza can induce additional problems for asthma-sufferers, it is not entirely unexpected that the ILIARI estimates were lower than the ARI estimates. In essence, we could expect some reductions in consultations for vaccinated individuals with asthma compared with unvaccinated individuals with asthma. This, in turn, causes some increase in the estimated VE.

Using the ILI outcome resulted in the highest estimates of VE out of the three types of consultations. The estimates were predominantly between 20–50% in the 2012–13 season. As ILI consultations are the most specific to influenza, we should expect the largest reductions in this type of consultation following vaccinations. Our VE results against ILI in 2012–13 are reasonably similar to estimates from other studies in Europe. A cohort study in France using volunteers registered on an online surveillance system for ILI estimated VE to be 49% (20% to 67%) for the overall population [85]. Mid-season estimates from Spain found VE to be 32% (15% to 46%) [53].

Using a highly specific outcome such as virology data should result in the most reliable estimates of VE but we still have to be wary around possible biases in testing. Looking at the virology data more closely indicated that higher proportions of tests came from people within certain age groups. For instance, in 2011–12 almost 30% of swab results came from those aged 0–5 and this fitted with the age group who had most consultations (with respect to the amount of patients aged 0–5 in the cohort). However, the proportion of tests on those aged 0–5 was still high which suggests that there may have been some bias in swabbing patterns towards certain age groups. Moreover, the largest proportion of positive test results within an age group came from those aged 6–15. From these findings, we felt that it was necessary to keep age in models used to estimate VE against the virology outcome.

As mentioned previously, a lack of influenza circulating in 2011–12 meant there was not much virology data to conduct analyses in 2011–12. Swab results were only available from 327 individuals in 2011–12. but in 2012–13 we had virology data from 846 individuals. The negative estimates we obtained from the virology outcome for 2011–12 were lower than UK-wide estimates obtained using a test-negative case-control study design [285]. They obtained a VE estimate of 23% (−10% to 47%) which was still fairly low. In England and Wales, influenza circulation hit a clear peak in 2011–12 while in Scotland, circulation was consistently low [156] which helps to explain this result. Furthermore, our estimates differ from other studies from around Europe. A study in Spain using virology data estimated VE to be 55% (3% to 79%) [191]. We do, however, have to note that they used a targeted population. A multicentre study using data from several hospitals around Europe provided a pooled VE estimate against hospitalisation with laboratory-confirmed influenza of 24.9% (1.8% to 44.6%) [311].

In the 2012–13 season, there did not seem to be as much of a tendency to swab individuals within certain age groups. The virology data gave VE estimates which were reasonably consistent with what was found when using ILI consultations – VE was 23% (−30% to 55%). Our estimates using this kind of data are again, lower than estimates found from studies in the UK and Europe. In the UK, a mid-season analysis found an overall VE estimate of 51% (27% to 68%) against laboratory-confirmed influenza [241]. A mid-season estimate is also available from Spain where VE against laboratory-confirmed influenza was found to be much higher at 86% (45% to 96%) [53].

We attempted to reduce possible bias in VE estimates via the construction of a PS model, which showed good predictive capability for distinguishing between vaccinated and non-vaccinated individuals. In the PS matched sample, we checked the level of covariate balance in the two groups using standardised differences [14]. Overall, a high level of balance on all background covariates was achieved which was a good indication that the matching procedure worked well. However, PS matching produced VE estimates that differed most from those produced by other methods. This can be expected as the method uses only a matched sample of the cohort while all the other methods we employed involved the whole cohort (with the exception of when using virology data). With PS stratification, VE estimates were more similar to what was produced from other methods. However, the level of balance within deciles of the PS score would ideally have been better. Generally, the PS model could be improved by having better frailty indicators which are important in determining who is more likely to get vaccinated.

It has to be noted that for the PS model and the Cox regression models used to find VE estimates, there was missing data on deprivation which was included as a separate category in those models. However, Vach and Blettner [361] showed that including individuals with missing data in models by creating an additional category results in a biased estimation. Hence, a better approach would have been to firstly investigate if the data was missing at random and then if that was the case, the data could be appropriately imputed to provide improved parameter estimates [317, 345].

The screening method produced estimates of VE which were in line with estimates from other estimates when measured against ILI consultations. However, when using ARI and ILIARI consultations, the screening method produced large negative estimates. The VE estimates from the screening method can be sensitive to the vaccine coverage [67] and it is difficult to appropriately align vaccination coverage with a time period for consultations, particularly as vaccine coverage is changing throughout the season. We limited the impact of a non-constant vaccination coverage by choosing to estimate vaccine coverage at a time when vaccine uptake had largely finished and using consultations from the peak influenza period in 2012–13.

Screening method VE estimates are dependent on the ratio of consultations (or events) from vaccinated individuals compared to unvaccinated individuals. When there are relatively many consultations from vaccinated individuals coupled with lower vaccine coverage, the screening method produces low estimates of VE. The reason that lower vaccine coverage affects this is because the relatively many consultations would be coming from a smaller pool of vaccinated individuals.

The best scenario for the screening method would be for the uptake of the seasonal vaccination to be finished before the influenza season starts, and then to look at aggregate numbers of events from vaccinated and unvaccinated during the peak of the influenza season. However, if we wish to obtain timely VE estimates for surveillance, this is not possible as we cannot be certain of when the peak will occur each season, and sometimes the uptake of the vaccine will be happening during times when influenza is already circulating. Hence a problem with the screening method that can occur frequently is to look at consultations at times when influenza is not circulating.

This study demonstrates how complicated measuring influenza VE can be and highlights a number of the issues involved when carrying out a study of influenza VE using observational study designs. We have shown that VE estimates vary most widely with the use of different clinical outcomes but the choice of methodology used to estimate VE can also have a substantial impact on VE estimates. The key point is that having a variety of VE estimates to compare and contrast against is beneficial as the weight of evidence can guide us to a consensus on the overall VE. Very low estimates of VE were obtained for the 2011–12 season, which was most likely due to the low amounts of influenza that were circulating in that season. Although 2012–13 was also a fairly mild influenza season, the amount of the virus circulating was much higher in comparison to the previous season [157], and we obtained higher VE estimates for that season. Although we attempted to adjust for potential bias using various methodologies, we cannot rule out residual confounding from unmeasured sources. One of the main ways that the study could be improved in this regard is by having better frailty indicators, particularly for those aged 65+. This would be of benefit to the PS model as well as aiding in the principal goal of producing more reliable VE estimates.

# 7.A VE Appendix

## Read Codes

| Read Code | Description | Count |
|---|---|---:|
| 65E.. | Influenza vaccination | 33,820 |
| 65E2. | Influenza vaccination given by other healthcare provider | 499 |
| ZV048 | [V]Influenza vaccination | 139 |
| 65E9. | PANDEMRIX - first influenza A (H1N1v) 2009 vaccination given | 65 |
| 65EB. | PANDEMRIX - 1st flu A (H1N1v) 2009 vac by othr hlth provider | 57 |
| 65E0. | First pandemic influenza vaccination | 28 |
| 65E3. | 1st pandemic influenza vacc give by other healthcare providr | 9 |
| 65EC. | PANDEMRIX - 2nd flu A (H1N1v) 2009 vac by othr hlth provider | 2 |
| 65E4. | 2nd pandemic influenza vacc give by other healthcare providr | 1 |
| 65E7. | CELVAPAN - 1st flu A (H1N1v) 2009 vacc by othr hlth provider | 1 |
| 65EA. | PANDEMRIX - second influenza A (H1N1v) 2009 vaccination give | 1 |

**Table 7.18:** Read Codes with descriptions for all influenza vaccinations (seasonal and pandemic) received by patients in the cohort between 1st July 2011 and 31st March 2012. The study start date for the 2011–12 season was 1st September 2011 and the end date was 31st March 2012 but individuals were also considered vaccinated for the season if they received vaccination up to 2 months prior to the study start date. Note that patients can have more than one vaccination.

| **Read Code** | **Description** | **Count** |
|---|---|---|
| 65ED. | Seasonal influenza vaccination | 37,070 |
| 65E.. | Influenza vaccination | 2,641 |
| 65E9. | PANDEMRIX - first influenza A (H1N1v) 2009 vaccination given | 72 |
| 65EB. | PANDEMRIX - 1st flu A (H1N1v) 2009 vac by othr hlth provider | 55 |
| ZV048 | [V]Influenza vaccination | 48 |
| 65E2. | Influenza vaccination given by other healthcare provider | 27 |
| 65E0. | First pandemic influenza vaccination | 11 |
| 65E3. | 1st pandemic influenza vacc give by other healthcare providr | 8 |
| 65ED0 | Seasonal influenza vaccination given by pharmacist | 3 |
| 65E7. | CELVAPAN - 1st flu A (H1N1v) 2009 vacc by othr hlth provider | 2 |
| 65EC. | PANDEMRIX - 2nd flu A (H1N1v) 2009 vac by othr hlth provider | 2 |
| ZV048-1 | [V]Flu - influenza vaccination | 2 |
| 65E.. | Query Influenza vaccination | 1 |
| 65E5. | CELVAPAN - first influenza A (H1N1v) 2009 vaccination given | 1 |
| 65E6. | CELVAPAN - second influenza A (H1N1v) 2009 vaccination given | 1 |

**Table 7.19:** Read Codes with descriptions for all influenza vaccinations (seasonal and pandemic) received by patients in the cohort between 1st July 2012 and 31st March 2013. The study start date for the 2012–13 season was 1st September 2012 and the end date was 31st March 2013 but individuals were also considered vaccinated for the season if they received vaccination up to 2 months prior to the study start date. Note that patients can have more than one vaccination.

<table>
<tr><th>Consultation Type</th><th>Readcodes</th></tr>
<tr><td>ILI</td><td>G5203 | H2... | H27.. | H270. | H2700 | H2701 | H270z | H271. | H2710 | H2711 | H271z | H27y. | H27y0 | H27y1 | H27yz | H27z. | H2y.. | H2z.. | Hyu05 | Hyu06 | Hyu07</td></tr>
<tr><td>ILIARI</td><td>H0... | H05.. | H05z. | H06.. | H06z. | H07.. | H0y.. | H22.. | H22y. | H23.. | H260. | H3... | Hyu1. | Hyu10 | H04.. | H05y. | H0z.. | H22z. | H23z. | H25.. | H26.. | Hyu0. | Hyu11 | H00.. | H01.. | H010. | H011. | H012. | H014. | H01y. | H01z. | H02.. | H022. | H023. | H024. | H02z. | H03.. | H030. | H031. | H035. | H036. | H037. | H03z. | H040. | H0400 | H0402 | H0403 | H040w | H040z | H041. | H0410 | H041z | H042. | H042z | H043. | H0432 | H043z | H044. | H04z. | H050. | H051. | H052. | H054. | H055. | H060. | H0603 | H0604 | H0605 | H0606 | H060C | H060w | H060x | H060z | H061. | H0612 | H0615 | H061z | H062. | H06z0 | H06z1 | H06z2 | H20.. | H201. | H20y. | H20z. | H21.. | H220. | H223. | H224. | H22y2 | H22yz | H231. | H24y2 | H2600 | H261. | H262. | H263. | H27z. | H28.. | H2A.. | H30.. | H300. | H301. | H302. | H30z. | H3y0. | H3y1. | Hyu0A | Hyu0H</td></tr>
<tr><td>Asthma</td><td>H33.. | H330. | H3300 | H3301 | H330z | H331. | H3310 | H3311 | H331z | H332. | H333. | H334. | H33z. | H33z0 | H33z1 | H33z2 | H33zz</td></tr>
</table>

**Table 7.20:** Read Codes for different GP consultation types used in the study. Abbreviations used are:
**ILI**: influenza-like illness
**ILIARI**: influenza-like illness or acute respiratory infection excluding asthma
Note that acute respiratory infections (ARI) were also used in the study and were made up of the same Read Codes as ILIARI but ARI included consultations for asthma.

# Chapter 8

# Conclusions

## 8.1 Summary and Discussion

The work we have undertaken in this thesis has considered some of the ways in which statistical methods can be applied to assist in the management of disease outbreaks that are vaccine preventable. We have focussed our attention on infectious disease outbreaks in Scotland where the National Health Service (NHS) holds a number of high quality healthcare databases, making it an ideal candidate country to facilitate the achievement of our goals. In sum, we have applied and developed methods to aid in the surveillance and management of two infectious diseases which are to different extents, vaccine preventable. These are tuberculosis (TB) and influenza (pandemic and seasonal strains), with the bulk of the work being on influenza. Specifically, one chapter was on TB, three chapters were on pandemic influenza (with one chapter that reviewed methods of estimating reproductive numbers with a view to applying those methods to pandemic influenza), and two chapters were on seasonal influenza and vaccine effect.

A prevailing theme throughout was the use of statistical methods to correct for issues of data quality and quantity which frequently appear when analysing disease outbreak data. This happens because national routine surveillance systems may often not be designed to collect the types of data that are required for analysis of outbreaks. Hence, these statistical methods are essential for producing improved parameter estimates in the face of imperfect data.

The first topic we considered looked at how statistical methods could be applied to genetic typing data to detect clusters of potentially linked TB cases. Here, the principal contribution was of practical importance as a bespoke system was created to detect potential genetic strain clusters of TB. At the time of writing, the system is being run regularly by Health Protection Scotland (HPS) and a database of all known strain clusters involving cases living in Scotland is being maintained. This detection system was necessary as the strain typing procedure is imperfect and often does not always capture the number of repeats in all 24 loci to form a complete MIRU (Mycobacterial interspersed repetitive units) profile. Hence, clustering based only on complete data is inadequate as this would fail to detect a large number of links between cases; our system rectified this issue by alerting investigators to potential links involving cases with incomplete genetic typing data (up to a maximum of two missing digits in MIRU profiles).

The system is important for two main reasons. Firstly, by notifying epidemiologists of many more potential links between cases, this ensures that less clusters fail to be detected. With active follow up of cases in identified clusters, this can help to prevent further spread of the disease. Furthermore, finding more linked cases can improve our understanding of how TB is being transmitted between persons in Scotland. Secondly, repeating the strain typing procedure in every instance where the number of repeats in all 24 loci is not captured would be prohibitively expensive, time-consuming and may not even be possible at all in some instances. However, the system helps in rapidly finding *close* matches between cases and thus, it helps in directing attention to instances where the strain typing procedure can be be performed again with a reasonable chance of finding potential strain clusters.

When the system finds that a case can belong to multiple potential clusters, we assigned probabilities that they belong to each potential cluster in two different ways. However, these probability measures were fairly crude. For example, one measure looked at the frequencies of numbers of repeats in each locus over all MIRU profiles and assumed independence in joint-missing loci. We examined where missing loci most often appeared and found that this happened far more often in loci 23 and 24. Therefore the independence assumption is likely to be invalid and better probability estimates could be attained by accounting for the positions of the missing loci. Due to the limitations of the probability measures, our recommendation is that in practice, the probability measures should only be used for preliminary investigation purposes and that qualitative information on cases should take precedence in follow up cluster investigations.

We then analysed data from the 2009 pandemic influenza outbreak in Scotland. More precisely, we got access to two novel datasets that were linked; one containing early cases of pandemic influenza in Scotland who answered detailed questionnaires and the other containing contacts of those cases. Since the data was collected in haste during the height of the pandemic, it was challenging to analyse due to the quantities of missing data and the amount of data that was entered in free-text form. In future outbreaks, additional thought should be given, prior to data collection, on how to record case data in a structured format to make rapid analysis more straightforward. Moreover, one of the reasons for the amount of missing data in the dataset could have been because investigators attempted to capture information on too many specific symptoms from cases – the questionnaire looked for answers on 21 symptoms which we then analysed as a set of 18 symptoms by grouping some similar symptoms together. Perhaps a simpler approach that categorised symptoms into larger groupings from the beginning may have been more beneficial in the sense that cases may have been more willing to respond to all questions. Previous work on shortening questionnaires in a US census has shown that simplification can indeed improve response rates [89].

To deal with the missing data, we made reasonable assumptions on the mechanism behind how missing data on symptoms was generated and this allowed us to use all cases in the dataset for analysis except those cases that did not provide any information on symptoms (over 92% of cases). Exploring the data in detail allowed us to gain a more vivid picture of how the disease affected the Scottish population. For instance, we looked at differences in the symptoms reported by gender and by those in different age groups using a series of logistic regression models, with adjustments made for multiple testing. This showed that symptoms reported did not really differ by gender but did by age group with infants and school children reporting different symptoms from adults. Furthermore, looking at the deprivation category of where cases lived showed no apparent indication that deprivation affected severity of illness or access to treatment. In general, pandemic influenza caused fairly mild illness on the majority of individuals and the types of illness reported by cases was not dissimilar to seasonal influenza.

The simple Bonferroni correction [35] was used to account for multiple testing. This method was used as we wished to limit the number of spurious statistically significant differences that would be found. Specifically, it would limit the number of false positive results (type one errors). However, we note that when conducting a large number of tests, the Bonferroni correction can be overly conservative and consequently increases the chance of false negative results. Other adjustment methods for multiple testing that do not reduce statistical power to such an extent could have been used instead. Two alternative methods include the Holm-Bonferroni method [173] and the Benjamini-Hochberg procedure [31], but many more methods exist [30]. These two procedures control the false discovery rate (FDR) and are more forgiving of type one errors compared with the Bonferroni correction. Using these may have yielded more interesting results that could have been further scrutinised for verification.

A more thorough understanding around case-contact mixing patterns during the explosive phase of an epidemic would help to prevent disease transmission. We wanted to achieve this using the data on contacts of cases but our analysis was limited. We were able to ascertain that children reported more contacts than adults and that most reported contacts were from areas with similar deprivation status to the case. However, the data was likely to have suffered from recall bias [71]. The fact that the majority of reported contacts were household contacts is perhaps an indication that the cases were far more likely to remember interactions with people they live with rather than transient contacts.

After looking at the pandemic influenza outbreak from a descriptive and analytical epidemiology perspective, we considered disease outbreak surveillance by estimating reproductive numbers along with the serial interval distribution. Both are important parameters, but reproductive numbers are of particular interest as they can neatly summarise how fast an epidemic is growing at a given time. Methods that can track the effective reproductive number over time with a relatively short time lag can be employed as surveillance systems for disease outbreaks. In addition to this, they can be used not only to alert authorities to situations when interventions are needed, but can also be used to monitor the impact of interventions.

For our work on this, we again looked at data on early case reports from pandemic influenza in Scotland and estimated the parameter using existing methods as well as by developing a new method that utilises spatial data on cases. Due to data issues related to case ascertainment during certain time periods and reporting of dates of onset of symptoms, we worked with three datasets: one dataset using the original data only; one dataset that imputed dates of onset of symptoms using the reporting delay distribution for cases that only had dates of illness reported; and one dataset that created additional case reports to correct for under-reporting on days when laboratory testing was suspended. Once the datasets were established, we estimated the reproductive number for pandemic influenza for all three datasets using a couple of established methods that make use of only the serial interval distribution and the dates of onset of symptoms.

The first of these methods, used the moment generating function of the serial interval distribution and assumes exponential growth of new cases [371] which is reasonable in the early phase; the second method can be thought of as a "back calculation" method [372] that also uses the serial interval distribution to infer infection probabilities for each case and then sums these probabilities to find reproductive numbers on a given day. All of the estimates, averaged over the early phase, were under 1.5, indicating that the disease was growing at a fairly manageable rate and this finding was consistent with what was being found regarding pandemic influenza from other countries. However, one peak in infection was found in May 2009 due to the detection of a cluster of cases that all travelled together on the same bus.

As postcode of residence was routinely collected for cases, we investigated combining data on the distance between where cases lived together with the time difference of when they first had onset of symptoms to see if this would result in better estimates of the reproductive number over time. The reason for developing this method was motivated by the argument that given the observed times of infection and outbreak size, not all *infection trees* are equally likely [140] which is an assumption in the method of Wallinga and Teunis [372]. For instance, this assumption is violated when there is clear variation in contact rates between different types of individuals. The assumption we made with the spatial method was that cases living nearer to each other would be more likely to be part of the same transmission chain. For the estimation, we included the spatial component together with the temporal component in a simple multiplicative fashion. If two cases were from the same household, we further increased the probability that one of them acquired infection from the other.

By developing an epidemic simulation model to simulate epidemics similar to the early phase of the pandemic influenza outbreak in Scotland, we were able to show that the spatial method could capture changes in the reproductive number over time more closely than using the temporal component alone, which produces more stable estimates over time. However, the limitation here was that we had knowledge of the parameters used in epidemic simulations and this helped us to choose parameters for the spatial method that would return favourable results. In practice it is difficult to provide strong evidence that the spatial method produces much better estimates without knowing the true transmission chain.

The last part of the thesis was dedicated to statistical methods for estimating vaccine effect (VE). This can be a complex affair in the context of large-scale influenza VE studies. The aims of our work here were twofold – one aim was simply to provide estimates for the seasonal influenza vaccine in two consecutive post-pandemic seasons (2011–12 and 2012–13) in Scotland, and another aim was to gain a better understanding of how estimates of VE vary by methodology, observational study design and clinical outcomes considered. This is important as estimates derived from observational study designs can often suffer from bias issues, if not appropriately accounted for. For instance, it is essential to make adjustments for the health-status and health-seeking behaviour of individuals. The data provided for our study came from a set of general practitioner practices that was considered to be representative of the Scottish population.

In total, five outcomes were considered: all-cause mortality, three different types of GP consultations for respiratory illness with differing levels of specificity for influenza and laboratory-confirmed influenza. Additionally, three study designs were applied: retrospective cohort (for all outcomes except laboratory-confirmed influenza), nested case-control (for laboratory-confirmed influenza only) and the screening method design (on consultations only). The methodologies applied for these study designs varied. Cox proportional hazards models and propensity score methods (matching and stratification) were applied in the retrospective cohort design; a generalised additive model was used for the nested case-control study; and a mixed effects logistic regression model was used for the screening method design.

Both time periods that we examined were relatively weak influenza seasons but in 2011–12, influenza activity was especially low. As a consequence, VE estimates for that season were close to zero and not particularly informative. Hence we gave more focus to the 2012–13 season when more influenza was circulating. VE estimates for that season were around 20-50% when measured against outcomes with higher specificity. Using different clinical outcomes was found to have a larger impact on the VE estimates than the methodologies applied. However one method affected our VE estimates fairly drastically and this was using propensity score matching. The main reason for this was because the method used only a relatively small sample of matched-patients from the cohort while most other methods were applied on the whole cohort.

We investigated the issue of biased VE estimates against all-cause mortality in seniors [182] by splitting our influenza season into a pre-influenza and during influenza period. The expectation is that there should be no VE in the pre-influenza periods when almost no influenza is circulating. However, we still found positive VE estimates against all-cause mortality in those aged 65+ in the pre-influenza period which is a strong indication of residual bias. For those aged 65+, we had no information on their underlying illnesses, which was possibly the main reason for these biased estimates. Modelling the effect of an unobserved confounder (smoking) showed that including this variable would help reduce the bias in those estimates but not sufficiently to bring VE in the pre-influenza period down to zero. Hence, to measure VE against all-cause mortality, it is absolutely necessary to have better variables that inform us about their health levels.

## 8.2 Future Work

The cluster detection system we created for TB works well within the confines of what it is capable of and is flexible enough to allow for different numbers of missing digits or different lengths of coding systems. This means that it can be applied to other diseases that gather genetic data in the form of a numeric code. For example, a number of disease organisms are coded using multilocus sequence typing (MLST) [234] including *campylobacter*, *Neisseria meningitidis* and *Staphylococcus aureus*. Note that at this time, only one of those organisms mentioned causes a disease for which, we could term as vaccine-preventable. This is *Neisseria meningitidis* which causes meningitis.

For TB, the 24 digit MIRU system has been universally accepted and has the great advantage of providing genetic typing data in a portable and comparable format. Therefore it should continue to be used as the main system for genetic typing for at least, a number of years in Scotland. However, as MIRU only looks at a number of specific sections of the *M. tuberculosis* genome, it can fail detect smaller changes that occur over time such as microevolutionary changes [370]. Moreover, it is possible for two cases to be directly linked but have non-identical MIRU profiles for this reason.

It is becoming increasingly important to understand these changes, particularly with the threat of multidrug-resistant TB looming ever larger. Hence, in eventuality, MIRU is likely to be replaced by whole genome sequencing (WGS) which offers more discriminatory power and is capable of detecting these smaller changes. Evidence in favour of WGS has already been observed in a study by Török et al. [358] that found two linked TB cases in the UK using WGS when epidemiological investigation missed the transmission event. WGS will be used more widely as the cost of the procedure continues to decrease and the time taken for the process to be completed also continues to decrease. In response to this, mathematical and statistical methods will need to be developed to rapidly analyse the large quantities of data that will be produced.

The work that we done on estimating reproductive numbers could be improved and extended further. The first area to mention here is the spatial model for estimating reproductive numbers that we produced as it had remaining issues. There are two key parameters in the spatial model – $\lambda_0$ which is the increase in probability of transmission due to cases being in the same household, and $\lambda$ which is the rate of exponential decay in probability of disease transmission as distance between cases increases.

Finding optimal parameters for the spatial model proved to be challenging. In our work, we simply used *sensible* parameters aligned closely against parameters used in epidemic simulations rather than estimating best-fit parameters from the data itself. The ideal model would be able to simultaneously identify optimal parameter estimates for $\lambda_0$ and $\lambda$ while estimating reproductive numbers. However, the optimisation procedure we used from the `optim()` function in R is computationally demanding and frequently encountered problems when trying to estimate best-fitting parameters by minimising the root mean square error (RMSE) between predicted and observed case counts in the epidemic curve. Furthermore, the optimisation procedure was sensitive to the initial conditions set. The consequence of this was that it would return variable parameter estimates dependent on those initial conditions.

There are other feasible ways we could have estimated parameters that may have yielded better or more consistent results such as using a Bayesian approach [129]. This would allow the $\lambda_0$ and $\lambda$ parameters to be treated as random variables and suitable prior distributions could be specified for the parameters. As an example, we may wish to specify a distribution that allows the $\lambda$ parameter (exponential decay rate for distance) to vary between 0–1, which is the range we looked at in Section 5.4.2.

In a different direction, there is also potential to extend our work in other ways which may be important in certain disease outbreak scenarios. For instance, there is the possibility for the serial interval distribution to vary over time [202]. We did not find significant evidence for this in Chapter 5 but we must keep in mind that we only looked at the initial phase of the epidemic and when looking at a disease outbreak over a longer duration, these changes could be important. In those circumstances, these changes may have to be taken into consideration, particularly when using methods of estimating the effective reproductive number that rely so heavily on serial intervals such as those that we applied. The variations can occur due to changes in the population including differing proportions of infected individuals over time as well as depletion of susceptible individuals (assuming that individuals become immune for at least a period of time, following recovery from infection). Given that time-dependant serial interval data exists, a simple way to account for these changes would be to split an outbreak into distinct time periods and apply different serial interval distributions for those periods. A more sophisticated approach would be able to gradually transition over time from an initially specified distribution.

Another way that our estimates could be improved would come if detailed data is available on imported cases. If we have knowledge of when imported cases got infected along with the country that they got infected in, then we can adjust the serial interval distribution for these cases. Specifically, a portion of time has to be removed for imported cases as they spend a portion of time outside the country under consideration and are thus, unable to infect individuals in that population during that time. It has to be noted that in the situation where there are relatively few imported cases and they acquired their infection in countries where the transit time to the country of interest is short, the points we make here can probably be ignored. However, in the opposite scenario, this could have a reasonable impact on estimates.

There is also plenty of scope to produce better estimates of VE for seasonal influenza. The most obvious starting point for this would be to have better frailty indicators for individuals, especially for those aged 65+, for whom we had no information on chronic disease. Furthermore, we were missing a number of the confounding factors that Valenciano et al. [362] suggested that researchers should have for studies of influenza VE. These include variables to measure smoking history, severity of chronic illness and functional status. From these, severity of chronic index could be measured using the Charlson index [64] or by looking at numbers of hospitalisations related to chronic illness within a timeframe (e.g. last 12 months); functional status could be measured using the Barthel index [69], which measures how well individuals can perform daily activities such as walking and bathing. As a minimum, to reduce bias in estimates of VE against all-cause mortality in seniors, we should have information on the number of underlying chronic diseases for those aged 65+.

VE could also have been measured against more clinical outcomes, with some being more specific for influenza. For instance, instead of looking at all-cause mortality, we could have looked specifically at those that died from causes related to pneumonia and influenza. Data on cause of death is recorded by the General Register Office for Scotland (GROS) in ICD-10 (international classification of diseases) form [275]. The main issue with trying to have a much more specific death outcome is that for a small country such as Scotland, we may only have very few deaths resulting from these specific causes in a given season which would make analysis difficult. To get round the issue of small numbers, we could also consider measuring VE against hospitalisations resulting from influenza, which we are likely to have more data on. Again, causes for hospitalisation are recorded in Scotland using ICD-10, which allows this information to be extracted. Measuring VE against these additional outcomes would bring us closer in our search to find the most reliable influenza VE estimates using data that is readily available in Scotland.

For the analysis of disease outbreaks, the confidence that investigators and analysts have in the conclusions they draw will always vary, dependant on key factors such as the data available and the methodologies considered. All of the suggestions made to improve the work in this thesis follow those themes. Recent infectious disease outbreaks such Zika virus and Ebola [113, 125] provide clear reminders that we must continue to improve the ways in which we deal with infectious diseases. One of the pivotal ways in which this will be possible comes from the fact that the amount as well as types of data continues to increase, potentially offering more avenues to analyse disease outbreaks – for instance, social media data has been used to alert authorities to cases of influenza in China [321]. Therefore, it seems certain that the role of the statistician in epidemiology will become ever more prominent and thus, investigating the ways in which statistics can be applied will continue to be a vibrant area of research.

# Bibliography

[1] Adler, F. R. (1992). The effects of averaging on the basic reproduction ratio. *Mathematical Biosciences 111*(1), 89–98.

[2] Agresti, A. (2007). *An introduction to categorical data analysis*, Volume 423. Wiley-Interscience.

[3] Agresti, A. and B. A. Coull (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician 52*(2), 119–126.

[4] Ahmed, A., K. G. Nicholson, and J. S. Nguyen-Van-Tam (1995). Reduction in mortality associated with influenza vaccine during 1989-90 epidemic. *The Lancet 346*(8975), 591–595.

[5] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

[6] Akobeng, A. (2005). Understanding randomised controlled trials. *Archives of Disease in Childhood 90*(8), 840–844.

[7] American Diabetes Association and others (2013). Standards of medical care in diabetes – 2013. *Diabetes Care 36*(Suppl 1).

[8] Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

[9] Anderson, R. M., R. M. May, and B. Anderson (1992). *Infectious diseases of humans: dynamics and control*, Volume 28. Wiley Online Library.

[10] Andrews, N., P. A. Waight, R. Borrow, S. Ladhani, R. C. George, M. P. Slack, and E. Miller (2011). Using the indirect cohort design to estimate the effectiveness of the seven valent pneumococcal conjugate vaccine in England and Wales. *PLoS One 6*(12).

[11] Andrews, N. J., P. A. Waight, R. C. George, M. P. Slack, and E. Miller (2012). Impact and effectiveness of 23-valent pneumococcal polysaccharide vaccine against

invasive pneumococcal disease in the elderly in England and Wales. *Vaccine 30*(48), 6802–6808.

[12] Athreya, K. B. and P. E. Ney (2012). *Branching processes*, Volume 196. Springer Science & Business Media.

[13] Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine 27*(12), 2037–2049.

[14] Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine 28*, 3083–3107.

[15] Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research 46*(3), 399–424.

[16] Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics 10*(2), 150–161.

[17] Austin, P. C., P. Grootendorst, and G. M. Anderson (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine 26*(4), 734–753.

[18] Awofisayo, A., G. Smith, B. Olowokure, Y. Rehman, H. Mohammed, H. Duggal, K. Janmohamed, V. de Souza, F. Wilson, S. Ibbotson, M. Catchpole, H. Osman, E. Smit, N. Phin, J. Watson, S. Palmer, R. Pebody, J. Ellis, A. Bermingham, and M. Zambon (2009, July). Preliminary descriptive epidemiology of a large school outbreak of influenza A(H1N1)v in the West Midlands, United Kingdom, May 2009. *Euro Surveillance 14*(27).

[19] Ayele, W., S. Neill, J. Zinsstag, M. Weiss, and I. Pavlik (2004). Bovine tuberculosis: an old disease but a new threat to Africa. *The International Journal of Tuberculosis and Lung Disease 8*(8), 924–937.

[20] Barreto, M. L., M. G. Teixeira, and E. H. Carmo (2006). Infectious diseases epidemiology. *Journal of Epidemiology and Community Health 60*(3), 192–195.

[21] Batel, P., F. Pessione, C. Maitre, and B. Rueff (1995). Relationship between alcohol and tobacco dependencies among alcoholics who smoke. *Addiction 90*(7), 977–980.

[22] Bates, J. H. and W. W. Stead (1993). The history of tuberculosis as a global epidemic. *The Medical Clinics of North America 77*(6), 1205–1217.

[23] Baxter, R., G. Ray, and B. Fireman (2010). Effect of influenza vaccination on hospitalizations in persons aged 50 years and older. *Vaccine 28*(45), 7267–7272.

[24] Bazin, H. (2000). *The eradication of smallpox: Edward Jenner and the first and only eradication of a human infectious disease.* Academic Press San Diego, CA.

[25] Becker, N. (1974). On parametric estimation for mortal branching processes. *Biometrika 61*(2), 393–399.

[26] Becker, N. and I. Marschner (1990). The effect of heterogeneity on the spread of disease. In *Stochastic Processes in Epidemic Theory*, pp. 90–103. Springer.

[27] Becker, N. G. (1989). *Analysis of infectious disease data*, Volume 33. Chapman & Hall/CRC.

[28] Belshe, R. B. (2009). Implications of the emergence of a novel H1 influenza virus. *New England Journal of Medicine 360*(25), 2667–2668.

[29] Benjamin, R. M. (2010). Oral health: the silent epidemic. *Public Health Reports 125*(2), 158.

[30] Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal 52*(6), 708–721.

[31] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

[32] Bennett, D. E., J. M. Courval, I. Onorato, T. Agerton, J. D. Gibson, L. Lambert, G. M. McQuillan, B. Lewis, T. R. Navin, and K. G. Castro (2008). Prevalence of tuberculosis infection in the United States population: the national health and nutrition examination survey, 1999–2000. *American Journal of Respiratory and Critical Care Medicine 177*(3), 348–355.

[33] Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 179–195.

[34] Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *BMJ: British Medical Journal 312*(7040), 1215.

[35] Bland, J. M. and D. G. Altman (1995). Multiple significance tests: the Bonferroni method. *BMJ 310*(6973), 170.

[36] Bloch, R., A. Dooneief, A. Buchberg, and S. Spellman (1954). The clinical effect of isoniazid and iproniazid in the treatment of pulmonary tuberculosis. *Annals of Internal Medicine 40*(5), 881–900.

[37] Boelle, P., P. Bernillon, J. Desenclos, et al. (2009). A preliminary estimation of the reproduction ratio for new influenza A (H1N1) from the outbreak in Mexico, March-April 2009. *Eurosurveillance 14*(19).

[38] Boivin, G., Z. Coulombe, and C. Wat (2003). Quantification of the influenza virus load by real-time polymerase chain reaction in nasopharyngeal swabs of patients treated with oseltamivir. *Journal of Infectious Diseases 188*(4), 578–580.

[39] Borg, I. and P. J. Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

[40] Bothamley, G. (2005). Smoking and tuberculosis: a chance or causal association? *Thorax 60*(7), 527–528.

[41] Boum, Y., D. Atwine, P. Orikiriza, J. Assimwe, A.-L. Page, J. Mwanga-Amumpaire, and M. Bonnet (2014). Male gender is independently associated with pulmonary tuberculosis among sputum and non-sputum producers people with presumptive tuberculosis in Southwestern Uganda. *BMC Infectious Diseases 14*(1), 638.

[42] Breman, J. G. and I. Arita (1980). The confirmation and maintenance of smallpox eradication. *The New England Journal of Medicine 303*(22), 1263–1273.

[43] Brookmeyer, R. (1998). Incubation period of infectious diseases. *Encyclopedia of Biostatistics*.

[44] Broome, C. V., R. R. Facklam, and D. W. Fraser (1980). Pneumococcal disease after pneumococcal vaccination: an alternative method to estimate the efficacy of pneumococcal vaccine. *New England Journal of Medicine 303*(10), 549–552.

[45] Brosch, R., S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, et al. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences 99*(6), 3684–3689.

[46] Buchholz, U., H. Bernard, D. Werber, M. M. Böhmer, C. Remschmidt, H. Wilking, Y. Deleré, M. an der Heiden, C. Adlhoch, J. Dreesman, et al. (2011). German outbreak of Escherichia coli O104: H4 associated with sprouts. *New England Journal of Medicine 365*(19), 1763–1770.

[47] Burnet, F. M. and D. O. White (1972). *Natural history of infectious disease.* CUP Archive.

[48] Butcher, J. C. (1987). *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods.* Wiley-Interscience.

[49] Butler, C. C., S. Rollnick, P. Kinnersley, L. Tapper-Jones, and H. Houston (2004). Communicating about expected course and re-consultation for respiratory tract infections in children: an exploratory study. *The British Journal of General Practice 54*(504), 536–538.

[50] Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing 16*(5), 1190–1208.

[51] Carcione, D., C. Giele, G. K. Dowse, D. B. Mak, L. Goggin, K. Kwan, S. Williams, D. Smith, and P. Effler (2010). Comparison of pandemic (H1N1) 2009 and seasonal influenza, Western Australia, 2009. *Emerging Infectious Diseases 16*(9), 1388.

[52] Carlson, A., S. F. Thung, and E. R. Norwitz (2009). H1N1 influenza in pregnancy: what all obstetric care providers ought to know. *Reviews in Obstetrics and Gynecology 2*(3), 139.

[53] Castilla, J., I. Martinez-Baz, V. Martinez-Artola, M. Fernandez-Alonso, G. Reina, M. Guevara, M. Garcia Cenoz, F. Elia, N. Alvarez, A. Barricarte, et al. (2013). Early estimates of influenza vaccine effectiveness in Navarre, Spain: 2012/13 mid-season analysis. *Euro Surveillance 18*(7), 2.

[54] Castilla, J., J. Morán, V. Martínez-Artola, M. Fernández-Alonso, M. Guevara, M. Cenoz, G. Reina, N. Alvarez, M. Arriazu, F. Elía, et al. (2011). Effectiveness of the monovalent influenza A (H1N1) 2009 vaccine in Navarre, Spain, 2009-2010: Cohort and case control study. *Vaccine*.

[55] Castro-Jiménez, M., J. Castillo-Pabón, G. Rey-Benito, P. Pulido-Domínguez, J. Barbosa-Ramírez, D. Velandia-Rodriguez, E. Angulo-Martínez, et al. (2009). Epidemiologic analysis of the laboratory-confirmed cases of influenza A (H1N1) v in Colombia. *Eurosurveillance 14*(30).

[56] Cauchemez, S., P.-Y. Boëlle, C. A. Donnelly, N. M. Ferguson, G. Thomas, G. M. Leung, A. J. Hedley, R. M. Anderson, and A.-J. Valleron (2006). Real-time estimates in early detection of SARS. *Emerging Infectious Diseases 12*(1), 110.

[57] Cauchemez, S., P.-Y. Boëlle, G. Thomas, and A.-J. Valleron (2006). Estimating in real time the efficacy of measures to control emerging communicable diseases. *American Journal of Epidemiology 164*(6), 591–597.

[58] Cauchemez, S., N. M. Ferguson, C. Wachtel, A. Tegnell, G. Saour, B. Duncan, and A. Nicoll (2009). Closure of schools during an influenza pandemic. *The Lancet Infectious Diseases 9*(8), 473–481.

[59] Cavanagh, R., M. Begon, M. Bennett, T. Ergon, I. M. Graham, P. E. de Haas, C. Hart, M. Koedam, K. Kremer, X. Lambin, et al. (2002). Mycobacterium microti infection (vole tuberculosis) in wild rodent populations. *Journal of Clinical Microbiology 40*(9), 3281–3285.

[60] Centers for Disease Control and Prevention (2009, May). Outbreak of swine-origin influenza A (H1N1) virus infection - Mexico, March-April 2009. *MMWR Morbidity and Mortality Weekly Report 58*(17), 467–470.

[61] Centers for Disease Control and Prevention (2011, October). A CDC Framework for Preventing Infectious Diseases. `http://www.cdc.gov/oid/docs/ID-Framework.pdf`. Accessed 2015 May 04.

[62] Centers for Disease Control and Prevention (2012). Steps of an Outbreak Investigation. `http://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson6/Section2.html`. Accessed 2015 May 11.

[63] Chan, M. (2009). World now at the start of 2009 influenza pandemic. *World Health Organization 11*.

[64] Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases 40*(5), 373–383.

[65] Christianson, S., J. Wolfe, P. Orr, J. Karlowsky, P. N. Levett, G. B. Horsman, L. Thibert, P. Tang, and M. K. Sharma (2010). Evaluation of 24 locus MIRU-VNTR genotyping of Mycobacterium tuberculosis isolates in Canada. *Tuberculosis 90*(1), 31–38.

[66] Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*(403), 596–610.

[67] Cohen, A. L., T. Taylor Jr, M. M. Farley, W. Schaffner, L. J. Lesher, K. A. Gershman, N. M. Bennett, A. Reingold, A. Thomas, J. Baumbach, et al. (2012). An assessment of the screening method to evaluate vaccine effectiveness: the case of 7-valent pneumococcal conjugate vaccine in the United States. *PloS One 7*(8).

[68] Colditz, G. A., C. S. Berkey, F. Mosteller, T. F. Brewer, M. E. Wilson, E. Burdick, and H. V. Fineberg (1995). The efficacy of bacillus Calmette-Guerin vaccination of newborns and infants in the prevention of tuberculosis: meta-analyses of the published literature. *Pediatrics 96*(1), 29–35.

[69] Collin, C., D. Wade, S. Davies, and V. Horne (1988). The Barthel ADL Index: a reliability study. *Disability & Rehabilitation 10*(2), 61–63.

[70] Comas, I., S. Homolka, S. Niemann, and S. Gagneux (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS One 4*(11).

[71] Coughlin, S. S. (1990). Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology 43*(1), 87–91.

[72] Cowling, B. J., V. J. Fang, S. Riley, J. M. Peiris, and G. M. Leung (2009). Estimation of the serial interval of influenza. *Epidemiology (Cambridge, Mass.) 20*(3), 344.

[73] Cowling, B. J., E. H. Lau, C. L. Lam, C. K. Cheng, J. Kovar, K. H. Chan, J. M. Peiris, and G. M. Leung (2008). Effects of school closures, 2008 winter influenza season, Hong Kong. *Emerging Infectious Diseases 14*(10), 1660.

[74] Crozier, G. and A. I. Schulte-Hostedde (2014). The ethical dimensions of wildlife disease management in an evolutionary context. *Evolutionary Applications 7*(7), 788–798.

[75] Cruz-Pacheco, G., L. Duran, L. Esteva, A. Minzoni, M. Lopez-Cervantes, P. Panayotaros, A. Ahued Ortega, I. Villasenor Ruiz, et al. (2009). Modelling of the influenza A (H1N1) v outbreak in Mexico City, April-May 2009, with control sanitary measures. *Eurosurveillance 14*(26).

[76] D'Agostino, R. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine 17*(19), 2265–2281.

[77] Davies, H. T. O., I. K. Crombie, and M. Tavakoli (1998). When can odds ratios mislead? *BMJ 316*(7136), 989–991.

[78] de Groot, R. J., S. C. Baker, R. S. Baric, C. S. Brown, C. Drosten, L. Enjuanes, R. A. Fouchier, M. Galiano, A. E. Gorbalenya, Z. A. Memish, et al. (2013). Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. *Journal of Virology 87*(14), 7790–7792.

[79] de Jong, B. C., I. Adetifa, B. Walther, P. C. Hill, M. Antonio, M. Ota, and R. A. Adegbola (2010). Differences between TB cases infected with M. africanum, West-African type 2, relative to Euro-American M. tuberculosis-an update. *FEMS Immunology and Medical Microbiology 58*(1), 102.

[80] De Jong, M. and T. G. Kimman (1994). Experimental quantification of vaccine-induced reduction in virus transmission. *Vaccine 12*(8), 761–766.

[81] de Quadros, C. A., B. S. Hersh, J.-M. Olivé, J. K. Andrus, C. M. da Silveira, and P. A. Carrasco (1997). Eradication of wild poliovirus from the Americas: acute flaccid paralysis surveillance, 1988–1995. *Journal of Infectious Diseases 175*(Supplement 1).

[82] De Santis, M., A. Cavaliere, G. Straface, and A. Caruso (2006). Rubella infection in pregnancy. *Reproductive Toxicology 21*(4), 390–398.

[83] De Serres, G., T. Pilishvili, R. Link-Gelles, A. Reingold, K. Gershman, S. Petit, M. M. Farley, L. H. Harrison, R. Lynfield, N. M. Bennett, et al. (2012). Use of surveillance data to estimate the effectiveness of the 7-valent conjugate pneumococcal vaccine in children less than 5 years of age over a 9 year period. *Vaccine 30*(27), 4067–4072.

[84] De Serres, G., D. Skowronski, X. Wu, and C. Ambrose (2013). The test-negative design: validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. *Euro Surveillance 18*(37), 20585.

[85] Debin, M., V. Colizza, T. Blanchon, T. Hanslik, C. Turbelin, and A. Falchi (2013). Effectiveness of 2012–2013 influenza vaccine against influenza-like illness in general population: Estimation in a French web-based cohort. *Human Vaccines & Immunotherapeutics 10*(3), 0–1.

[86] Des Roches, A., L. Paradis, R. Gagnon, C. Lemire, P. Bégin, S. Carr, E. S. Chan, J. Paradis, L. Frenette, M. Ouakki, et al. (2012). Egg-allergic patients can be safely vaccinated against influenza. *Journal of Allergy and Clinical Immunology 130*(5), 1213–1216.

[87] DeWitte, S. N. (2014, May). Mortality Risk and Survival in the Aftermath of the Medieval Black Death. *PLoS One 9*(5).

[88] Diekmann, O., J. Heesterbeek, and J. Metz (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology 28*(4), 365–382.

[89] Dillman, D. A., M. D. Sinclair, and J. R. Clark (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly 57*(3), 289–304.

[90] Dobson, A. J., K. Kuulasmaa, E. Eberle, and J. Scherer (1991). Confidence intervals for weighted sums of Poisson parameters. *Statistics in Medicine 10*(3), 457–462.

[91] Donaldson, L. J., P. D. Rutter, B. M. Ellis, F. E. Greaves, O. T. Mytton, R. G. Pebody, and I. E. Yardley (2009). Mortality from pandemic A/H1N1 2009 influenza in England: public health surveillance study. *BMJ 339*.

[92] Donnelly, C. A., R. Woodroffe, D. Cox, F. J. Bourne, C. Cheeseman, R. S. Clifton-Hadley, G. Wei, G. Gettinby, P. Gilks, H. Jenkins, et al. (2005). Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature 439*(7078), 843–846.

[93] Dublin, L. I. and A. J. Lotka (1925). On the True Rate of Natural Increase: As Exemplified by the Population of the United States, 1920. *Journal of the American Statistical Association 20*(151), 305–339.

[94] Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics 6*(3), 241–252.

[95] Dworkin, M. S. (2011). *Cases in Field Epidemiology: A Global Perspective.* Jones & Bartlett Learning.

[96] Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap.* CRC press.

[97] Erlewyn-Lajeunesse, M., L. P. Hunt, P. T. Heath, and A. Finn (2012). Anaphylaxis as an adverse event following immunisation in the UK and Ireland. *Archives of Disease in Childhood*.

[98] Erlewyn-Lajeunesse, M., J. Lucas, and J. O. Warner (2011). Influenza immunization in egg allergy: an update for the 2011–2012 season. *Clinical & Experimental Allergy 41*(10), 1367–1370.

[99] Eurich, D., T. Marrie, J. Johnstone, and S. Majumdar (2008). Mortality reduction with influenza vaccine in patients with pneumonia outside "flu" season. *American Journal of Respiratory and Critical Care Medicine 178*(5), 527–533.

[100] European Centre for Disease Control and Prevention (2009). Protocol for cohort database studies to measure pandemic and seasonal influenza vaccine effectiveness in the European Union and European Economic Area Member States. `http://ecdc.europa.eu/en/publications/Publications/0907_TER_Influenza_AH1N1_Measuring_Influenza_Vaccine_Effectiveness_Protocol_Cohort_Database_Studies.pdf`. Accessed 2016 Feb 07.

[101] European Centre for Disease Prevention and Control (2008). ECDC Guidance: Priority risk groups for influenza vaccination. `http://ecdc.europa.eu/en/`

publications/Publications/0808_GUI_Priority_Risk_Groups_for_Influenza_Vaccination.pdf. Accessed 2016 Mar 09.

[102] Euser, A. M., C. Zoccali, K. J. Jager, and F. W. Dekker (2009). Cohort studies: prospective versus retrospective. *Nephron Clinical Practice 113*(3).

[103] Evans, A. S. (1976). Causation and disease: the Henle-Koch postulates revisited. *The Yale Journal of Biology and Medicine 49*(2), 175.

[104] Evans, A. S. (1993). *Causation and disease: a chronological journey.* Springer.

[105] Evans, R. J. (1988). Epidemics and revolutions: cholera in nineteenth-century Europe. *Past and Present*, 123–146.

[106] Everitt, B., S. Rabe-Hesketh, B. S. Everitt, G. B. Statisticien, B. S. Everitt, and G. B. Statistician (1997). *The analysis of proximity data.* Arnold London.

[107] Fairchild, A. L. and G. M. Oppenheimer (1998). Public health nihilism vs pragmatism: history, politics, and the control of tuberculosis. *American Journal of Public Health 88*(7), 1105–1117.

[108] Faraway, J. J. (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* CRC press.

[109] Farrington, C. (1993). Estimation of vaccine effectiveness using the screening method. *International Journal of Epidemiology 22*(4), 742–746.

[110] Farrington, C., M. Kanaan, and N. Gay (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(3), 251–292.

[111] Farrington, C., M. Kanaan, and N. Gay (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics 4*(2), 279–295.

[112] Farrington, C. P., E. Miller, and B. Taylor (2001). MMR and autism: further evidence against a causal association. *Vaccine 19*(27), 3632–3635.

[113] Fauci, A. S. and D. M. Morens (2016). Zika virus in the Americas – yet another arbovirus threat. *New England Journal of Medicine 374*(7), 601–604.

[114] Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology 186*(5), 1518–1530.

[115] Ferguson, N. M., C. A. Donnelly, and R. M. Anderson (2001). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science 292*(5519), 1155–1160.

[116] Fine, M. J., M. A. Smith, C. A. Carson, F. Meffe, S. S. Sankey, L. A. Weissfeld, A. S. Detsky, and W. N. Kapoor (1994). Efficacy of pneumococcal vaccination in adults: a meta-analysis of randomized controlled trials. *Archives of Internal Medicine 154*(23), 2666–2677.

[117] Fine, P., K. Eames, and D. L. Heymann (2011). Herd immunity: a rough guide. *Clinical Infectious Diseases 52*(7), 911–916.

[118] Fine, P. E. (2003). The interval between successive cases of an infectious disease. *American Journal of Epidemiology 158*(11), 1039–1047.

[119] Fireman, B., J. Lee, N. Lewis, O. Bembom, M. Van Der Laan, and R. Baxter (2009). Influenza vaccination and mortality: differentiating vaccine effects from bias. *American Journal of Epidemiology 170*(5), 650–656.

[120] Fleming, D., N. Andrews, J. Ellis, A. Bermingham, P. SebastianPillai, A. Elliot, E. Miller, and M. Zambon (2010). Estimating influenza vaccine effectiveness using routinely collected laboratory data. *Journal of Epidemiology and Community Health 64*(12), 1062.

[121] Fouchier, R. A., T. Kuiken, M. Schutten, G. Van Amerongen, G. J. van Doornum, B. G. van den Hoogen, M. Peiris, W. Lim, K. Stöhr, and A. D. Osterhaus (2003). Aetiology: Koch's postulates fulfilled for SARS virus. *Nature 423*(6937), 240–240.

[122] Francisco, A. P., M. Bugalho, M. Ramirez, and J. A. Carriço (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics 10*(1), 152.

[123] Fraser, C., C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, et al.

(2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science 324*(5934), 1557–1561.

[124] Fraser, C., S. Riley, R. M. Anderson, and N. M. Ferguson (2004). Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the United States of America 101*(16), 6146–6151.

[125] Frieden, T. R., I. Damon, B. P. Bell, T. Kenyon, and S. Nichol (2014). Ebola 2014 – new challenges, new global response and responsibility. *New England Journal of Medicine 371*(13), 1177–1180.

[126] Funk, S., T. L. Bogich, K. E. Jones, A. M. Kilpatrick, and P. Daszak (2013). Quantifying trends in disease impact to produce a consistent and reproducible definition of an emerging infectious disease. *PloS One 8*(8).

[127] Funk, S., E. Gilad, C. Watkins, and V. A. Jansen (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences 106*(16), 6872–6877.

[128] Gefenaite, G., J. Rahamat-Langendoen, A. Ambrozaitis, A. Mickiene, L. Jancoriene, M. Kuliese, D. Velyvyte, H. Niesters, R. P. Stolk, K. Zagminas, et al. (2014). Seasonal influenza vaccine effectiveness against influenza in 2012–2013: A hospital-based case-control study in Lithuania. *Vaccine 32*(7), 857–863.

[129] Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. Taylor & Francis.

[130] General Register Office for Scotland (2012, August). High Level Summary of Statistics: Population and Migration, Deaths - Variation within Scotland. URL `http://www.gro-scotland.gov.uk/files2/stats/high-level-summary/j11198/j1119815.htm`.

[131] Gillespie, S. H. (2006). Non-Tuberculosis Mycobacteria. *Principles and Practice of Clinical Bacteriology, Second Edition*, 171–181.

[132] Ginsberg, M., J. Hopkins, A. Maroufi, G. Dunne, D. Sunega, J. Giessick, P. McVay, K. Lopez, P. Kriner, S. Munday, et al. (2009). Swine influenza A (H1N1) infection in two children-Southern California, March-April 2009. *Morbidity and Mortality Weekly Report 58*(15), 400–402.

[133] Gjini, E. and M. G. M. Gomes (2015). Expanding vaccine efficacy estimation with dynamic models fitted to cross-sectional prevalence data post-licensure. *Epidemics*.

[134] GmbH, M. S. (2013). *XLConnect: Excel Connector for R*. R package version 0.2-5.

[135] Goeyvaerts, N., N. Hens, B. Ogunjimi, M. Aerts, Z. Shkedy, P. V. Damme, and P. Beutels (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 255–277.

[136] Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

[137] Gower, J. C. and G. Ross (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 54–64.

[138] Gray, L. (2007). Comparisons of health-related behaviours and health measures between Glasgow and the rest of Scotland. *Glasgow: Glasgow Centre for Population Health*.

[139] Greenacre, M., R. Primicerio, and Fundación BBVA (2014). *Multivariate analysis of ecological data*. Bilbao: Fundación BBVA.

[140] Griffin, J. T., T. Garske, A. C. Ghani, and P. S. Clarke (2011). Joint estimation of the basic reproduction number and generation time parameters for infectious disease outbreaks. *Biostatistics 12*(2), 303–312.

[141] Groenwold, R., E. Hak, O. Klungel, and A. Hoes (2010). Instrumental Variables in Influenza Vaccination Studies: Mission Impossible?! *Value in Health 13*(1), 132–137.

[142] Gu, X. S. and P. R. Rosenbaum (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics 2*(4), 405–420.

[143] Haight, F. A. and M. A. Breuer (1960). The Borel-Tanner distribution. *Biometrika 47*(1/2), 143–150.

[144] Hak, E., T. Verheij, D. Grobbee, K. Nichol, and A. Hoes (2002). Confounding by indication in non-experimental evaluation of vaccine effectiveness: the example of prevention of influenza complications. *Journal of Epidemiology and Community Health 56*(12), 951–955.

[145] Halloran, M. and I. Longini Jr (2001). Using validation sets for outcomes and exposure to infection in vaccine field studies. *American Journal of Epidemiology 154*(5), 391–398.

[146] Halloran, M. E. (2006). Invited commentary: Challenges of using contact data to understand acute respiratory disease transmission. *American Journal of Epidemiology 164*(10), 945–946.

[147] Halloran, M. E., M. Haber, I. M. Longini, and C. J. Struchiner (1991). Direct and indirect effects in vaccine efficacy and effectiveness. *American Journal of Epidemiology 133*(4), 323–331.

[148] Halloran, M. E., I. M. Longini, C. J. Struchiner, I. M. Longini, and C. J. Struchiner (2010). *Design and analysis of vaccine studies.* Springer.

[149] Hancock, K., V. Veguilla, X. Lu, W. Zhong, E. N. Butler, H. Sun, F. Liu, L. Dong, J. R. DeVos, P. M. Gargiullo, et al. (2009). Cross-reactive antibody responses to the 2009 pandemic H1N1 influenza virus. *New England Journal of Medicine 361*(20), 1945–1952.

[150] Hansen, E. and T. Day (2010). Optimal antiviral treatment strategies and the effects of resistance. *Proceedings of the Royal Society B: Biological Sciences*.

[151] Hardelid, P., D. Fleming, J. McMenamin, N. Andrews, C. Robertson, P. SebastianPillai, J. Ellis, W. Carman, T. Wreghitt, J. Watson, et al. (2011). Effectiveness of pandemic and seasonal influenza vaccine in preventing pandemic influenza A (H1N1) 2009 infection in England and Scotland 2009–2010. *Euro Surveillance 16*(2), 19763.

[152] Hardy, A. B., R. Varma, T. Collyns, S. J. Moffitt, C. Mullarkey, and J. P. Watson (2010). Cost-effectiveness of the NICE guidelines for screening for latent tuberculosis infection: the QuantiFERON-TB Gold IGRA alone is more cost-effective for immigrants from high burden countries. *Thorax 65*(2), 178–180.

[153] Harvala, H., R. Gunson, P. Simmonds, A. Hardie, S. Bennett, F. Scott, H. Roddie, J. McKnight, T. Walsh, D. Rowney, et al. (2010). The emergence of oseltamivir-resistant pandemic influenza A (H1N1) 2009 virus amongst hospitalised immunocompromised patients in Scotland, November-December, 2009. *Euro surveillance 15*(14).

[154] Hauer, B., B. Brodhun, D. Altmann, L. Fiebig, R. Loddenkemper, and W. Haas (2011). Tuberculosis in the elderly in Germany. *European Respiratory Journal 38*(2), 467–470.

[155] Health Protection Agency (2009, June). Epidemiology of new influenza A (H1N1) virus infection, United Kingdom, April-June 2009. *Euro Surveillance 14*(22).

[156] Health Protection Agency (2012). Surveillance of influenza and other respiratory pathogens in the UK, 2011/12. `http://webarchive.nationalarchives.gov.uk/20140714084352/http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317134705939`. Accessed 2016 Feb 24.

[157] Health Protection Agency (2013). Surveillance of influenza and other respiratory pathogens, including novel respiratory viruses, in the United Kingdom: Winter 2012/13. `http://webarchive.nationalarchives.gov.uk/20140714084352/http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317139321787`. Accessed 2016 Mar 09.

[158] Health Protection Scotland. ECOSS (The Electronic Communication of Surveillance in Scotland). `http://www.hps.scot.nhs.uk/surveillance/SystemsDetail.aspx?id=248`. Accessed 2013 Nov 10.

[159] Health Protection Scotland. Influenza Surveillance Systems. `http://www.hps.scot.nhs.uk/resp/influenzasurveillancesystems.aspx`. Accessed 2013 Nov 10.

[160] Health Protection Scotland (2007). Surveillance Systems – Enhanced Surveillance of Mycobacterial Infections in Scotland (ESMI). `http://www.hps.scot.nhs.uk/resp/ssdetail.aspx?id=15`. Accessed 2013 Dec 07.

[161] Health Protection Scotland (2010, December). The Pandemic of Influenza A(H1N1) Infection in Scotland 2009-2010 A Report on the Health Protection Response. `http://www.documents.hps.scot.nhs.uk/respiratory/swine-influenza/outbreak-report/flu-a-h1n1-hp-response-2010-12.pdf`. Accessed 2015 Sep 06.

[162] Health Protection Scotland (2013). Respiratory Infections – Tuberculosis. `http://www.hps.scot.nhs.uk/resp/tuberculosis.aspx`. Accessed 2013 Dec 07.

[163] Heesterbeek, J. (2002). A brief history of R0 and a recipe for its calculation. *Acta Biotheoretica 50*(3), 189–204.

[164] Heffernan, J., R. Smith, and L. Wahl (2005). Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface 2*(4), 281–293.

[165] Hens, N., Z. Shkedy, M. Aerts, C. Faes, P. Van Damme, and P. Beutels (2012). *Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective*, Volume 63. Springer Science & Business Media.

[166] Herrador, B., P. Aavitsland, B. Feiring, M. Bergsaker, and K. Borgen (2012). Usefulness of health registries when estimating vaccine effectiveness during the Influenza A (H1N1) pdm09 pandemic in Norway. *BMC Infectious Diseases 12*(1), 63.

[167] Hershkovitz, I., H. D. Donoghue, D. E. Minnikin, G. S. Besra, O. Y. Lee, A. M. Gernaey, E. Galili, V. Eshed, C. L. Greenblatt, E. Lemma, et al. (2008). Detection and molecular characterization of 9000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. *PloS One 3*(10).

[168] Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review 42*(4), 599–653.

[169] Hide, G., O. Gerwash, E. Morley, R. Williams, J. Hughes, D. Thomasson, M. Elmahaishi, K. Elmahaishi, R. Terry, and J. Smith (2007). Does vertical transmission contribute to the prevalence of toxoplasmosis? *Parassitologia 49*(4), 223–226.

[170] Higgins, C. A., R. J. McClean, and D. W. Conrath (1985). The accuracy and biases of diary communication data. *Social Networks 7*(2), 173–187.

[171] Hirotsu, N., H. Ikematsu, N. Iwaki, N. Kawai, T. Shigematsu, O. Kunishima, and S. Kashiwagi (2004). Effects of antiviral drugs on viral detection in influenza patients and on the sequential infection to their family members–serial examination by rapid diagnosis (Capilia) and virus culture. In *International Congress Series*, Volume 1263, pp. 105–108. Elsevier.

[172] Höhle, M. (2003). R0 estimation by the martingale method.

[173] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

[174] Hosmer, D. W. and S. Lemeshow (2004). *Applied logistic regression*, Volume 354. Wiley-Interscience.

[175] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417.

[176] Hothorn, T. and B. S. Everitt (2009). *A handbook of statistical analyses using R*. CRC Press.

[177] Influenza A(H1N1)v Investigation Teams (2009). Modified surveillance of influenza A(H1N1)v virus infections in France. *Euro Surveillance 14*, 29.

[178] International Agency for Research on Cancer and World Health Organization and others (2014). *GLOBOCAN: Estimated Cancer Incidence, Mortality, and Prevalence Worldwide in 2012*. IARC.

[179] ISD Scotland (2013a). General Practice - Practice Team Information. `http://www.isdscotland.org/Health-Topics/General-Practice/GP-Consultations`. Accessed 2013 Nov 10.

[180] ISD Scotland (2013b). NHS Board (1st April 2006 configuration) Population Estimates 1982-2013. `http://www.isdscotland.org/Products-and-Services/GPD-Support/Population/Estimates/index.asp`. Accessed 2015 May 15.

[181] ISD Scotland (2015). Coding and Terminology Systems. `http://www.isdscotland.org/Products-and-Services/Terminology-Services/Coding-and-Terminology-Systems`. Accessed 2015 Dec 10.

[182] Jackson, L., M. Jackson, J. Nelson, K. Neuzil, and N. Weiss (2006). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology 35*(2), 337–344.

[183] Jackson, L., J. Nelson, P. Benson, K. Neuzil, R. Reid, B. Psaty, S. Heckbert, E. Larson, and N. Weiss (2006). Functional status is a confounder of the association of influenza vaccine and risk of all cause mortality in seniors. *International Journal of Epidemiology 35*(2), 345–352.

[184] Jackson, M., J. Nelson, N. Weiss, K. Neuzil, W. Barlow, and L. Jackson (2008). Influenza vaccination and risk of community-acquired pneumonia in immunocompetent

elderly people: a population-based, nested case-control study. *The Lancet 372*(9636), 398–405.

[185] Jain, S., L. Kamimoto, A. M. Bramley, A. M. Schmitz, S. R. Benoit, J. Louie, D. E. Sugerman, J. K. Druckenmiller, K. A. Ritger, R. Chugh, et al. (2009). Hospitalized patients with 2009 H1N1 influenza in the United States, April–June 2009. *New England Journal of Medicine 361*(20), 1935–1944.

[186] Jamieson, D. J., M. A. Honein, S. A. Rasmussen, J. L. Williams, D. L. Swerdlow, M. S. Biggerstaff, S. Lindstrom, J. K. Louie, C. M. Christ, S. R. Bohm, et al. (2009). H1N1 2009 influenza virus infection during pregnancy in the USA. *The Lancet 374*(9688), 451–458.

[187] Jansen, V. A., N. Stollenwerk, H. J. Jensen, M. Ramsay, W. Edmunds, and C. Rhodes (2003). Measles outbreaks in a population with declining vaccine uptake. *Science 301*(5634), 804–804.

[188] Jefferson, T., M. Jones, P. Doshi, and C. Del Mar (2009). Neuraminidase inhibitors for preventing and treating influenza in healthy adults: systematic review and meta-analysis. *BMJ: British Medical Journal 339*.

[189] Jewell, N. P. (2003). *Statistics for epidemiology*, Volume 58. Chapman & Hall/CRC.

[190] Jhung, M. A., D. Swerdlow, S. J. Olsen, D. Jernigan, M. Biggerstaff, L. Kamimoto, K. Kniss, C. Reed, A. Fry, L. Brammer, et al. (2011). Epidemiology of 2009 pandemic influenza A (H1N1) in the United States. *Clinical Infectious Diseases 52*(suppl 1).

[191] Jimenez-Jorge, S., S. d. Mateo, F. Pozo, I. Casas, M. Garcia Cenoz, J. Castilla, V. Gallardo, E. Pérez, T. Vega, C. Rodriguez, et al. (2012). Early estimates of the effectiveness of the 2011/12 influenza vaccine in the population targeted for vaccination in Spain, 25 December 2011 to 19 February 2012. *Eurosurveillance*.

[192] Johnson, N. P. and J. Mueller (2002). Updating the accounts: global mortality of the 1918–1920 "Spanish" influenza pandemic. *Bulletin of the History of Medicine 76*(1), 105–115.

[193] Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

[194] Jones, C. L., K. M. Milsom, P. Ratcliffe, A. Wyllie, T. V. Macfarlane, and M. Tickle (2011). Clinical outcomes of single-visit oral prophylaxis: a practice-based randomised controlled trial. *BMC Oral Health 11*(1), 1.

[195] Kamerbeek, J., L. Schouls, A. Kolk, M. Van Agterveld, D. Van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, et al. (1997). Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *Journal of Clinical Microbiology 35*(4), 907–914.

[196] Kavanagh, K., C. Robertson, and J. McMenamin (2011). Assessment of the Variability in Influenza A (H1N1) Vaccine Effectiveness Estimates Dependent on Outcome and Methodological Approach. *PloS One 6*(12).

[197] Keeling, M. J. and L. Danon (2009). Mathematical modelling of infectious diseases. *British Medical Bulletin 92*(1), 33–42.

[198] Keeling, M. J. and P. Rohani (2011). *Modeling infectious diseases in humans and animals.* Princeton University Press.

[199] Kelly, H., K. Carville, K. Grant, P. Jacoby, T. Tran, and I. Barr (2009). Estimation of influenza vaccine effectiveness from routine surveillance data. *PLoS One 4*(3).

[200] Kelly, H. A., K. A. Grant, S. Williams, J. Fielding, D. Smith, et al. (2009). Epidemiological characteristics of pandemic influenza H1N1 2009 and seasonal influenza infection. *Medical Journal of Australia 191*(3), 146–149.

[201] Kelly, L., A. Brouwer, A. Wilson, P. Gale, E. Snary, D. Ross, and C. De Vos (2013). Epidemic Threats to the European Union: Expert Views on Six Virus Groups. *Transboundary and Emerging Diseases 60*(4), 360–369.

[202] Kenah, E., M. Lipsitch, and J. M. Robins (2008). Generation interval contraction and epidemic data analysis. *Mathematical Biosciences 213*(1), 71.

[203] Kenfield, S. A., M. J. Stampfer, B. A. Rosner, and G. A. Colditz (2008). Smoking and smoking cessation in relation to mortality in women. *JAMA 299*(17), 2037–2047.

[204] Kermack, W. O. and A. G. McKendrick (1932). Contributions to the mathematical theory of epidemics. II. The problem of endemicity. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Volume 138, pp. 55–83. The Royal Society.

[205] Khan, M. S., O. Dar, C. Sismanidis, K. Shah, and P. Godfrey-Faussett (2007). Improvement of tuberculosis case detection and reduction of discrepancies between men and women by simple sputum-submission instructions: a pragmatic randomised controlled trial. *The Lancet 369*(9577), 1955–1960.

[206] King, G. E., L. E. Markowitz, P. A. Patriarca, and L. G. Dales (1991). Clinical efficacy of measles vaccine during the 1990 measles epidemic. *The Pediatric Infectious Disease Journal 10*(12), 883–888.

[207] Kissling, E., M. Valenciano, J. Falcao, A. Larrauri, K. Widgren, D. Pitigoi, B. Oroszi, B. Nunes, C. Savulescu, A. Mazick, et al. (2009). I-MOVE towards monitoring seasonal and pandemic influenza vaccine effectiveness: lessons learnt from a pilot multi-centric case-control study in Europe, 2008-9. *Eurosurveillance*, 1–8.

[208] Knol, M. J., S. Le Cessie, A. Algra, J. P. Vandenbroucke, and R. H. Groenwold (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. *Canadian Medical Association Journal 184*(8), 895–899.

[209] Koep, T. H., F. T. Enders, C. Pierret, S. C. Ekker, D. Krageschmidt, K. L. Neff, M. Lipsitch, J. Shaman, and W. C. Huskins (2013). Predictors of indoor absolute humidity and estimated effects on influenza virus survival in grade schools. *BMC Infectious Diseases 13*(1), 71.

[210] Kopf, E. W. (1916). Florence Nightingale as statistician. *Quarterly publications of the American Statistical Association 15*(116), 388–404.

[211] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika 29*(2), 115–129.

[212] Last, J. M., I. E. Association, et al. (2001). *A dictionary of epidemiology*, Volume 44. Oxford University Press.

[213] Lawn, S. D. and A. I. Zumla (2011, July). Tuberculosis. *Lancet 378*(9785), 57–72.

[214] Lederberg, J., R. E. Shope, S. C. Oaks Jr, et al. (1992). *Emerging Infections:: Microbial Threats to Health in the United States*. National Academies Press.

[215] Lederman, M. (1981). The early history of radiotherapy: 1895–1939. *International Journal of Radiation Oncology* Biology* Physics 7*(5), 639–648.

[216] Lee, N., P. K. Chan, D. S. Hui, T. H. Rainer, E. Wong, K.-W. Choi, G. C. Lui, B. C. Wong, R. Y. Wong, W.-Y. Lam, et al. (2009). Viral loads and duration of viral shedding in adult patients hospitalized with influenza. *Journal of Infectious Diseases 200*(4), 492–500.

[217] Legrand, J., E. Vergu, and A. Flahault (2006). Real-time monitoring of the influenza vaccine field effectiveness. *Vaccine 24*(44), 6605–6611.

[218] Leung, G. M. and A. Nicoll (2010). Reflections on pandemic (H1N1) 2009 and the international response. *PLoS Medicine 7*(10).

[219] Lewallen, S. and P. Courtright (1998). Epidemiology in practice: case-control studies. *Community Eye Health 11*(28), 57.

[220] Lifson, A. and M. Rogers (1986). Vertical transmission of human immunodeficiency virus. *The Lancet 328*(8502), 337.

[221] Lipsitch, M. and C. T. Bergstrom (2004). Invited commentary: real-time tracking of control measures for emerging infections. *American Journal of Epidemiology 160*(6), 517–519.

[222] Lipsitch, M., T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, et al. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science 300*(5627), 1966–1970.

[223] Lipsitch, M., E. Tchetgen, and T. Cohen (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.) 21*(3), 383.

[224] Lone, N., C. Simpson, K. Kavanagh, C. Robertson, J. McMenamin, L. Ritchie, and A. Sheikh (2012a). Seasonal Influenza Vaccine Effectiveness in the community (SIVE): protocol for a cohort study exploiting a unique national linked data set. *BMJ Open 2*(2).

[225] Lone, N. I., C. Simpson, K. Kavanagh, C. Robertson, J. McMenamin, L. Ritchie, and A. Sheikh (2012b). Seasonal Influenza Vaccine Effectiveness in the community (SIVE): protocol for a cohort study exploiting a unique national linked data set. *BMJ Open 2*(2).

[226] Longini Jr, I. M., E. Ackerman, and L. R. Elveback (1978). n optimization model for influenza A epidemics. *Mathematical Biosciences 38*(1), 141–157.

[227] López-Cervantes, M., A. Venado, A. Moreno, R. L. Pacheco-Domínguez, and G. Ortega-Pierres (2009). On the spread of the novel influenza A (H1N1) virus in Mexico. *The Journal of Infection in Developing Countries 3*(05), 327–330.

[228] Luo, Z., L. Li, and B. Ruan (2012). Impact of the implementation of a vaccination strategy on hepatitis B virus infections in China over a 20-year period. *International Journal of Infectious Diseases 16*(2).

[229] Macartney, K. (2014). Prevention of varicella: time for two-dose vaccination. *The Lancet 383*(9925), 1276–1277.

[230] Macfarlane, T., G. Macfarlane, N. Thakker, S. Benhamou, C. Bouchardy, W. Ahrens, H. Pohlabeln, P. Lagiou, A. Lagiou, X. Castellsague, et al. (2012). Role of medical history and medication use in the aetiology of upper aerodigestive tract cancers in Europe: the ARCAGE study. *Annals of Oncology 23*(4), 1053–1060.

[231] Macfarlane, T. V., M. M. Kawecki, C. Cunningham, I. Bovaird, R. Morgan, K. Rhodes, and R. Watkins (2011). Mouthwash use in general population: results from adult dental health survey in Grampian, Scotland. *Journal of Oral & Maxillofacial Research 1*(4).

[232] Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik (2012). *cluster: Cluster Analysis Basics and Extensions.* R package version 1.14.3 — For new features, see the 'Changelog' file (in the package source).

[233] Mahon, B. E., K. Hsu, S. Karumuri, S. L. Kaplan, E. O. Mason, S. I. Pelton, U. P. M. P. S. Group, et al. (2006). Effectiveness of abbreviated and delayed 7-valent pneumococcal conjugate vaccine dosing regimens. *Vaccine 24*(14), 2514–2520.

[234] Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences 95*(6), 3140–3145.

[235] Mardia, K. V., J. T. Kent, and J. M. Bibby (1980). Multivariate analysis.

[236] Massad, E., M. N. Burattini, F. A. B. Coutinho, and L. F. Lopez (2007). The 1918 influenza A epidemic in the city of Sao Paulo, Brazil. *Medical Hypotheses 68*(2), 442–445.

[237] McBryde, E., I. Bergeri, C. v. Gemert, J. Rotty, E. Headley, K. Simpson, R. Lester, M. Hellard, J. Fielding, M. Baker, et al. (2009). Early transmission characteristics of influenza A (H1N1) v in Australia: Victorian state, 16 May-3 June 2009. *Eurosurveillance 14*(42).

[238] McCormick, D., S. Thorn, D. Milne, C. Evans, J. Stevenson, M. Llano, and M. Donaghy (2011). Public health response to an outbreak of Legionnaires' disease in Edinburgh, United Kingdom, June 2012. *Euro Surveillance 17*(28), 905–913.

[239] McCullagh, P. and J. A. Nelder (1989). *Generalized linear model*, Volume 37. Chapman & Hall/CRC.

[240] McLean, E., R. Pebody, C. Campbell, M. Chamberland, C. Hawkins, J. Nguyen-Van-Tam, I. Oliver, G. Smith, C. Ihekweazu, S. Bracebridge, et al. (2010). Pandemic (H1N1) 2009 influenza in the UK: clinical and epidemiological findings from the first few hundred (FF100) cases. *Epidemiology and Infection 138*(11), 1531–1541.

[241] McMenamin, J., N. Andrews, C. Robertson, D. Fleming, H. Durnall, B. Von Wissmann, J. Ellis, A. Lackenby, S. Cottrell, B. Smyth, et al. (2013). Effectiveness of seasonal 2012/13 vaccine in preventing laboratory-confirmed influenza infection in primary care in the United Kingdom: mid-season analysis 2012/13. *Euro Surveillance 18*(5).

[242] McNutt, L.-A., C. Wu, X. Xue, and J. P. Hafner (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology 157*(10), 940–943.

[243] Menzies, D. (2013). Molecular methods for tuberculosis trials: time for whole-genome sequencing? *The Lancet Respiratory Medicine 1*(10), 759–761.

[244] Miller, E., K. Hoschler, P. Hardelid, E. Stanford, N. Andrews, and M. Zambon (2010). Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. *The Lancet 375*(9720), 1100–1108.

[245] Mishra, S., D. N. Fisman, and M.-C. Boily (2010). The ABC of terms used in mathematical models of infectious diseases. *Journal of Epidemiology and Community Health*.

[246] Morabia, A. (2004). *A history of epidemiologic methods and concepts*. Springer Science & Business Media.

[247] Morens, D. M., E. C. Holmes, A. S. Davis, and J. K. Taubenberger (2011). Global rinderpest eradication: lessons learned and why humans should celebrate too. *Journal of Infectious Diseases*.

[248] Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine 5*(3).

[249] Mukherjee, A., I. Saha, A. Sarkar, and R. Chowdhury (2012). Gender differences in notification rates, clinical forms and treatment outcome of tuberculosis patients under the RNTCP. *Lung India: Official Organ of Indian Chest Society 29*(2), 120.

[250] Munayco, C., J. Gomez, V. Laguna-Torres, J. Arrasco, T. Kochel, V. Fiestas, J. Garcia, J. Perez, I. Torres, F. Condori, et al. (2009). Epidemiological and transmissibility analysis of influenza A (H1N1) v in a southern hemisphere setting: Peru. *Euro Surveillance 14*(32).

[251] Murphy, K. R., A. Eivindson, K. Pauksens, W. J. Stein, G. Tellier, R. Watts, P. Léophonte, S. J. Sharp, and E. Loeschel (2000). Efficacy and safety of inhaled zanamivir for the treatment of influenza in patients with asthma or chronic obstructive pulmonary disease. *Clinical Drug Investigation 20*(5), 337–349.

[252] National Center for Health Statistics (US and others (2014). Health Risk Factors.

[253] National Institute for Health and Clinical Excellence (2009). Amantadine, oseltamivir and zanamivir for the treatment of influenza. Review of NICE technology appraisal guidance 58. Technical report. URL `http://www.nice.org.uk/nicemedia/live/11774/43268/43268.pdf`.

[254] National Institute of Allergy and Infectious Disease (2010). Community Immunity ("Herd" Immunity). `http://www.niaid.nih.gov/topics/pages/communityimmunity.aspx`. Accessed 2015 May 10.

[255] National Records of Scotland (2013). Mid-2011 and Mid-2012 Population Estimates Scotland. http://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/2012/list-of-tables. Accessed 2016 Feb 17.

[256] National Records of Scotland (2014). 2013 Mid-Year Population Estimates. http://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population. Accessed 2015 May 17.

[257] National Records of Scotland (2015a). Mid-Year Population Estimates. http://nationalrecordsofscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates. Accessed 2015 November 22.

[258] National Records of Scotland (2015b). Population Estimates by Urban Rural Classification (2001 Data Zone based). http://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/special-area-population-estimates/population-estimates-by-urban-rural-classification. Accessed 2016 Feb 17.

[259] Nchinda, T. C. (1998). Malaria: a reemerging disease in Africa. *Emerging Infectious Diseases 4*(3), 398.

[260] Nelson, K. E. and C. M. Williams (2012). *Infectious disease epidemiology.* Jones & Bartlett Publishers.

[261] Nelson, K. E. and C. M. Williams (2014). *Infectious disease epidemiology: theory and practice.* Jones & Bartlett Publishers.

[262] Newhouse, J. and M. McClellan (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health 19*(1), 17–34.

[263] Nichol, K. (2009). Challenges in evaluating influenza vaccine effectiveness and the mortality benefits controversy. *Vaccine 27*(45), 6305–6311.

[264] Nichol, K., J. Nordin, D. Nelson, J. Mullooly, and E. Hak (2007). Effectiveness of influenza vaccine in the community-dwelling elderly. *New England Journal of Medicine 357*(14), 1373–1381.

[265] Nishiura, H. (2007). Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. *Emerging Themes in Epidemiology 4*(2), 1–12.

[266] Nishiura, H., C. Castillo-Chavez, M. Safan, G. Chowell, et al. (2009). Transmission potential of the new influenza A (H1N1) virus and its age-specificity in Japan. *Euro Surveillance 14*(22), 19227.

[267] Nishiura, H., G. Chowell, M. Safan, C. Castillo-Chavez, et al. (2010). Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza A (H1N1) 2009. *Theoretical Biology and Medical Modelling 7*(1).

[268] Nishiura, H., M. Schwehm, M. Kakehashi, and M. Eichner (2006). Transmission potential of primary pneumonic plague: time inhomogeneous evaluation based on historical documents of the transmission network. *Journal of Epidemiology and Community Health 60*(7), 640–645.

[269] Obadia, T., R. Haneef, and P.-Y. Boëlle (2012). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making 12*(1), 147.

[270] Ohmit, S., J. Victor, J. Rotthoff, E. Teich, R. Truscon, L. Baum, B. Rangarajan, D. Newton, M. Boulton, and A. Monto (2006). Prevention of antigenically drifted influenza by inactivated and live attenuated vaccines. *New England Journal of Medicine 355*(24), 2513–2522.

[271] Oluyomi-Obi, T., L. Avery, C. Schneider, A. Kumar, S. Lapinsky, S. Menticoglou, and R. Zarychanski (2010). Perinatal and maternal outcomes in critically ill obstetrics patients with pandemic H1N1 Influenza A. *J Obstet Gynaecol Can 32*(5), 443–447.

[272] Orenstein, E., G. De Serres, M. Haber, D. Shay, C. Bridges, P. Gargiullo, and W. Orenstein (2007). Methodologic issues regarding the use of three observational study designs to assess influenza vaccine effectiveness. *International Journal of Epidemiology 36*(3), 623–631.

[273] Orenstein, W. A., R. H. Bernier, T. J. Dondero, A. R. Hinman, J. S. Marks, K. J. Bart, and B. Sirotkin (1985). Field evaluation of vaccine efficacy. *Bulletin of the World Health Organization 63*(6), 1055.

[274] Orenstein, W. A., G. J. Nigel, G. De Serres, P. C. Farrington, S. B. Redd, and M. J. Papania (2004). Assessment of the status of measles elimination from reported outbreaks: United States, 1997–1999. *Journal of Infectious Diseases 189*(Supplement 1).

[275] Organization, W. H. et al. (2012). International classification of diseases (ICD).

[276] Ormerod, L. (2005). Multidrug-resistant tuberculosis (MDR-TB): epidemiology, prevention and treatment. *British Medical Bulletin 73*(1), 17–24.

[277] Osterhaus, A., G. Rimmelzwaan, B. Martina, T. Bestebroer, and R. Fouchier (2000). Influenza B virus in seals. *Science 288*(5468), 1051–1053.

[278] Paine, S., G. Mercer, P. Kelly, D. Bandaranayake, M. Baker, Q. Huang, G. Mackereth, A. Bissielo, K. Glass, and V. Hope (2010). Transmissibility of 2009 pandemic influenza A (H1N1) in New Zealand: effective reproduction number and influence of age, ethnicity and importations. *Euro Surveillance 15*(24), 9–17.

[279] Paneth, N. (2004). Assessing the contributions of John Snow to epidemiology: 150 years after removal of the broad street pump handle. *Epidemiology 15*(5), 514–516.

[280] Pareek, M., I. Baussano, I. Abubakar, C. Dye, and A. Lalvani (2012). Evaluation of immigrant tuberculosis screening in industrialized countries. *Emerging Infectious Diseases 18*(9), 1422.

[281] Pareek, M., J. P. Watson, L. P. Ormerod, O. M. Kon, G. Woltmann, P. J. White, I. Abubakar, and A. Lalvani (2011). Screening of immigrants in the UK for imported latent tuberculosis: a multicentre cohort study and cost-effectiveness analysis. *The Lancet Infectious Diseases 11*(6), 435–444.

[282] Patil, G. (1962). Certain properties of the generalized power series distribution II. *Annals of the Institute of Statistical Mathematics 14*(1), 179–182.

[283] Patrozou, E. and L. A. Mermel (2009). Does influenza transmission occur from asymptomatic infection or prior to symptom onset? *Public Health Reports 124*(2), 193.

[284] Pearson, T. A. et al. (1996). Alcohol and heart disease. *Circulation 94*(11), 3023–3025.

[285] Pebody, R., N. Andrews, J. McMenamin, H. Durnall, J. Ellis, C. Thompson, C. Robertson, S. Cottrell, B. Smyth, M. Zambon, et al. (2013). Vaccine effectiveness of 2011/12 trivalent seasonal influenza vaccine in preventing laboratory-confirmed influenza in primary care in the United Kingdom: evidence of waning intra-seasonal protection. *Euro Surveillance 18*(5).

[286] Pebody, R., C. Joseph, E. McLean, C. Hawkins, G. Kafatos, M. Catchpole, J. Van Tam, P. Kaye, J. Green, P. White, N. Phin, B. Evans, J. Watson, J. Ellis, A. Bermingham, A. Lackenby, G. Smith, S. Palmer, S. Malur Sudhanva, D. Brown, L. Miller, M. Zambon, J. McMenamin, C. Ramsay, O. Blatchford, D. Goldberg, J. Cowden, M. Donaghy, A. Eastaway, and B. Carmen (2009, May). Epidemiology of new influenza A(H1N1) in the United Kingdom, April-May 2009. *Euro Surveillance 14*(19).

[287] Pebody, R., E. McLean, H. Zhao, P. Cleary, S. Bracebridge, K. Foster, A. Charlett, P. Hardelid, P. Waight, J. Ellis, et al. (2010). Pandemic Influenza A (H1N1) 2009 and mortality in the United Kingdom: risk factors for death, April 2009 to March 2010. *Euro Surveillance 15*(20).

[288] Pedroni, E., M. Garcia, V. Espinola, A. Guerrero, C. Gonzalez, A. Olea, M. Calvo, B. Martorell, M. Winkler, M. Carrasco, et al. (2010). Outbreak of 2009 pandemic influenza A (H1N1), Los Lagos, Chile, April-June 2009. *Euro Surveillance 15*(1).

[289] Perez-Padilla, R., D. De La Rosa-zamboni, S. Ponce de Leon, M. Hernandez, F. Quiñones-Falconi, E. Bautista, A. Ramirez-Venegas, J. Rojas-Serrano, C. E. Ormsby, A. Corrales, et al. (2009). Pneumonia and respiratory failure from swine-origin influenza A (H1N1) in Mexico. *New England Journal of Medicine 361*(7), 680–689.

[290] Pinheiro, R. O., J. de Souza Salles, E. N. Sarno, and E. P. Sampaio (2011). Mycobacterium leprae-host-cell interactions and genetic determinants in leprosy: an overview. *Future Microbiology 6*(2), 217–230.

[291] Pitondo-Silva, A., A. C. B. Santos, K. A. Jolley, C. Q. F. Leite, and A. L. da Costa Darini (2013). Comparison of three molecular typing methods to assess genetic diversity for Mycobacterium tuberculosis. *Journal of Microbiological Methods 93*(1), 42–48.

[292] Potter, J., D. J. Stott, M. A. Roberts, A. G. Elder, B. O'Donnell, P. V. Knight, and W. F. Carman (1997). Influenza vaccination of health care workers in long-term-care hospitals reduces the mortality of elderly patients. *Journal of Infectious Diseases 175*(1), 1–6.

[293] Public Health England (2012, July). Drug resistant TB on the increase. http://www.hpa.org.uk/NewsCentre/NationalPressReleases/2012PressReleases/120705DrugresistantTBincrease/. Accessed 2013 Dec 11.

[294] Public Health England (2013a). Pneumococcal: the green book, chapter 25. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/263318/Green-Book-Chapter-25-v5_2.pdf. Accessed 2015 May 05.

[295] Public Health England (2013b). Tuberculosis (TB): annual notification and mortality data (1913 onwards). https://www.gov.uk/government/statistics/tuberculosis-tb-annual-notifications-1913-onwards. Accessed 2015 May 18.

[296] Public Health England (2013c). UK TB Cluster Naming Resource. http://www.hpa-bioinformatics.org.uk/TBCluster/tbhome.php. Accessed 2015 May 18.

[297] Public Health England (2014a, August). Communicable Disease Outbreak Management: Operational Guidance. https://www.gov.uk/government/publications/communicable-disease-outbreak-management-operational-guidance. Accessed 2015 June 08.

[298] Public Health England (2014b). Influenza: the green book, chapter 19. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/385226/Green_Book_Chapter_19_v8_2.pdf. Accessed 2015 May 10.

[299] Public Health England (2014c, February). TB Strain Typing and Cluster Investigation Handbook 3rd Edition.

[300] Public Health England (2014d, November). Tuberculosis. https://www.gov.uk/government/collections/tuberculosis-and-other-mycobacterial-diseases-diagnosis-screening-management-and-data. Accessed 2015 July 08.

[301] Qian, L., H. Tseng, L. Sy, and S. Jacobsen (2012). Confounder adjustment in vaccine safety studies: Comparing three offset terms for case-centered approach. *Vaccine*.

[302] R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

[303] Ramsay, M., J. McVernon, N. Andrews, P. Heath, and M. Slack (2003). Estimating Haemophilus influenzae type b vaccine effectiveness in England and Wales by use of the screening method. *Journal of Infectious Diseases 188*(4), 481–485.

[304] Raviglione, M. C. and I. M. Smith (2007). XDR tuberculosis-implications for global public health. *New England Journal of Medicine 356*(7), 656–659.

[305] Rehm, J., A. V. Samokhvalov, M. G. Neuman, R. Room, C. Parry, K. Lönnroth, J. Patra, V. Poznyak, and S. Popova (2009). The association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic review. *BMC Public Health 9*(1), 450.

[306] Remschmidt, C., T. Rieck, B. Bödeker, and O. Wichmann (2015). Application of the screening method to monitor influenza vaccine effectiveness among the elderly in Germany. *BMC Infectious Diseases 15*(1), 1.

[307] Richter, E., M. Weizenegger, S. Rüsch-Gerdes, and S. Niemann (2003). Evaluation of genotype MTBC assay for differentiation of clinical Mycobacterium tuberculosis complex isolates. *Journal of Clinical Microbiology 41*(6), 2672–2675.

[308] Ripley, B. D. (2002). *Modern applied statistics with S.* Springer.

[309] Roberts, M. G. and H. Nishiura (2011). Early estimation of the reproduction number in the presence of imported cases: pandemic influenza H1N1-2009 in New Zealand. *PLoS One 6*(5).

[310] Roetzer, A., R. Diel, T. A. Kohl, C. Rückert, U. Nübel, J. Blom, T. Wirth, S. Jaenicke, S. Schuback, S. Rüsch-Gerdes, et al. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Medicine 10*(2).

[311] Rondy, M., J. Puig-Barbera, O. Launay, X. Duval, J. Castilla, M. Guevara, S. Costanzo, K. de Gaetano Donati, and A. Moren (2013). 2011–12 Seasonal Influenza Vaccines Effectiveness against Confirmed A (H3N2) Influenza Hospitalisation: Pooled Analysis from a European Network of Hospitals. A Pilot Study. *PloS One 8*(4).

[312] Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

[313] Rosenbaum, P. and D. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 516–524.

[314] Rosenbaum, P. and D. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39*(1), 33–38.

[315] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association 82*(398), 387–394.

[316] Roush, S. W., T. V. Murphy, V.-P. D. T. W. Group, et al. (2007). Historical comparisons of morbidity and mortality for vaccine-preventable diseases in the United States. *JAMA 298*(18), 2155–2163.

[317] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, Volume 81. John Wiley & Sons.

[318] Rückinger, S., M. Van der Linden, R. R. Reinert, and R. Von Kries (2010). Efficacy of 7-valent pneumococcal conjugate vaccination in Germany: An analysis using the indirect cohort method. *Vaccine 28*(31), 5012–5016.

[319] Rupert Jr, G. et al. (2012). *Simultaneous statistical inference*. Springer Science & Business Media.

[320] Sa, M. and A. Rm (2014, May). An evaluation of the indirect cohort method to estimate the effectiveness of the pneumococcal polysaccharide vaccine. *Journal of Vaccines & Immunization 2*(1), 4–6.

[321] Salathé, M., C. C. Freifeld, S. R. Mekaru, A. F. Tomasulo, and J. S. Brownstein (2013). Influenza A (H7N9) and the importance of digital epidemiology. *The New England Journal of Medicine 369*(5), 401.

[322] Salisbury, D., M. Ramsay, J. White, and D. Brown (1997). Polio eradication: surveillance implications for the United Kingdom. *Journal of Infectious Diseases 175*(Supplement 1).

[323] Sanderson, S., I. D. Tatt, and J. P. Higgins (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology 36*(3), 666–676.

[324] Sartwell, P. E. (1995). The distribution of incubation periods of infectious disease. 1949. *American Journal of Epidemiology 141*(5), 386–94.

[325] Schatz, A. and S. A. Waksman (1944). Effect of Streptomycin and Other Antibiotic Substances upon Mycobacterium tuberculosis and Related Organisms. *Experimental Biology and Medicine 57*(2), 244–248.

[326] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

[327] Schwoebel, V., D. Antoine, and J. Veen (1999). Surveillance of tuberculosis in Europe. *Medicinski Arhiv 53*(3 Suppl 1), 9.

[328] Scottish Executive (2009). Scottish Index of Multiple Deprivation 2009 Technical Report. *Edinburgh Scottish Executive*.

[329] Scottish Executive (2015). Health of Scotland's population - Smoking. `http://www.gov.scot/Topics/Statistics/Browse/Health/TrendSmoking`. Accessed 2016 Mar 09.

[330] Scottish Government (2011). A TB Action Plan for Scotland. Technical report, Edinburgh.

[331] Seagar, L. and I. F. Laurenson (2012, October). MIRU-VNTR typing of M.tuberculosis complex isolates. [Presentation given by the Scottish Mycobacteria Reference Laboratory].

[332] Shorten, R., A. McGregor, S. Platt, C. Jenkins, M. Lipman, S. Gillespie, B. Charalambous, and T. McHugh (2013). When is an outbreak not an outbreak? Fit, divergent strains of Mycobacterium tuberculosis display independent evolution of drug resistance in a large London outbreak. *Journal of Antimicrobial Chemotherapy 68*(3), 543–549.

[333] Simona, L. and T. MIHAESCU (2013). History of BCG Vaccine. *Mædica 8*(1), 53.

[334] Simonsen, L., R. Taylor, C. Viboud, M. Miller, and L. Jackson (2007). Mortality benefits of influenza vaccination in elderly people: an ongoing controversy. *The Lancet*

[342] Smith, D. J., A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science 305*(5682), 371–376.

[343] Smith, G. J., D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani, S. Bhatt, et al. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature 459*(7250), 1122–1125.

[344] Smith, R. D. (2006). Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. *Social science & medicine 63*(12), 3113–3123.

[345] Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ 338*.

[346] Stone, S. P., C. Fuller, J. Savage, B. Cookson, A. Hayward, B. Cooper, G. Duckworth, S. Michie, M. Murray, A. Jeanes, et al. (2012). Evaluation of the national Cleanyourhands campaign to reduce Staphylococcus aureus bacteraemia and Clostridium difficile infection in hospitals in England and Wales by improved hand hygiene: four year, prospective, ecological, interrupted time series study. *BMJ 344*.

[347] Stroup, D. F., R. Lyerla, et al. History of statistics in public health at CDC 1960–2010: the rise of statistical evidence.

[348] Sudfeld, C. R., A. M. Navar, and N. A. Halsey (2010). Effectiveness of measles vaccination and vitamin A treatment. *International Journal of Epidemiology 39*(suppl 1).

[349] Supply, P., E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, and C. Locht (2000). Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. *Molecular Microbiology 36*(3), 762–771.

[350] Surveillance Group for New Influenza A(H1N1) Virus Investigation and Control in Spain (2009, May). New influenza A(H1N1) virus infections in Spain, April-May 2009. *Euro Surveillance 14*(19).

[351] Sussman, J. and R. Hayward (2010). Using instrumental variables to adjust for treatment contamination in randomised controlled trials. *British Medical Journal 340*, 1181–1184.

[352] Svensson, Å. (2007). A note on generation times in epidemic models. *Mathematical Biosciences 208*(1), 300–311.

[353] Talbot, H., M. Griffin, Q. Chen, Y. Zhu, J. Williams, and K. Edwards (2011). Effectiveness of seasonal vaccine in preventing confirmed influenza-associated hospitalizations in community dwelling older adults. *Journal of Infectious Diseases 203*(4), 500.

[354] Taylor, B., E. Miller, C. Farrington, M.-C. Petropoulos, I. Favot-Mayaud, J. Li, and P. A. Waight (1999). Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *The Lancet 353*(9169), 2026–2029.

[355] Therneau, T. M. and P. M. Grambsch (2000). *Modeling survival data: extending the Cox model.* Springer.

[356] Thomas, Y., G. Vogel, W. Wunderli, P. Suter, M. Witschi, D. Koch, C. Tapparel, and L. Kaiser (2008). Survival of influenza virus on banknotes. *Applied and Environmental Microbiology 74*(10), 3002–3007.

[357] Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, N. Cox, L. J. Anderson, and K. Fukuda (2003). Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA 289*(2), 179–186.

[358] Török, M. E., S. Reuter, J. Bryant, C. U. Köser, S. V. Stinchcombe, B. Nazareth, M. J. Ellington, S. D. Bentley, G. P. Smith, J. Parkhill, et al. (2013). Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. *Journal of Clinical Microbiology 51*(2), 611–614.

[359] Turner, P. J., J. Southern, N. J. Andrews, E. Miller, M. Erlewyn-Lajeunesse, S. S. Investigators, et al. (2015). Safety of live attenuated influenza vaccine in atopic children with egg allergy. *Journal of Allergy and Clinical Immunology*.

[360] UNAIDS (2014). UNAIDS report on the global AIDS epidemic 2013. `http://www.unaids.org/sites/default/files/en/media/unaids/contentassets/`

`documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf`. Accessed 2015 May 04.

[361] Vach, W. and M. Blettner (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology 134*(8), 895–907.

[362] Valenciano, M., E. Kissling, B. Ciancio, and A. Moren (2010). Study designs for timely estimation of influenza vaccine effectiveness using European sentinel practitioner networks. *Vaccine 28*(46), 7381–7388.

[363] van Soolingen, D., P. Hermans, P. De Haas, D. Soll, and J. Van Embden (1991). Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *Journal of Clinical Microbiology 29*(11), 2578–2586.

[364] Vandenbroucke, J. P. (2008). Observational research, randomised trials, and two views of medical science. *PLoS Medicine 5*(3).

[365] Velji, P., V. Nikolayevskyy, T. Brown, and F. Drobniewski (2009). Discriminatory ability of hypervariable variable number tandem repeat loci in population-based analysis of Mycobacterium tuberculosis strains, London, UK. *Emerging Infectious Diseases 15*(10), 1609.

[366] Venzon, D. and S. Moolgavkar (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 87–94.

[367] Voordouw, B., P. van der Linden, S. Simonian, J. van der Lei, M. Sturkenboom, and B. Stricker (2003). Influenza vaccination in community-dwelling elderly: impact on mortality and influenza-associated morbidity. *Archives of Internal Medicine 163*(9), 1089.

[368] Vu, T., S. Farish, M. Jenkins, and H. Kelly (2002). A meta-analysis of effectiveness of influenza vaccine in persons aged 65 years and over living in the community. *Vaccine 20*(13), 1831–1836.

[369] Vynnycky, E. and P. E. Fine (2000). Lifetime risks, incubation period, and serial interval of tuberculosis. *American Journal of Epidemiology 152*(3), 247–263.

[370] Walker, T. M., C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, et al. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases 13*(2), 137–146.

[371] Wallinga, J. and M. Lipsitch (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences 274*(1609), 599–604.

[372] Wallinga, J. and P. Teunis (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology 160*(6), 509–516.

[373] Wallinga, J., P. Teunis, and M. Kretzschmar (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology 164*(10), 936–944.

[374] Wallinga, J., M. van Boven, and M. Lipsitch (2010). Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences 107*(2), 923–928.

[375] Walsh, C. et al. (2003). *Antibiotics: actions, origins, resistance.* American Society for Microbiology (ASM).

[376] Wang, Y., M. A. Beydoun, L. Liang, B. Caballero, and S. K. Kumanyika (2008). Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. *Obesity 16*(10), 2323–2330.

[377] Ward, P., I. Small, J. Smith, P. Suter, and R. Dutkowski (2005). Oseltamivir (Tamiflu®) and its potential for use in the event of an influenza pandemic. *Journal of Antimicrobial Chemotherapy 55*(suppl 1).

[378] Watkins, R. and A. Plant (2006). Does smoking explain sex differences in the global tuberculosis epidemic? *Epidemiology and Infection 134*(02), 333–339.

[379] Watson, H. W. and F. Galton (1875). On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland 4*, 138–144.

[380] Watson, R. (2007). Antibiotic resistant microbes biggest threat to European health. *BMJ: British Medical Journal 334*(7605), 1187.

[381] Weinberger, D. M., R. Malley, and M. Lipsitch (2011). Serotype replacement in disease after pneumococcal vaccination. *The Lancet 378*(9807), 1962–1973.

[382] White, L. F., J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, and M. Pagano (2009). Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza and Other Respiratory Viruses 3*(6), 267–276.

[383] WHO. Tuberculosis Fact Sheet. `http://www.who.int/mediacentre/factsheets/fs104/en/`. Accessed 2013 Dec 02.

[384] WHO Ebola Response Team (2014). Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. *New England Journal of Medicine 371*(16), 1481–95.

[385] WHO, Joint and Consultation, FAO Expert (2003). Diet, nutrition and the prevention of chronic diseases. *WHO Technical Report Series* (916), 1–60.

[386] Wichmann, O., P. Stocker, G. Poggensee, D. Altmann, D. Walter, W. Hellenbrand, G. Krause, and T. Eckmanns (2010). Pandemic influenza A (H1N1) 2009 breakthrough infections and estimates of vaccine effectiveness in Germany 2009-2010. *Euro Surveillance 15*(18), 19561.

[387] Williams, D. (1991). *Probability with martingales*. Cambridge university press.

[388] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association 22*(158), 209–212.

[389] Wong, K., M. Campitelli, T. Stukel, and J. Kwong (2012). Estimating Influenza Vaccine Effectiveness in Community-Dwelling Elderly Patients Using the Instrumental Variable Analysis Method. *Archives of Internal Medicine*.

[390] Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

[391] World Health Organisation (1980). A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bulletin of the World Health Organization 58*, 585–591.

[392] World Health Organisation (2015). World Malaria Report 2014.

[393] World Health Organization (2011). Recommended composition of influenza virus vaccines for use in the 2011-2012 northern hemisphere influenza season. Technical report, Geneva: World Health Organization. URL `http://www.who.int/influenza/vaccines/2011_02_recommendation.pdf`.

[394] World Health Organization (2012a). Global tuberculosis control: WHO report 2012. Technical report, Geneva: World Health Organization.

[395] World Health Organization (2012b). Recommended composition of influenza virus vaccines for use in the 2012-2013 northern hemisphere influenza season. Technical report, Geneva: World Health Organization. URL `http://www.who.int/influenza/vaccines/virus/recommendations/201202_recommendation.pdf`.

[396] Wu, S., L. Shen, and G. Liu (1999). Study on vertical transmission of Chlamydia trachomatis using PCR and DNA sequencing. *Chinese Medical Journal 112*(5), 396–399.

[397] Yang, D. and J. Dalton (2012). A unified approach to measuring the effect size between two groups using SAS. *SAS Global Forum 2012: Statistics and Data Analysis*.

[398] Yoo, B. and K. Frick (2006). The instrumental variable method to study self-selection mechanism: a case of influenza vaccination. *Value in Health 9*(2), 114–122.

[399] Zhang, J. and F. Y. Kai (1998). What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA 280*(19), 1690–1691.

[400] Ziegler, P. (2013). *The black death.* Faber & Faber.