Hindcasting trends of infection

using crossectional diagnostic test data

Gustaf Rydevik

Doctor of Philosophy
University of York

Environment

June 2015

ABSTRACT

Infectious diseases are a major threat to the wellbeing of humans, livestock, and wildlife. However, there is often a paucity of information for responding to these threats, and thus a need to make efficient use of existing data. This thesis shows how to use Bayesian analysis to maximise the information gained from already collected diagnostic test data.

First, the commonly used latent class analysis of multiple binary diagnostic tests is extended to account for vaccinated individuals, and used to estimate the effect of study size on sensitivity and specificity estimates of DIVA (”Distinguishing Infected and Vaccinated Animals”) tests for bovine Tuberculosis.

It is then shown how quantitative test responses can be used as clocks indicating the time since infection to “hindcast” historic trends of disease incidence using cross-sectional data. This is used to determine whether an endemic disease is increasing or decreasing up to the time of sampling, enabling the tracking of trends in populations where routine surveillance data is not available.

It is further demonstrated how to hindcast the rise and fall of disease outbreaks. Using the 2007 UK Bluetongue virus outbreak and a whooping cough outbreak as examples, it is shown that hindcasting can be used to determine whether an outbreak is increasing or past its peak at the time of sampling, thus informing potential outbreak responses.

In the light of these methods for analysing quantitative test data, the challenges of generating data on test kinetics are discussed. Suggestions are given for how to improve on current methods by modelling the development of paired diagnostic tests as a dynamic host-pathogen system.

This thesis demonstrates that multiple quantitative tests can be used to recover disease trends in a population. These methods have far-reaching consequences for the design and practice of disease surveillance in all contexts.

Till min älskade farmor

. . .
To have a chance to learn and
to grow
to be skillful in your
profession or craft
practicing the precepts and
loving speech -
this is the greatest happiness.
. . .

Mahamangala Sutta,
Sutta Nipata 2.4

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

Thanks to my ″Fantastic Five″ group of supervisors: Glenn Marion, Giles Innocent, Ross Davidson, Piran White, and Mike Hutchings. Glenn, for your permanent smile, your patience despite a clash in cultures - Swede vs Brit and statistician vs modeller - and for helping me express the methods and the math in a coherent and logical manner. Giles, for your never-ending positive support, your thesis coaching, and for the talks we've had about all sorts of statistics and life questions. Ross, for asking the technical questions, and being calm and persistent in the face of my perennial stubbornness. Piran, for the way you demonstrate that one can be a successful academic as well as full of joy, for pointing out the holes in my sloppy writing (any remaining holes are my own fault!), and for always pushing me to go back and find references in the literature - a habit I hope to keep in the future! Mike, for always asking for the "bigger picture", and for making me understand and explain what my work really means.

Thanks to EU framework programme 7 and the WildTech project for providing the funding for this PhD, and to all the WildTech scientists I've met over the past four years - Babbis, Duncan, Dolores, Debra, George, Jonas, Lisa, Peter, Tim, and everyone else!

Building an infrastructure for a life far away from home has been crucial to enjoying this experience and Edinburgh to the extent that I have. Thanks to my office mates over the years for chats, support and laughs - Naomi, Laura, Max, and Katie! Thanks to Promoe and Burial for producing the soundtrack of this thesis, fuelling countless hours of writing and working. Thanks to Artisan Roast for creating an oasis of perfect coffee, caffeinating those morning sessions.

Thanks to Edinburgh Aikido club for providing a place to channel my frustrations and feel the joy of being tossed on a green mat, and thanks to Wild Geese Sangha for the opportunities for peace and reflections.

Ett stort tack till mina kära föräldrar, som har stöttat och hejat på!

Thanks to Lisa Francis-Jennings, project manager extraordinaire, for coaching me during the final push. Thanks to Molly Uzzell for offering to copy-edit, massively improving the quality of my writing. Thanks to Linda Arvedal for sharing my life and supporting me during the first half of this PhD, and for dealing with the long flights and the stressed-out me. And finally, thanks to Ainsley Francis, for giving me joy and for sharing my life these final years, despite me being alternatively absorbed in work, frustrated and grumpy, or exhausted. Soon, you'll be able to meet the hopefully more rested and joyful Dr Gustaf!

AUTHOR'S DECLARATION

Chapter four is a reproduction of the paper
Using combined diagnostic tests to Hindcast trends of infection from cross-sectional data
by Gustaf Rydevik, Giles T. Innocent, Glenn Marion, Ross S. Davidson,Piran C. L. White, Charalambos Billinis, Paul Barrow, Peter P. C. Mertens, Dolores Gavier-Widén, and Michael R. Hutchings. The paper was submitted to *PloS Computational Biology* on the 18th of June, 2015. For this paper, I developed the methodology, ran the experiments, and generated the figures. I was also the main contributor of the writing of the paper. I declare that, with the exception of chapter four, the work contained in this thesis is my own, and has not been submitted for any other degree or award at this, or any other, University.

CHAPTER 1: INTRODUCTION AND OVERVIEW

On the 17th of April 2009 the Mexican Institute for Social Security sounded an epidemiological alert regarding an unusual pattern of reported influenza cases (Echevarría-Zuno et al., 2009). Later studies suggested this outbreak started in February of 2009 (Echevarría-Zuno et al., 2009) or possibly earlier (Smith et al., 2009). On April 23rd, the cause was identified as a novel influenza virus with serotype H1N1 (Girard et al., 2010) and by the end of April, cases infected with the identical strain had been reported in several different countries (Fineberg, 2014). The presented symptoms in this flu outbreak were severe: of 899 patients hospitalised in Mexico, 6.5% suffered critical illness; 41% of these critically ill patients died.

In the following months and for most of 2009, this novel strain, spread from country to country under the moniker "swine flu".

On the 11th of June 2009, WHO officially classified H1N1 as a pandemic flu (Chan, 2009), thereby triggering emergency responses by governments across the world.

Despite the severity of initial cases, the pandemic H1N1 flu was soon found to have similar morbidity to regular seasonal flu (Tuite et al., 2010). In light of this, the responses by WHO and others has been criticised (Durodié, 2011; Flynn, 2010), but most official reports have subsequently concluded that the response was appropriate given the data available at the time (Fineberg, 2014; Leung and Nicoll, 2010).

The case of the H1N1 outbreak highlights the difficulty of making decisions in the face of an infectious disease outbreak, and the importance of having enough information available so that action can be taken proportionately to the true rather than perceived threat. Disease surveillance can tell us what pathogens occur at what time and in which hosts and environments. It is thus a critical link in the chain between awareness and action. However, surveillance data is often patchy and incomplete. Statistical techniques can contribute substantially to the interpretation of these data, and in-depth statistical analyses make it possible to maximise the information gained. The statistical frameworks developed in this thesis have implications for and relevance to multiple aspects of the interactions between diagnosing disease, developing new diagnostics and designing surveillance systems and appropriate control policies.

The specific types of information required differ depending on the epidemiology of the disease. Section 1.1 discusses different types of disease epidemiology, and how the goals of surveillance differ depending on the pattern of spread. Section 1.2 provides a brief disease description of how surveillance systems function, and typical gaps and limitations in performance. Section 1.3 focusses on diagnostic tests and how more comprehensive analyses of collected test results could help mitigate key gaps in the performance of existing surveillance systems. Section 1.4 highlights major current and historical uses of statistical analysis in disease surveillance. Section 1.5 introduces key notations and the foundations necessary for conducting the statistical analyses used in this thesis. Finally, section 1.6 describes the focus of the thesis, and lays out the structure of the following chapters.

1.1 CHARACTERISING PATTERNS OF INFECTIOUS DISEASE

The epidemiology of infectious diseases vary greatly in the particular. In general, however, there are a few categories of epidemiological patterns that are often used for describing diseases. These are not distinct classes, but describe different aspects of observed patterns of spread.

1.1.1 *The basic reproduction number R_0*

A natural starting point is to consider the infectiousness of a pathogen. This can be quantified by defining the seemingly theoretical concept of the "basic reproduction number", or R_0 as the average number of new cases infected by a single original case in a large host population in which every individual is susceptible to infection. Intuitively, it is clear that if R_0 is less than 1, i.e. if each case infects on average less than a single new host before recovering (i.e. ceasing to be infectious, dying or leaving the population under study), the pathogen will die out. If $R_0 > 1$, the pathogen will instead spread exponentially. As the pathogen spreads through the population the number of susceptible individuals declines and the number of cases infected by each new infective (the R number) is lower than R_0. The development of the R number over time determines the type of epidemiological pattern the pathogen follows.Observed values of R_0 vary tremendously across infectious pathogens; studies have indicated that in a totally susceptible population, a single measles case can infect (i.e. have an R_0 of) between 12 and 18 (Fine, 1993), while for scrapie, R_0 has been estimated to be 3.9 (Heffernan et al., 2005).

1.1.2 *Endemic and epidemic disease*

The terms endemic and epidemic are frequently used to characterize disease. However, the terms are not exclusive and depend on the spatial and temporal scale being considered.

An *endemic pathogen* is a pathogen that is continually present in a population. This implies that the pathogen has reached a (possibly dynamic) steady state, where an individual infected with the pathogen will on average spread the pathogen to one other individual before recovering or dying (i.e. $R \approx 1$). Note that this says nothing about the absolute prevalence (i.e. proportion of the population infected) or incidence (number of individuals infected per time unit) of the pathogen. The incidence of new infections could exhibit a relatively stable trend in the population if the pathogen spreads slowly and/or is chronic. Alternatively, it could show a regular or irregular pattern of fluctuations over a period of time, as the infectiousness and/or the number of susceptible individuals change or simply due to stochastic variation in disease contact. One example of such a disease is Nephropatia epidemica, caused by the *Puumala* virus (PUUV) (Family *Bunyaviridae*, genus *hantavirus*). It is a zoonotic pathogen that has voles as its main host, but has an annual season of a few months in winter when the voles are forced inside buildings, and the pathogen can infect humans, causing seasonal epidemics (Zeimes et al., 2012). That the pathogen is endemic does not mean that the average incidence is stable; if the point of equilibrium increase or decrease, so does the incidence or prevalence. In this context a major concern is the impact of climate change on infectious pathogens, where it is feared that increasing temperatures will lead to higher rates of transmission (e.g by increasing the numbers of disease vectors such as ticks or mosquitos (McMichael et al., 2006)), as well as a shift in the endemic regions for a wide range of pathogens, leading to changes in the incidence for many pathogens in both in humans and animals (Fox et al., 2011).

An *epidemic pathogen* is characterised by the change in incidence being dominated by the dynamics of the pathogen spread, rather than changes in the equilibrium point. This could mean a classic outbreak scenario, such as the rapid spread of Norovirus (winter vomiting disease) in a naive population aboard cruise ships after food contamination (Isakbaeva et al., 2005). But it could also refer to a pathogen such as HIV, where the number of cases are still increasing, only lately showing some signs of having reached equilibrium (Nagelkerke et al., 2014); in this case, the increase is slow and steady, but undeniably positive, and only rarely exhibits fluctuations.

As noted above the terms "epidemic" and "endemic" are complementary, not in contradiction. A pathogen could be endemic and epidemic at the same time. Such pathogens tend to circulate in a population (or in a nearby reservoir host) at a low level, causing occasional outbreaks once some epidemic threshold is reached. Within the wider population the pathogen would be technically endemic but may remain undetected at low levels, only being detected during 'epidemic' outbreaks. Alternatively, if the pathogen remains in a reservoir population (where it is endemic) it may cause true epidemic outbreaks in connected and completely susceptible populations. A classic example of this is the epidemiology of measles in the UK pre-vaccination. At some scale, it would always be present, causing occasional cases in the population. On a regular basis, however, the number of children who had not been exposed would increase to a critical level, and a measles epidemic would occur. Interestingly, these epidemics nearly always started in London before spreading to the rest of the country, probably because high population density meant that a critical mass was reached earlier than in other cities (Grenfell et al., 2001).

The terms *emerging disease* or *reemerging disease* have been increasingly used, in particular in connection with climate change and some of the high-profile outbreaks seen in the last decades. The term is usually used to refer to diseases caused by pathogens that have a high risk of being introduced or have recently been introduced in a population, but where there is no (recent) history of occurrence in that particular population. Many examples of emerging diseases are arthropod-borne pathogens; Bluetongue (Wilson and Mellor, 2009), Chikungunya (Thiboutot et al., 2010), and West Nile Virus (Bernard et al., 2001) are all spread via arthropod bites. Since the range and population densities of insects are highly affected by micro- and macro- climate, a changing climate can be expected to have a substantial effect on the locations and populations where arthropod vectors are to be found (Daszak et al., 2000). An emerging (re-emerging) disease can also refer to diseases that have been present in the past but were either not discovered or did not cause problems until some recent change in circumstances increased the incidence or the consequences of infection. The foot-and-mouth disease outbreak in 2001 in the UK (Scudamore and Harris, 2002) can be considered as a classic example of both an epidemic and an emerging disease of concern for northern Europe, while in many parts of the world, such as sub-saharan Africa, foot-and-mouth disease is endemic (Bronsvoort et al., 2004).

1.2 THE IMPLEMENTATION OF SURVEILLANCE SYSTEMS

Disease surveillance can be defined as the systematic collection, analysis and dissemination of disease information with the aim of informing action towards the management of the disease (CDC, 2001). It can thus refer to any kind of data collection; covering an entire population or just a representative sample, aggregated or individual-based information, the registration of laboratory-confirmed cases or monitoring high concentrations of pathogens that are likely to cause disease. Given the above definition, it is crucial that collected data are used to inform a suitable course of action with regards to a particular disease or group of diseases.

The purpose of disease surveillance may differ depending on the epidemiological pattern of the pathogen of interest. For endemic pathogens, the role of disease surveillance is to keep track of the trend (for example changing incidence due to climate change), to understand the burden of disease (Kosek et al., 2003; Parashar et al., 2003), the epidemiology of the pathogen (Ali et al., 2002), and to evaluate the effectiveness of any implemented control measures. Infections might be underreported, which does not affect trend estimates as long as the proportion of cases being reported remains stable, but can cause problems for comparisons over time, between regions or between different surveillance systems. For epidemic diseases, the role of disease surveillance is to catch an outbreak as early as possible (Chan et al., 2010), provide information from which to act to to limit its spread (Zhang et al., 2013), and help in the process of understanding the likely cause of the outbreak (Communications, 2006) . It is usually (at least in the western world) the case that there are existing surveillance systems and routines for reporting cases and tracking the incidence curves. However, those surveillance systems might well have substantial reporting delay, which can limit their usefulness for rapid response. Depending on the disease, it is more or less likely that surveillance systems will fail to detect cases; some, like measles, have clear and relatively severe symptoms, while others such as chickenpox will probably not go unnoticed but might go unreported as they usually have mild symptoms. For truly emerging diseases i.e. those that have not been experienced previously anywhere, we know know almost nothing initially; there has been no opportunity for extensive study, and while it is possible to use disease models to predict behaviour uncertainty in any predictions is large. In this scenario, the role of disease surveillance is first to detect the initial incursion, and second to track development as the disease spreads (or not) in the community (Chan et al., 2010). Both of these are crucial in informing decisions as to what measures are appropriate, enabling an effective response to be mounted.

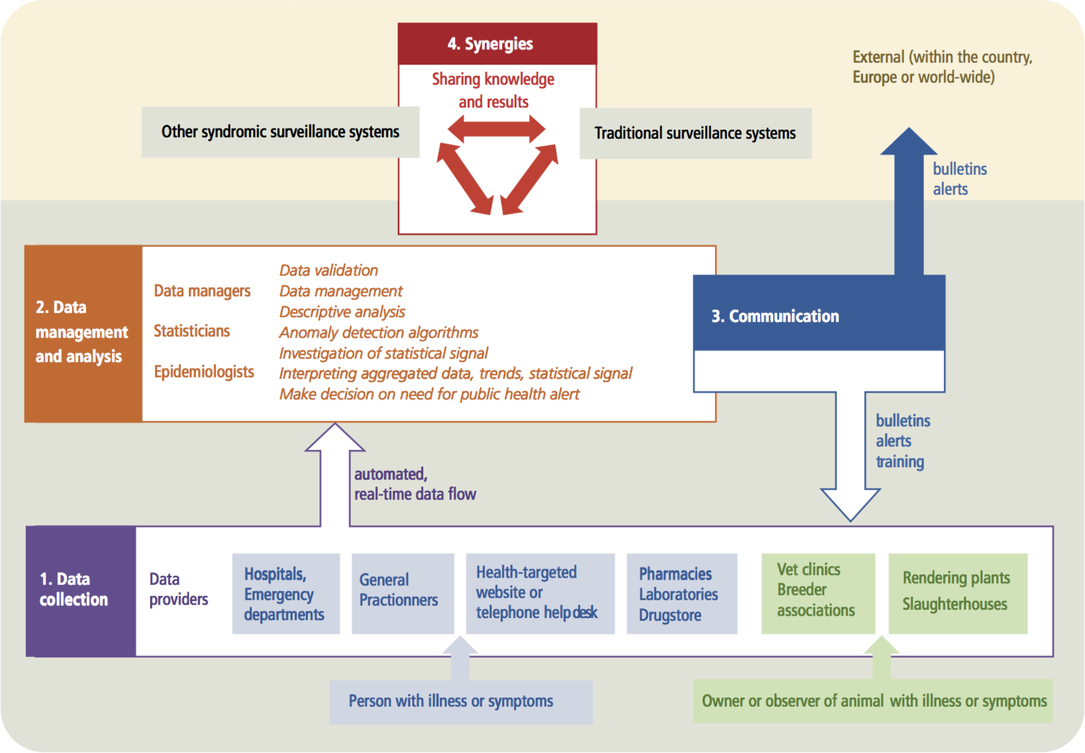
4. Synergies
Sharing knowledge and results
Other syndromic surveillance systems
Traditional surveillance systems
External (within the country, Europe or world-wide)
bulletins
alerts
2. Data management and analysis
Data managers
Statisticians
Epidemiologists
Data validation
Data management
Descriptive analysis
Anomaly detection algorithms
Investigation of statistical signal
Interpreting aggregated data, trends, statistical signal
Make decision on need for public health alert
3. Communication
bulletins
alerts
training
automated, real-time data flow
1. Data collection
Data providers
Hospitals, Emergency departments
General Practionners
Health-targeted website or telephone help desk
Pharmacies Laboratories Drugstore
Vet clinics Breeder associations
Rendering plants Slaughterhouses
Person with illness or symptoms
Owner or observer of animal with illness or symptoms

Figure 1.1: *Generic structure of a disease surveillance system, showing the various tasks, stakeholder, and flows of information that needs to come together in a functioning surveillance system, whether based on diagnosed cases or registration of syndromic indicators. Figure reproduced from Triple S Consortium (2013) .*

However, no matter the objective, all surveillance systems are composed of several different tasks linked together. Figure 1.1 highlights the task flow that needs to be successfully implemented to form a functioning surveillance system, using the surveillance of wildlife disease as an example. This starts with the occurrence of a case, follows through to the case being recorded, and then being (correctly!) communicated to statisticians and epidemiologists. Thus, the number of different people and organizations that are involved in producing disease information is large and susceptible to miscommunications and errors. While the idea of taking note of infected individuals can be traced back to the 1300s, the resources for coordinating systematic population-level registration of deaths and cases of illness only became available in the 19th century; in the UK, the first disease registry was introduced in 1838 by William Farr (Langmuir, 1976). The concept of surveillance as it is known today, in which the informing of action is a critical aspect, was only set out in 1968 by the WHO technical assembly (Declich and Carter, 1994). Computerization and automation has made the implementation of surveillance systems easier and faster, but setting up large-scale surveillance programs is still a challenge requiring substantial expertise and capital (CDC, 2001).

While it is easy to assume that data collection is a simple and painless process, in reality it is one of the major hurdles within surveillance, and worth describing in more detail. Starting from the occurrence of infection, a chain of events need to occur before a diagnosis is recorded in the database. First, the individual needs to either show symptoms that are severe enough to be noticed, or be sampled at random as part of a screening program. For many pathogens, a substantial number of infected hosts exhibit mild or subclinical symptoms, and are therefore not captured by passive surveillance. Subclinical infections, while of limited consequence for the individ-

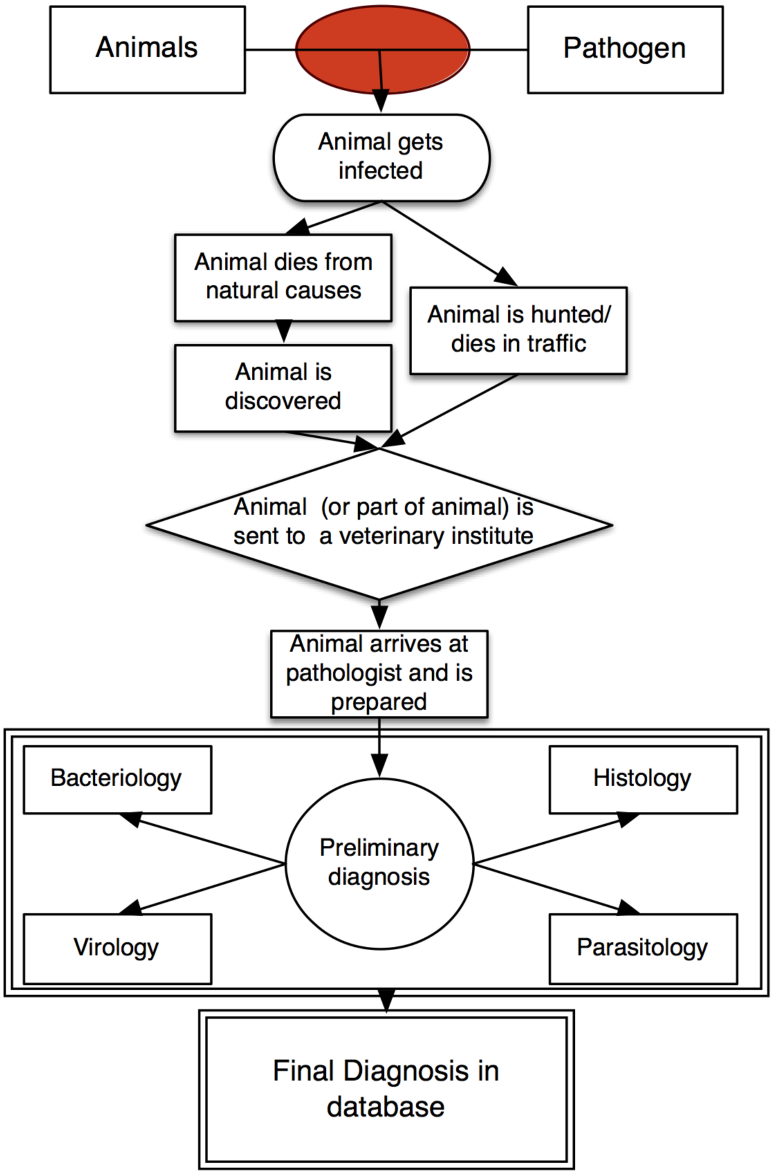
Animals
Pathogen
Animal gets infected
Animal dies from natural causes
Animal is hunted/ dies in traffic
Animal is discovered
Animal (or part of animal) is sent to a veterinary institute
Animal arrives at pathologist and is prepared
Bacteriology
Histology
Preliminary diagnosis
Virology
Parasitology
Final Diagnosis in database

Figure 1.2: *Flowchart of the process from infection of a wild animal to the registration of a case in a database. After personal communication with D. Gavier-Widen, Head of Pathology, Swedish National Veterinary Institute.*

ual, can be crucial for understanding the pattern of spread and the epidemiology of disease. Figure 1.2 show a schematic of the various phases that a wildlife infection goes through from occurrence up until the infection is registered in a database (D. Xavier-Widen, personal communication).

After a case has been recorded, the doctor or veterinarian needs to go through the effort of establishing a diagnosis, often by taking samples that are sent for testing at a regional biomedical veterinary or health care laboratory. Depending on available resources, including time and interest of the veterinarian or doctor, this might only happen for particularly unusual or severe cases. Once a sample is analysed, the diagnostic laboratory procedures used need to be sensitive enough to correctly establish the diagnosis. Depending on the pathogen, this can be more or less difficult, and choosing the optimal test would usually require that the referring doctor or veterinarian has a suspicion of which pathogen to look for. Some pathogens are predominantly present in a particular part of the body; as an example, Bovine TB in cattle is only present in localized encapsulated lesion, and the bacterium is unlikely to be found in other tissue samples (OIE, 2009). Such pathogens would thus require samples to be taken from the correct tissues in order for the presence of the pathogen to be detected. For wildlife diseases in particular, where animals are received without knowledge of their clinical history, it is often a substantial detective effort to identify the correct causative agent(s), and there is little doubt that a lot of infections go undetected (Mörner et al., 2002).

The net effect of this chain of events is to create a pyramid of cases (see figure 1.3, reproduced from Gibbons et al. (2014)). This can lead to the "iceberg" phenomenon of disease (Last, 1963). For many pathogens, only the very tip of the iceberg is detected and reported, the cases that are most severe and/or noticeable. For some pathogens,

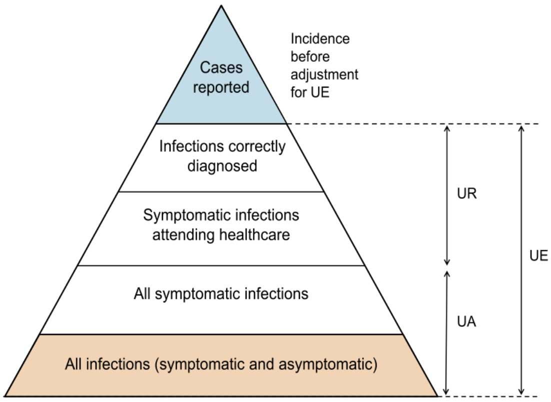
Cases reported
Incidence before adjustment for UE
Infections correctly diagnosed
Symptomatic infections attending healthcare
All symptomatic infections
All infections (symptomatic and asymptomatic)
UR
UA
UE

Figure 1.3: *The set of all cases of a particular disease, and the various levels of registration of those cases. UE stands for the overall extent of Under Estimation of cases which arise from Under Reporting of cases UR, and Under Ascertained cases UA. Ignoring the non-reported cases will lead to an underestimation of the disease incidence, if the disease suffer non-neglible levels of UR and UA cases. Reproduced from Gibbons et al. (2014) .*

seroepidemiology has indicated that the *majority* of cases go undetected; two examples are Salmonella (Simonsen et al., 2011) and Pertussis (Boven et al., 2004). These are pathogens for which infection is often asymptomatic or causing subclinical disease, leading to large number of undetected cases. In fact, for a number of pathogens, it was not possible to detect most infections before the advent of sensitive PCR-based diagnostics (Watzinger et al., 2006) and advanced antibody arrays (Uttamchandani et al., 2009).

In order to track cases from the lower levels of the pyramid, one can broaden the inclusion criteria, at the risk of including false positive cases. Another option is to target the lower levels via wider (longitudinal or cross-sectional) screening programs that target either some of the lower levels of the pyramid or the entire population (Gibbons et al., 2014). One of the hurdles to ongoing population level screening is that it can be expensive, and thus particularly difficult to implement, in particular in settings with limited resources, such as pathogen surveillance in developing countries. Occasional cross-sectional studies are cheaper to implement than continuous surveillance, but reliance on these limits the ability to track disease trends. Infectious diseases were seen as a solved problem during the latter part of the 20th century (Fauci, 2001), with the result that disease surveillance started lagging behind. The AIDS epidemic that began in the 1980s caused an increased awareness of the continuing risk posed by infectious diseases, but as late as 1994, Berkelman et al. (1994) referred to infectious disease surveillance as a "crumbling foundation". The capacity for infectious disease surveillance in general has improved greatly since then. With the advent of modern diagnostic measures such as those made possible by the PCR revolution (Yang and Rothman, 2004) the ability to rapidly identify pathogens has increased, and the effort of organizations such as the European Centre for Disease Control (ECDC) and the World Organisation for Animal Health (OIE) to standardize, unify and improve disease surveillance databases has improved our ability to compare pathogen occurrence and prevalence between different locations and countries. The development of the internet has also been a great help, with surveillance networks such as the ProMed mailing list (Madoff, 2004) and HealthMap (Freifeld et al., 2008) making even early warning signals of disease events available globally. However, the quality of disease surveillance is uneven across different nations, and in different contexts (The Institute of Medicine, 2007). Jones et al. (2008) analysed which areas of the world new pathogens are likely to emerge in, and compared that with how much research has been focused on the same areas. They concluded that the research and surveillance to a large extent are biased towards "the richer, developed countries of Europe, North America, Australia and some parts of Asia, than in developing regions". On the other hand, their analysis indicated that lower-latitude developing countries have the highest risk for emerging infectious disease events (EID). One implication of their study is that improving surveillance systems in highly resource-constrained countries would have a high pay-off for improving the ability of the global community to detect and respond to EID.

Wildlife disease surveillance is another area that has often been neglected, despite the growing awareness of the interdependence of diseases in livestock, humans, and wildlife, and thus the mutual importance of surveillance in all three types of hosts (Zinsstag et al., 2011). However, since wildlife disease research is constrained in funding (as few strong economical interests are directly affected), and is technically challenging (Mörner et al., 2002), improving the current state of surveillance systems has been difficult. Kuiken and Gortázar (2011) describes the state of wildlife surveillance across Europe, and concludes that a majority of countries have no general surveillance established beyond targeted efforts for a few key pathogens. In developing countries, surveillance systems for human diseases are often severely lacking, and animal disease surveillance has been referred to as "nonexistent" (Butler, 2006).

1.3 THE USE OF DIAGNOSTICS IN DISEASE SURVEILLANCE

A veterinarian or medical doctor can use symptoms to diagnose disease in a patient. However, in order to confidently establish that the disease is caused by a particular pathogen, rigorous and tested diagnostics procedures based on chemical or other assays are almost always necessary. The suggestion that infectious diseases might be caused by unseen organisms dates back to the 16th century, but only gained widespread acceptance in the 19th. A particularly important development was made by the german physician Robert Koch in 1890. Koch's postulates (Kaufmann and Schaible, 2005) set out criteria for concluding that a microorganism is the cause of a particular disease. In the original form, the postulates were: A disease is caused by a pathogen 1) if and only if it is present in a diseased host, and absent in non-diseased hosts; 2) The microorganism can be be isolated from an individual diagnosed with the disease, purified and grown; 3) After isolating the microorganism, introducing it into a healthy host causes the disease; and 4) from the experimentally infected host, the microorganism can be isolated again, and identified to be the same as the introduced organism. These postulates worked to put the identification of organisms that cause a particular disease on solid footing.

With respect to surveillance, the properties of a binary diagnostic test for a particular pathogen can be summarised by the *sensitivity* and the *specificity* of the test. The sensitivity refers to the probability of getting a positive result from an individual infected with the pathogen, while the specificity refers to the probability of getting a negative result from an individual not infected with the pathogen (the specificity). The terms of "Specificity" and "Sensitivity" were first used in the context of a screening program for cervical cancer (which we know now to largely be the consequence of an infectious pathogen) in the 1940s as described in a publication of the CDC in 1961 (Morabia and Zhang, 2004).

Following Koch, microbial culture is commonly used as the gold standard for diagnosis for infectious diseases (OIE, 2013), and treated as being 100% specific, with no false positive results. A sample of tissue or fluid is taken from the potential case, purified of substances that may inhibit growth, and a culture of organisms is grown that can be identified as a particular pathogen, via observation in microscope or some other approach through which it is possible to identify features unique to that pathogen. This method will nearly always (except in cases of sample contamination) be 100% specific, i.e. not show any false positives. However, growing a culture takes time and is therefore resource intensive. Depending on the pathogen, it might also be difficult due to biosecurity issues: only laboratories with biosecurity classification level four are allowed to culture highly pathogenic organisms such as Ebola or foot-and-mouth disease (US Department of Health and Human Services, 1999). For other pathogens, it might be difficult to create the right conditions of growth outside the host. For example, *Mycobacterium* spp. is notoriously difficult to culture, and if culture is successful, for taking several weeks to grow before being identifiable (OIE, 2014). Viruses and prions are even more challenging, requiring carefully prepared environments with the correct cells and proteins present to replicate. While culture is still used as a gold standard in studies of properties of new diagnostic methods, it is increasingly rare to use it as a diagnostic tool in itself, for the above mentioned reasons.

Another method of diagnosis relies on the existence of genetic material such as DNA or RNA from the pathogen in the test sample, as a proxy for the presence in the sample of the pathogen itself. This was made possible with the development of the polymerase chain reaction (PCR) procedure in 1985 (Mullis et al., 1986). A small length of a genetic sequence unique to the pathogen is used as a "primer" to amplify any genetic material containing that primer to the point where its presence can be easily detected. If amplified material is then found, that is evidence for a diagnosis. A particular "primer" tends to be unique to a particular pathogen; however, it is possible to use a combination of primers to be able to detect and identify a range of different pathogens simultaneously, including the identification of pathogen families, species, sub-species, or types. DNA- and RNA-based methods tend to be highly sensitive, able to detect concentrations of only a few tens or hundreds of organisms per sample. Because of their sensitivity, however, they also tend to be prone to false positive results, and can easily suffer from cross-contamination or cross-reactions with genetic material from non-pathogen sources, resulting in low specificity.

A third approach to diagnosis is to look for effects of the pathogen in the host (e.g. a detectable signature of the host responding to the presence of the pathogen). At its core, this is exactly what medical doctors do when diagnosing a patient in the traditional manner based on a collection of symptoms indicative of a particular disease. . Whilst such symptoms based diagnosis at times has rather low specificity i.e. several diseases may have similar symptoms, the laboratory methods exploit the fact that the immune system of animals and humans respond to infections in specific and predictable ways, generating antibodies that target a particular pathogen. These antibodies can be isolated from e.g. a blood sample, and their existence detected by mixing the test sample with a preparation of organic material derived from the pathogen that has been marked for easy identification with e.g a fluorescent protein. If there are antibodies in the sample that react with the organic preparation, this can then be measured, for example via enzyme-linked immunosorbent assay (ELISA) (Wright et al., 1993). The existence of antibodies are then taken as evidence for the individual having been infected with the pathogen. However, since antibodies are not 100% specific, there is usually a threshold concentration that the sample must display for the diagnosis to be reliable. A further complicating matter is that there are several types of immune response that can be used as a basis for a test. Some immune responses are specific to a particular disease, while other immune responses (such as interferon-γ) can react to a wide range of different infections. Some components of the immune response, such as memory cells, remain in the body for a long time, while others can subside quickly.

The different approaches of diagnosis measure different aspects of the host-pathogen interaction during the progression from initial pathogen infection via subsequent host response to recovery. Because of this, they will show a case as positive under somewhat different conditions (Greiner and I. A. Gardner, 2000a). For a culture approach to be feasible, there must still be live pathogens present in the host, and thus a positive test is evidence of an ongoing infection. A genetics-based approach will be positive if DNA or RNA is present; this mean that the pathogen could potentially have died out, but dead organisms still remain in the body. A positive DNA or RNA-based test is thus evidence of a recent infection, but not necessarily one that is ongoing (the genetic remains can be present in the body despite the infection having cleared). The presence of an immune response indicates that the host has responded to an infection at some point in the past. Some antibodies disappear quickly, and so their presence indicates a recent infection. Other types of antibodies can stay for weeks, months or years before their levels decline, and so their presence would only indicate that the individual has been infected at some point in the past. An additional complication is that individuals can develop antibodies just by being exposed to a pathogen, without developing an active infection. Using a positive antibody test to indicate infection thus usually implies a broader case definition than that based on using a culture test.

Knowing the details of what is being measured is thus important for understanding exactly what any given test data represents, and also for resolving possibly conflicting results between different tests. Classically, a test result is reported as a binary positive/negative for infection. More recent methods base such classifications on quantitative measurements of the indicator; concentration of antibodies (Uttamchandani et al., 2009), number of copies of a particular gene found (Caboche et al., 2014), or concentration of bacteria in a media

(Hammes and Egli, 2010). In such cases if the test result is above a certain cut-off value, it is classified as positive, but otherwise as negative.

However, quantitative results can also be analysed directly, in which case the statistical properties of the measurement becomes important. For example, many diagnostic methods use repeated dilutions to bring concentrations into a range that can be measured, and then the result is scaled by the dilution factor. If there is a constant error in measuring the concentrations, the scaling will then introduce a multiplicative error in the resulting data set. Such dilutions also frequently mean that observations are clumped at different multiples of the same number (2,4,8,. . .), which can lead to under dispersion which should be accounted for in any statistical analysis of the data.

The steps of the laboratory workflow can each introduce errors of different sorts, which need to be taken into account in subsequent statistical analysis (Greiner and I. A. Gardner, 2000b).

The use of the results of diagnostic testing can be approached from at least two different viewpoints. From the clinician's point of view, a diagnostic test is a tool supporting decision-making and helps in identifying appropriate treatment for the patient at the point of care. From this clinical perspective, information on when to administer treatment and when to abstain are the most important functions of diagnostic testing, and this includes understanding the limitations and strength of the available tests.

In order for a diagnostic test to be useful as a decision-support tool, a clear result indicating the likely status of the individual is clearly of great value: a message of "infected" or "not infected", with a known level of uncertainty. However, from a statistical or epidemiological as opposed to clinical prospective, the interest lies in the overall pattern of disease in the population. From these perspectives, we have the luxury of not having to worry about what the test result is for any particular individual. Understanding these two differing perspectives becomes important for the discussion of quantitative test results instead of binary results, and the use of more than one diagnostic test. For the clinician, a quantitative test response has little value unless it can be translated to inform a particular course of action, for instance by the use of a binary cutoff point. Likewise, in most cases, the use of multiple diagnostic tests is only of relevance to the extent that it can be used to pin down the appropriate diagnosis.

From a population perspective however, the overall distribution of the quantitative test results within the population can be of use for estimating test variability as well as variation in exposure and severity/response to pathogen infection between individuals. Multiple diagnostic tests, while providing limited additional information on the individual level, can be used in the aggregate to provide a description of the how the pathogen is spreading, a problem we address in this thesis (chapters 3 and 4). The increasing awareness of the high proportion of subclinical infections that play an important role in the epidemiology for a number of pathogens also complicates the usage of a binary "infected"/"non-infected" dichotomy. The collection of test results pre-cutoff can thus be invaluable to statisticians and epidemiologists, while being perceived as unnecessary by those closer to the point of care e.g clinicians, and it is important to be aware of this tension when conducting studies or setting up surveillance systems, so that both perspectives are acknowledged.

1.4 THE USE OF STATISTICAL METHODS IN DISEASE SURVEILLANCE

This section will not attempt a review of all types of statistical analysis used in the process of disease surveillance; there is far too much to cover in just a few paragraphs. Instead, it will describe a few examples of how statistical analysis can be used to extract important signals from limit surveillance data that can then be used to inform decisions on how to manage a particular disease.

At the most fundamental level, statistical analysis can be used to highlight patterns in collected data. In 1854, the physician John Snow identified the source of a Cholera outbreak by mapping the cases and noticing that there was a cluster of cases around a particular street pump, a study widely considered to mark the start of epidemiology (Hempel, 2013). Such cluster analysis has remained an important research topic to this day. One of the most well-known modern statistical tools for cluster analysis is the Kulldorf scan statistic (Kulldorff, 1997), implemented in the SatScan program (Kulldorff et al., 1998), among other statistical software packages. Cluster detection can look for either clusters in spatial dimensions with a number of cases occurring in one area, clusters in temporal dimensions where a number of cases occur within a short space of time but not necessarily in the same place or clusters in both space and time. Cluster analysis can also be based on metrics other than space and time, e.g. age, to detect unusual aspects of data. Another prominent type of pattern detection is to understand when reported cases of some disease reaches an "unusual" level, due to an outbreak or because of some other change in epidemiology, that might require a public health intervention. One important approach for doing so is known as a temporal scan statistic, first introduced in Farrington et al. (1996), which described the Farrington Algorithm. This algorithm continuously evaluates an incoming time series for statistically significant changes, sounding an alarm if the cumulative change over a specified time period reaches a significance threshold, according to some statistic; many different statistics have been suggested for this broad purpose (Unkel et al., 2012).

Statistical analysis can also be used to extract trends from indirect data sources. One example is the recent increasing usage of syndromic surveillance. The main motivation for syndromic surveillance is that there is a delay between infection and symptoms, and between symptoms and health-care-seeking behavior (in humans), symptoms severe enough to be noticed by a farmer or veterinarian(in livestock) or to cause death and the resulting carcass to be noticed by hunters or the general public (in wildlife). However, a symptomatic individual has an effect on its surroundings or habitat, and this effect can be tracked. Since symptoms occur before an infection is diagnosed, syndromic surveillance has the potential to increase the timeliness, sensitivity and robustnesss of existing systems (Buehler et al., 2003; Dórea et al., 2011). Hulth et al. (2009) developed a statistical model using partial least squares regression to predict officially reported influenza case numbers from web searches submitted to a medical web site. Google Flu Trends also predicts influenza levels, using a simple statistical regression model combined with model selection procedures and using google search queries as their data source (Ginsberg et al., 2009). Syndromic surveillance can be used for animals as well; Warns-Petit et al. (2010) describe the use of data mining to categorize recorded necropsy data pre-diagnosis in wild animals, identifying clusters that could signify wildlife disease outbreaks of re-emerging or emerging pathogens.

An increasing use of statistical analysis is to integrate different sources of information together with the surveillance data to make inference about the epidemiology of disease. For example, in the case of a disease outbreak, it is possible to combine data analysis with models of the epidemic process to, for example, predict the timing of the peak based on the present rate of change, estimate the duration of the outbreak, or the scale of the impact (Andersson et al., 2008). Epidemic models informed by statistical analysis were heavily used during the 2001 foot-and-mouth disease outbreak in Britain (Kao, 2002), and again during the 2007 bluetongue disease outbreak (Szmaragd et al., 2009) to inform the decisions on suitable control policies. Another approach is to combine information on diagnostic test behaviour with data analysis to estimate historic patterns of disease. For example, Giorgi et al. estimated the time of the start of an HIV outbreak under assumptions of exponential growth of viral load (Giorgi et al., 2010). Others have exploited information on diagnostic test kinetics, i.e., the pattern of diagnostic test values during the course of infection, to estimate average incidence rates. Examples include the use of antibody test kinetics to estimate seroincidence rates forinfluenza (Baguelin et al., 2011), *Salmonella* in cattle (Nielsen et al., 2011) and *Salmonella* in humans (Simonsen et al., 2008). Interestingly enough, because of the delay in surveillance systems, it may sometimes be necessary to estimate what is happening “now”. Höhle and An der Heiden (2014) describes a Bayesian model for “nowcasting” current levels of disease, based on the number of currently reported cases and the expected number of delayed reports that will arrive in the coming weeks, and demonstrates its use in the case of a large 2011 outbreak of Shiga toxin-producing *E. coli* 0104 in Germany.

1.5 FOUNDATIONS FOR THE STATISTICAL METHODS USED IN THE THESIS

The following sections give a general overview of the statistical foundations underlying the methods developed in the thesis, introducing the core terminology and describes some of the specific assumptions and approaches used.

1.5.1 *A brief introduction to Bayesian Statistics*

This thesis adopts a Bayesian approach to the analysis of diagnostic test data, instead of a classical, frequentist approach. The difference is partly in philosophy, but mostly show up as a difference in how results are reported and which tools tend to be used. In a Bayesian framework, probability distributions are encodings of our knowledge of the world. The Bayesian methodology described below provides a means to update these distributions when new information is available, and thus the Bayesian approach provides a natural framework for the integration of data. Quantities that we are interested in can be viewed as having probability distributions that encode our original knowledge about them, with a modal value representing the "best guess" for the true value, and the spread of the distribution describing our uncertainty of that guess. Assume that there are two quantities: A, which represents something you want to know, and B, which represents a quantity that you have some knowledge of e.g. via measurement or observation. As an example, consider a person named Ainsley, with an unknown gender but with the known height of 157cm. Then A would represent the gender of Ainsley, and B would represent the height.

There are three important types of probability distributions in Bayesian statistical analysis: prior distributions, posterior distributions, and likelihoods. A prior distribution $P(A)$ represents our knowledge about A before (or prior to) observing some (new) data B. In the example of Ainsley, it would be our best guess of the gender before learning about the height. For example, in the wikipedia page on the name "Ainsley", there are links to 5 male and 3 female people. Taking this information as our prior gives a binomial distribution $Bin(5/8, 3/8)$ for the outcomes (male,female).

A posterior distribution $P(A|B)$ represents the state of our knowledge incorporating the data B. The notation $P(x|y)$ refers to the probability of x conditional on the value of y. In the case of our example, the posterior distribution $P(A|B)$ would refer to our knowledge of Ainsley's gender given that we know the height of 157cm.
A likelihood $P(B|A)$, finally, describes the probability of observing the data B given a particular value for A. It is often treated as a function $L(x)$ of the possible values x of A, so that $L(x) := P(B|A = x)$. In our example, $P(B|A)$ would refer to the probability of the height being 157cm given that Ainsley was male/female, respectively. The likelihood is typically based on a set of assumptions underlying our model of the process of interest and describes how the data is generated.
The prior, the posterior, and the likelihood can be related to each other via Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

This equation can be interpreted as saying that what is known about some object A given that we make some observation B is encoded in the probability $P(A|B)$, and that this can be expressed by multiplying the prior with the likelihood $P(B|A)$ divided by the probability $P(B)$. The probability $P(B)$ is often called the normalization constant for the posterior distribution $P(A|B)$ and is the probability of observing the data B under the assumptions underlying the likelihood irrespective of the value of A. In intuitive terms, Bayes' theorem thus tells us how, under model assumptions encoded by the likelihood, to modify our prior state of knowledge about the world by how surprising (i.e how unlikely) newly observed data is.
It should be noted that the normalization constant $P(B)$, the unconditional probability distribution of the data, is rarely known. The posterior probability is therefore usually calculated as $P(A|B) \propto P(B|A)P(A)$. Since for most applications, only the relative probabil-

ity of different values of A are of interest (and in particular, finding the values of A with the highest posterior probability), this is rarely an issue in practice (e.g. see the section on Markov chain Monte Carlo below).

Bayesian inference make use of Bayes' theorem in the following way. Assume that there there is some data that is believed to be relevant to a quantity of interest. Set up a statistical model describing how the data is expected to be distributed depending on the various parameters that might might affect the observed data. Then calculate the likelihood (as defined in the previous section) of the observed data under the statistical model. Assign a prior distribution for all parameters representing our state of knowledge about them before conducting the statistical analysis. Finally, calculate the posterior distribution from the product of the likelihood and the prior distribution. This posterior distribution can also be interpreted as the prior distribution shifted in the direction of the data likelihood.

From the Bayesian perspective, the full posterior is of interest and the posterior distribution is what tends to be reported, possibly summarised, rather than point estimates of parameter values. An Analytical solution of the posterior distribution is only possible in fringe cases, and numerical algorithms are the only way to conduct Bayesian inference in practice. For this reason the application of Bayesian statistical analysis has expanded greatly with the advent of cheap widely available computing resources.

1.5.2 *Markov chain Monte Carlo*

The posterior distribution is usually described with a complex multidimensional integral that is not solvable analytically, and so needs to be approximated numerically. Many different algorithms have therefore been developed that implement Bayesian inference by providing approximate posterior distributions. The most common class of algorithms, and the kind used in this thesis, are Markov chain Monte Carlo (MCMC) type algorithms (Robert and Casella, 2011). In very general terms, MCMC algorithms work by selecting an initial point in the parameter space, and then jumping from point to point by calculating the value of the density at the current point (up to a normalizing constant, e.g. $P(B)$ in the example above), and then calculating the density at a second point drawn from a proposal distribution. The algorithm decides to move to the second point or stay with the current point by taking into account properties of the proposal distribution and the relative value of the density at the proposed second point compared to the current one. In this way, the algorithm generates a sequence of samples from the posterior, where each sampling point is only dependent on the preceding sampling point. The type of conditional dependence that this process exhibits is known as "Markov property", and sequence of samples having the Markov property is known as a Markov chain. Because we are generating the sequence at random, the whole procedure is thus known as Markov chain Monte Carlo(MCMC); MCMC was preceded by so called Monte Carlo algorithms that propose independent samples from the distribution of interest but are typically much less efficient. It can be proven that by carefully specifying the conditional jump probability, the distribution of the samples of the MCMC converge to the posterior distribution of interest (Asmussen and Glynn, 2011). There are many different implementations of MCMC , but the two oldest and most widely used are the Metropolis-Hastings algorithm and Gibbs sampling.

The Metropolis algorithm was published in a seminal paper by Metropolis et al. (1953). Using the Metropolis algorithm, sequential samples are generated as follows: Assume a starting point X_t and an expression for the posterior density $p(x) = p(A = x|B) \propto p(B|A =$

$x)p(A = x)$. Jump from $X_t = x_i$ to a new point x_j randomly selected in in the parameter space using an arbitrary transition probability distribution $T(x|X_t)$ that depends on the current point, and is symmetric so that $T(x_i|x_j) = T(x_j|x_i)$. Compare the posterior density X_i and X_j, and if the value at X_j is larger than that at X_i, set $X_{t+1} = X_i$. However, if $p(X_j) \leq p(X_i)$, then calculate the ratio $\alpha = p(X_j)/p(X_i)$, and set $X_{t+1} = X_j$ as the next value of the chain with probability α. If X_j is rejected, then set $X_{t+1} = X_i$. Once a value for X_{t+1} has been chosen, set X_{t+1} as the new point, and repeat the process. Metropolis et al. proved in their paper that the distribution of samples generated in this way will converge to the distribution defined by $p()$. In the Metropolis algorithm, the transition kernel $T(x|X_t)$ is arbitrary but must be symmetric, so that $T(X_i|X_j) = T(X_j|X_i)$. A generalization known as the Metropolis-Hastings (MH) algorithm that relaxes this symmetry requirement was later described in a paper by Hastings (1970), and it is this form that is most commonly implemented. Gibbs sampling can be seen as a special case of the Metropolis algorithm, introduced in Geman and Geman (1984), that can be used if it is possible to sample from the conditional posterior distribution for each variable (holding the others fixed) but not necessarily from the full joint posterior distribution. Assume a posterior density $p()$ and a starting point in the parameter space X_t, with components $X_t = (x_0, x_1, \ldots, x_n)$. Fix all but the first component of X_t, and generate a new value x'_0 by sampling from the conditional posterior distribution $p(x|x_1, x_2, x_3, \ldots, x_n)$. Then sample a new value x'_1 from $p(x|x'_0, x_2, x_3, \ldots, x_n)$, conditioning on the newly sampled value x'_0. Repeat the process for each component until a new vector $X_{t+1} = (x_0, x_1, x_2, x_3, \ldots, x_n)$ has been generated. Repeat the process from the beginning starting at X_{t+1}.

The use of MCMC algorithms is a computationally expensive process, but the realisation of Moore's law, other developments in CPU design, and improved design of proposal distribution such as Hamiltonian MCMC (Hanson, 2001) have made more and more problems tractable using MCMC methods. The MCMC algorithm can itself be implemented using different approaches. It is possible to code it explicitly using a programming language of your choice, such as R, C++ or Python. Alternatively, there are a number of computational packages that have implemented an engine for the translation of statistical models to program code running an MCMC algorithm, allowing the user to focus on encoding their model using the engine syntax.

1.5.3 *MCMC engines*

One of the earliest implemented engines for MCMC was the BUGS project, which was started in 1989, and popularized with the creation of the Windows software WinBUGS (Lunn et al., 2009). BUGS is a declarative language for describing statistical models as Directed Acyclical Graphs (DAGs) (Thomas et al., 1994). These are graphs where each statistical component is a node. Edges between a pair of nodes represent their conditional dependence. Hierarchical statistical models are well suited to a description in terms of DAGs, and so with the increasing prominence of hierarchical modelling (Steenbergen and Jones, 2002), BUGS and WinBUGS gained in popularity (though it may be that with the advent of WinBUGS, hierarchical modelling became easier and thus more popular). One downside of this approach, however, is that there is no concept of "order" in a BUGS program file, and thus only limited possibilities to implement if-then type statements. An alternative open source implementation of the BUGS language is JAGS (Just Another Gibbs Sampler) by Plummer (2003), and after ongoing development of WinBUGS ceased, JAGS has become one of the dominant programs with which Bayesian inference is implemented. Like BUGS, it is based on a declarative language describing DAGS, and as a result the kinds of models that are possible to implement are somewhat restricted.

A relatively recent development is STAN (Hoffman and Gelman, 2014), a probabilistic programming language with more sophisticated MCMC algorithm known as a "No-U-turn Hamiltonian MCMC". By adjusting the size of sampling steps based on information on the gradient of the posterior, STAN can be more more efficient and in general take fewer iterations to reach convergence, though each iteration requires more computations. For a wide range of models, this translates into STAN implementations of a given Bayesian inference problem being faster in terms of reaching a given effective sample size compared with corresponding implementations in JAGS or WinBUGS. As opposed to JAGS or WinBUGS, the language is not based on DAGS; the program code is interpreted sequentially and is therefore more expressive. As a side effect, the sequential nature combined with a requirement to declare the type of variable upon definition (continuous, integer, etc) makes for substantially easier debugging of code.

The different engines described have different strengths and weaknesses. Coding your own model has the advantages of allowing you to incorporate problem-specific information and shortcuts, and that it can result in faster inference, especially when coding is carried out using a lower-level language. A disadvantage is that the coding itself can take considerable time, and that the resulting program is more prone to bugs than using one of the higher-level languages that already has an MCMC engine implemented. WinBUGS has been around for a long time, and there are a wide variety of examples and models implemented to take inspiration from. JAGS is faster than WinBUGS, is portable across computing platforms, does not rely on a DOS-based software, and is still being developed and improved. On the downside, Dr. Martyn Plummer is the sole developer, which means that bug fixes and improvements can take time. STAN is faster, more expressive, easier to debug, and under very active development by a four-person team. However, at the time of this writing, there are some models that cannot be implemented, and there are not a large numbers of previous examples for how STAN can be used.

In this thesis, all the described models were implemented in JAGS, with the exception of chapter 4, which made use of STAN.

1.5.4 *Evaluation of convergence*

Running an MCMC algorithm results in a series of sequentially correlated draws of parameter values. The proofs of convergence of MCMC state that in the limit, i.e. after "enough" number of sequential parameter values have been drawn, the distribution of parameter values corresponds to the posterior distribution of interest. In order to evaluate if "enough" draws have been made, one needs to evaluate whether the MCMC procedure has reached a steady state. To do so in a strict formal manner is only possible in some simple analytically tractable special cases, but for the remainder of models there are a number of heuristics that can be used for this purpose. Summary statistics have been proposed that measure the amount of convergence. The Gelman-Rubin statistic (Brooks and Gelman, 1998) measures the within-chain variance of several (usually parallel) runs of the same MCMC algorithm, with initial values in different parts of the parameter space, and compares this to the overall variance of all chains taken as a whole. If the ratio of the overall variance to the within-chain variance is less than 1.15 (the original paper recommended 1.2 as a rough criteria for convergence), this is taken as evidence that the different chains are sampling from the same distribution, and thus have converged. The GR statistics relies on the existence of several different chains, and that the different chains initial values were chosen from sufficiently different parts of the parameter space that their starting between-variance is high. In this thesis, Initial starting values for the parameters in the various MCMC analyses conducted were chosen at random, either sampled from the prior distributions, or by selecting a central region of the support and sampling uniformly for this region. In either case, initial values independently sampled for each chain.

Another heuristic includes looking at autocorrelation plots for different parameters, as too high an autocorrelation indicates that the chain is still shifting towards the true value. A high autocorrelation can also indicate bad "mixing", where the chains have converged to the true posterior, but where the correlation between individual draws is so high that you need to generate a large number of draws to produce an unbiased estimate representative of the full posterior distribution. In addition to any quantitative evaluations, it is also essential to visualise the draws directly; obvious patterns over time tend to indicate non-convergence, and other problems such as parameters being correlated with each other also tend to show up when plotting.

For the results presented in this thesis, the Gelman-Rubin statistics has been used for the initial judgement of convergence and all parallel chains have GR values under the accepted threshold of 1.15. As a rule of thumb, five chains have been run in parallel to provide sufficient data on between-chain variability. After a draw passes this value, additional inspection of trace plots and autocorrelation plots have been used to further ensure that the estimated posterior is not misleading.

1.5.5 *On the selection of priors*

The choice of priors is a contentious one in the field of Bayesian statistical analysis, with many different schools, and the approach used in this thesis has evolved over the course of the years of work. As a general summary, there are three main categories of priors: informative priors, uninformative priors, and weakly informative priors (Robert (2007), p 105ff).

Informative priors explicitly try to incorporate previous information in a parametric form. This information might be results from previous studies or solicited from experts. This is a classic approach to Bayesian statistical analysis, as it is strongly influenced by the view that probability distributions should describe our state of knowledge of the world. In the work for this thesis, informative priors were used in the development phase of the hindcasting model for the variance of test diagnostics. At a later stage, these priors were changed to the uninformative kind. The hindcasting method is thus robust to the choice of prior, and the data used was informative enough to identify parameters without strong prior beliefs.

According to the "uninformative" school of thought, priors are chosen in a way that attempt to influence the posterior estimates as little as possible. A classic example of this is the use of a gamma(0.0001,0.0001) prior for the reciprocal of the variance of a normal distribution, thus putting approximately equal weight to a large subset of the positive real number line. (This choice of prior was introduced as a recommended choice in BUGS in 1994(Thomas et al., 1994), and subsequently gained popularity). This school is primarily governed by an attempt to be as scientifically unbiased and objective as possible.

A category of priors that can be used in both an informative and an uninformative setting is conjugate priors. These are priors chosen so that the prior distribution $P(A)$ and the posterior distribution $P(B|A)P(A)$ both belong to the same class of parametric distributions. For example, the Beta $\mathbb{B}(a, b)$-distribution is the conjugate prior to the proportion parameter of a binomial distribution $Bin(a, p)$, in that the posterior distribution for p given a beta prior $\mathbb{B}(a, b)$ and observed c success and d failures from a binomial distribution is $\mathbb{B}(a + c, b + d)$. Because of this property, conjugate priors simplify analytical calculations and were important for making numerical calculations tractable in the early days of lower computer power. In the case of Gibbs sampling (described above), using conjugate priors implies that the posterior conditional distribution of each component follows a known parametric form, and so can be easily sampled when updating the components. One reason the gamma prior became popular was that the gamma distribution is the conjugate prior for the inverse of the variance parameter in a normal distribution with unknown mean and variance.

The approach of weakly informative priors, championed by amongst other, Andrew Gelman (Gelman et al., 2008), was used during later phases of the thesis work, when investigating and trying to resolve convergence issues. When using weakly informative priors, the principle is to incorporate domain knowledge on as general a level as possible. So, for example, it is known that the height of humans are on the order of meters and not kilometers or millimeters, and so a lognormal distribution of $lognorm(1, \sqrt{10})$ would be used to incorporate this knowledge (a standard deviation of $10^{1/2}$ implies that 95% of the distribution lies within $[1/(10^{1/2*2}), 1x(10^{1/2*2})] = [1/10, 10)]$). This improves the behaviour of the numerical inference by regularizing the likelihood surface, while at the same time avoiding that posterior estimates are highly influenced by subjective expert judgements.

1.6 THESIS OUTLINE

The work conducted in this thesis is focused on ways to exploit the full potential of diagnostic data, leveraging the continuous and dynamic nature of test measurements, and the additional information gained by combining multiple diagnostic tests. A key part part of the argument put forward is that recording raw test data and conducting multiple tests on samples typically collected in current surveillance systems would make possible statistical analyses that can produce valuable information on the epidemiology of diseases. This is conceptualised as being of particular value where surveillance data is limited, or for diseases where little is known, either because they are emerging infections, or because they have been undetected or considered of low priority in the past.

The classical Hui-Walter latent class analysis (Hui and Walter, 1980) makes it possible to estimate the unknown sensitivity and specificity of two binary diagnostic tests by comparing their results in two settings with differing levels of incidence. Chapter two applies this approach to evaluate the use of multiple diagnostic tests to estimate sensitivity and specificity of diagnostic tests that can distinguish vaccinated and infected animals (DIVA tests) for *Mycobacterium bovis*, producing recommendations regarding a potential vaccine trial in cattle in the UK. It expands on the Hui Walter approach by showing how to estimate vaccine efficacy in addition to sensitivity and specificity.

Chapter three also looks at the additional benefit of using two diagnostic tests, but in a more general fashion by treating the diagnostic test response as continuous with known kinetics over the time since infection. The concept of "hindcasting" (inferring the historical trend of) historic disease dynamics from cross-sectional data is introduced, and applied to the situation of endemic diseases which exhibit linear trends. In this way, it is possible to estimate the long-term trends of pathogen incidence, for example to evaluate the impact of policy implementations using cross-sectional data.

Chapter four is primarily aimed at improving the early phase of disease surveillance, during the initial phase of an outbreak, or when a new pathogen has been introduced to a region or country. The focus is on emerging or epidemic pathogens where the interest is on finding out information on the dynamic of an outbreak. The approach is illustrated by application to two case studies, one based on a bluetongue outbreak in the UK in 2007, and one based on a whooping cough outbreak in Wisconsin in 2003..

Chapters three and four highlight the benefits gained by utilizing knowledge of test kinetics in disease surveillance, and thus make a case for considering ways to ensure greater access to such information. Chapter five is thus an exploration of strategies for improving estimation of test kinetics using observational data, with a particular focus on the potential advantages of estimating two or more test kinetics simultaneously.

Chapter six provides an overview of the results presented in the thesis, highlights potential avenues of research and argues for the greater utilisation of multiple testing in the field of disease surveillance.

CHAPTER 2: A LATENT CLASS ANALYSIS OF BTB DIVA TESTS

Abstract The practice of disease surveillance covers the process from discovering potential cases, via taking samples and testing these for disease with different diagnostic tests, to collecting the test results in databases, analysing the patterns of disease, and deciding on appropriate responses and policies. As a part of this process, establishing the properties of the diagnostic tests used is of critical importance. This chapter describes an application of classical binary diagnostic test analysis for estimating sensitivity and specificity in a setting of vaccinated individuals tested using two diagnostic tests with different properties in vaccinated and unvaccinated animals. Starting from the definitions of sensitivity and specificity of a diagnostic test, it discusses an approach to estimate these using a Bayesian Latent Class Analysis. The approach is then applied to estimate the effect of study size on sensitivity and specificity estimates for a trial of a test for bovine Tuberculosis that is able to distinguish between vaccinated and infected animals, a so-called DIVA ("Distinguishing Infected and Vaccinated Animals") test. The sample size analysis was commissioned by the Welsh Government and the British Department for Environment, Food and Rural Affairs (Defra). The result of the analysis shows that the sample size required to demonstrate that the test specificity is above the previously established threshold for cost-effectiveness, is in excess of 30 000 animals, even under the assumption that the real specificity is 99.9999%, and including a pilot study of animals tested with a gold-standard approach. The described framework expands on published studies by estimating vac-

cine efficacy in addition to diagnostic test properties, allowing for the application of latent class analysis in a wider range of settings than previously possible.

2.1 INTRODUCTION

Diagnostic testing plays a crucial role in the surveillance and detection of infectious diseases. In its simplest form, we take a sample from an individual, and use our test, for example an antibody test, to say whether the individual is infected. As all tests are imperfect, knowing the behaviour of the diagnostic tests used for surveillance is fundamental to evaluate collected surveillance data; of particular importance is the reliability and error rates of the tests used.
Classically, diagnostic tests have been treated as having a binary response, producing either a positive or a negative result. The diagnostic test gives a result that can be classified as "positive" if the individual is infected, and a result that can be classified as "negative" if the individual is not infected. Denote by $D_+(D_-)$ a infected (non-infected) individual, and t_+ (t_-) a positive (negative) test result. The probability that a test of an infected animal produces a positive result , $Se = P(t_+|D_+)$ is commonly referred to as the **sensitivity** of the test, and the probability that a test of a non-infected animal produces a negative test, $Sp = P(t_-|D_-)$ is commonly referred to as the **specificity** of the test. This terminology was first introduced in 1961 by Thorner and Remein (1961) in a US dept. of Health publication. Together, sensitivity and specificity fully describe the expected behavior of a binary test, when used in a population and on a disease where both sensitivity and specificity can be assumed to be the same for all individuals.
In order to estimate the specificity of a test, the ideal situation is to test a population that is known to be free from disease, and produce

statistical estimates and confidence intervals of the specificity from the proportion of negative test results and the number of individuals in the population. Similarly, in order to evaluate the sensitivity of a test, the ideal situation is to test a population where all individuals are known to be infected, and estimate the sensitivity from the proportion of positive test results. However, such ideal situations are rare, as it requires either a perfect test or a controlled infection study.

A somewhat more realistic approach is to test a population with unknown prevalence with two tests, one reference test where the sensitivity and specificity is already known, and another test that is being evaluated. In such a situation, it is possible to estimate the unknown sensitivity and specificity of the new test. However, this creates a bit of a catch-22 situation, as the properties of the reference test would have to have been evaluated at some point in the past.

A statistical method for estimating the properties of two diagnostic tests at once was pioneered by Hui and Walter (1980). In this seminal paper, they show that it is possible to estimate the unknown sensitivity and specificity of two different diagnostic tests simultaneously if both tests are used on all individuals in two populations with different (but unknown) prevalences of disease. The approach used in their paper is known as "Latent Class Analysis".

This chapter extends such latent class analysis to a situation where a vaccine can be expected to interfere with the diagnostic test performance. The assumptions are that two populations with different prevalences of disease are available, each population is split up into one vaccinated and one unvaccinated subpopulation, and two diagnostic tests are used on all four subpopulations. We allow sensitivity and specificity to vary for vaccinated and unvaccinated individuals. It is demonstrated that it is possible to extend the latent class framework to estimate unknown vaccine efficacy in addition to the unknown test parameters, and population prevalences. The framework is applied to estimate the effect of study size on sensitivity and specificity estimates for a trial of a test for bovine Tuberculosis that is able to distinguish between vaccinated and infected animals, a so-called DIVA (”Distinguishing Infected and Vaccinated Animals”) test. The results of this study are currently being considered by the Welsh Government and the British Department for Environment, Food and Rural Affairs) (Defra).

2.2 A LATENT-CLASS ANALYSIS OF VACCINATED POPULATIONS

In this section, the notation follows that of Johnson et al. (2001) closely, with the addition of vaccinated and unvaccinated subpopulations, and a vaccine effectiveness parameter.

Consider two populations in which there are cases of the pathogen of interest; one ”high prevalence” population and one ”low prevalence” population, but where the exact prevalences are unknown in both populations. Each population is split into two subpopulations; one consisting of vaccinated individuals, and one consisting of non vaccinated individuals. Assume there are two different tests, with possibly differing sensitivity and specificity. Assume that the results of the tests are independent conditional on the status of the animal, and that sensitivity and specificity are affected by the vaccination status of the individual.

Ignoring vaccination, we would define $P(t = +|D = +)$ as the probability of a test t being positive given that the true status D is positive. We would also have $P(t = +|D = -)$, $P(t = -|D = +)$, and $P(t = -|D = -)$ denote the probabilities of the different combinations of test results and true status. However, here we assume that each test t can have different sensitivity Se_{tv}, defined as $Se_{tv} = P(t = +|D = +, v)$ when used to test animals with vaccination status v (i.e. vaccinated or unvaccinated). Likewise, we assume that the specificity $Sp_{tv} = P(t = -|D = -, v)$ can differ between tests and between vaccinated and unvaccinated animals. Se_{tv} and Sp_{tv} are treated as unknown properties of the tests in the model. Denote the prevalence in each unvaccinated population p by I_{p1}, and the prevalence in each vaccinated population by $I_{p2} = \lambda * I_{p1}$, where λ denotes the vaccine efficacy defined as the relative reduction in incidence in vaccinated animals. Assume that the relative reduction of incidence is the same regardless of the original incidence rate.

Finally, consider the event that a particular individual is tested with the two diagnostic tests. Denote by $++$ the event that the two tests, according to a specified ordering, are positive. In a similar fashion $+-$ denotes the event in which the first test is positive and the second negative, etcetera. Note that that a test can be positive either because the true status of an animal $D = +$ and the test gives the correct answer (with probability Se_{tv}), or because the true status is negative, $D = -$, and the test gives the wrong answer (with probability $1 - Sp_{tv}$).

Using this, it is possible to write the probability for a combined test result given population and vaccination status as:

$$P(++|p,v) = (Se_{1v} * I_{pv})(Se_{2v} * I_{pv}) +$$

$$(1 - Sp_{1v}) * (1 - I_{pv})(1 - Sp_{2v}) * (1 - I_{pv})$$

$$P(+-|p,v) = (Se_{1v} * I_{pv})((1 - Se_{2v}) * I_{pv}) +$$

$$(1 - Sp_{1v}) * (1 - I_{pv}p)(Sp_{2v}) * (1 - I_{pv})$$

$$P(-+|p,v) = ((1 - Se_{1v}) * I_{pv})(Se_{2v} * I_{pv}) +$$

$$(Sp_{1v}) * (1 - I_{pv})(1 - Sp_{2v}) * (1 - I_{pv})$$

$$P(--|p,v) = ((1 - Se_{1v}) * I_{pv})((1 - Se_{2v}) * I_{pv}) +$$

$$(Sp_{1v}) * (1 - I_{pv})(Sp_{2v}) * (1 - I_{pv})$$

When tabulating test data, denote by $N_{++,pv}$ the number of observations where both tests were positive in population p with vaccination status v, and similarly for the number of observations with only the first test positive, only the second test, or neither test positive. Then the vector

$$\mathbb{N} = \{N_{++,pv}, N_{+-,pv}, N_{-+,pv}, N\text{-}_{,pv}\}$$

is multinomial-distributed with probability vector

$$\{P(++|p,v), P(+-|p,v), P(-+|p,v), P(--|p,v)\}$$

There are four counts for each subpopulation, and four subpopulations, giving a total of 16 data points, and 12 degrees of freedom. The probability vector is a function of $\{Se_{tv}\}$ and $\{Sp_{tv}\}$, with four distinct values each, I_p, with two values, and by λ, giving a total of of 11 parameters. We therefore have one degree of freedom when estimating the parameters.

Given observed counts $\{N\}$, the data likelihood for the parameters in the model is then simply:

$$\prod_{\forall p,v} \left(P(++|p,v)^{N_{++,pv}} \times P(+-|p,v)^{N_{+-,pv}} \times\right.$$

$$\left.P(-+|p,v)^{N_{-+,pv}} \times P(--|p,v)^{N_{--,pv}})\right)$$

We evaluated the likelihood using a Bayesian Markov Chain Monte Carlo (MCMC) approach implemented in the JAGS language (Plummer, 2003) using beta priors for Se_{pv} and Sp_{pv}, with parameters derived from the results of "gold standard" tests. Specifically, if the number of true positive animals, as detected by a gold standard test that were detected by test 1 was $s1$, and the number incorrectly tested as negative was $n1$, then a $\beta(s1+1, n1+1)$ prior was used for the sensitivity of test 1. A similar approach was used to define beta priors for the sensitivity of test 2 and the two specificities. Starting values for the parameters in the MCMC were chosen at random from the prior distributions, independently sampled for each chain. Inspection of trace plots indicated that for all scenarios, the chains had fully converged after a burn-in of 1000 iterations. The following 1000 iterations were then used as basis for the presented results. There is a potential technical issues with the implementing the above approach, known as label switching. In the expressions above, swapping Se for $1 - Sp$, and Sp for $1 - Se$ would give the exact same result, and so the posterior could converge to either of these interpretations, leading to a bimodal distribution. One solution for both of these problems to require that Sensitivity plus Specificity is above 1 (as otherwise swapping the interpretation of the test around, treating positive test response as a negative result and vice versa, would result in a better test) (Toft et al., 2004). In the case of this study, the use of prior information from gold standard tests proved enough to resolve the potential bimodality, so constraining the sensitivity and specificity was not needed.

2.3 applying the model to inform an m.bovis diva test trial

The use of widespread vaccination is one of the most powerful methods available for the control of endemic and epidemic infectious diseases. The success stories are many, including polio, smallpox (WHO and Global Commission for the Certification of Smallpox, 1980), and rinderpest (FAO, 2013). In some cases however, the use of vaccination can be complicated by diagnostic tests failing to distinguish between infected and vaccinated individuals. Unless the vaccine is perfectly efficacious, this means that any breakthrough cases that occur despite vaccination would go undetected. In the case of an outbreak of cases, the lack of a functioning test would severely hamper control efforts.

In the case of bovine Tuberculosis ("bTB"), the standard test recommended for screening by the World Organisation for Animal Health is the so-called tuberculin skin test (OIE, 2009), which consists of injecting purified tuberculin proteins in the neck of the animal and looking for the resulting inflammation that indicates the presence of antibodies against *M.bovis*. Vaccinated animals will show the same reactions, and so the tuberculin test is unable to distinguish between vaccinated animals without infection, and vaccinated animals where breakthrough infections have occurred. Because of this, the European Council have made it mandatory for countries to abstain from vaccination if they want to retain an export classification (EEC, 1977). The concern is that infected animals will be exported and thus spread the disease to countries that are disease free. Bovine Tuberculosis is endemic in the UK and there has been a steady increase in incidence since 1984 (Gilbert et al., 2005). This increasing incidence has resulted in a similarly increasing cost to farmers and society (DEFRA, 2014). Suitable policies for the control of bTB has been the subject for intense debate (Schiller et al., 2011), including the controversial proposals for the culling of badgers which can transmit infections to cattle (Godfray et al., 2013), and which many believe represent a reservoir for bTB. One option is the large-scale vaccination of the cattle population (Waters et al., 2012). However, because of the EEC regulation, this has so far not been an option. Recently, new tests for bTB have become available that can diagnose infection in vaccinated animals. These tests are able to distinguish between antibodies produced in response to vaccination, and antibodies produced in response to a natural infection, so-called DIVA tests. They have been evaluated in lab settings as well as in controlled experimental infections, and have been found to have a very high specificity, as well as a good sensitivity (AHAW, 2013; Conlan et al., 2015). The next step in the evaluation is to test their performance in realistic field conditions, necessitating a large, long-term trial involving production farms in the UK.

This section describes the effect that design choices, vaccine efficacy, true test sensitivity and true test specificity have on the sample size required to attain a certain level of precision for the estimated test sensitivity and specificity, using the latent class analysis framework described above. A particular goal for the analysis is to estimate the sample size needed to able to be able to reliably prove that the novel DIVA test reaches a 99.85% threshold.

This threshold was established in (Conlan et al., 2015) by using individual-based modelling to investigate the cost-effectiveness of introducing vaccination of bTB in the UK under a range of different scenarios. False positive tests incur a large cost in terms of unnecessary use of containment efforts and full-herd tests. Conlan et al. thus concluded that a DIVA test would need to be reach 99.85% specificity, for the balance between a reduced clinical burden and an increased cost of false-positive diagnosis (with subsequent actions) to be beneficial.

The previous studies of the DIVA test indicate that the true specificity is likely above 99.9%, making it likely that it would be possible to begin routine vaccinations. However, in order to provide reliable evidence that this is the case in real-world conditions, any estimates of specificity produced by the trial should ideally indicate that the 95% credible interval for the posterior estimate is above 99.85.

2.3.1 *Data simulation*

As in the description of the framework, the following assumptions were made when analyzing the potential outcomes of a DIVA trial:

- The study is composed of one "high prevalence" population and one "low prevalence" population, but the exact prevalences are unknown.

- Each population is then split into two subpopulations; one consisting of vaccinated individuals, and one consisting of non vaccinated individuals.
- It is assumed that there are two different tests used in the trial, both DIVA capable. It is further assumed that these tests had differing sensitivity and specificity, but that sensitivity and specificity was not affected by the vaccination status of the animal. The analysis used a balanced design, with equal number of individuals in each subgroup, as it was believed that this represented the most likely protocol to be adopted for the field trials.

Data were simulated using the R software package (R Core Team, 2012). During data generation, it was assumed that the trial was run on four groups of equal size: vaccinated and unvaccinated cattle from either a high prevalence (modelled as a prevalence of 5%), or a low prevalence (modelled as 2%) population. In addition to this data, it was also assumed that there would be a number of animals taken from breakdown herds that had been tested positive using the established skin test and could be considered "gold standard" positive, as well as a number of animals taken from certified bTB-free farms that could be considered "gold standard" negative. These smaller populations of known positive and known negative animals was included because the results from animals with known disease status could be expected to provide information on sensitivity and specificity that could replace test results from a large number of tests from animals with unknown status, thereby decreasing the required size of the study. The other reason for including such animals is to provide a safe lower bound for the specificity, estimated in a way that is incontrovertible; an additional safeguard to guarantee a baseline of useful results. Finally, it was felt that having preliminary estimates of the test properties would be useful for farmers to be convinced of the benefit to being part in the vaccination study. Data was generated from scenarios with different combinations of parameters: eight different values of sample size were used between 30 000 and 100 000 animals in 10 000 step intervals; 10 different values for true specificity between 0.9990 and 0.9999, and for the gold standard combinations of either 30 positive and 100 negative animals, 300 positive and 1000 negative, or 1500 positive and 5000 negative animals were used. 100 different data sets were randomly generated for each unique combination of parameter values.

2.3.2 *Results*

Recall that we assume that the true specificity of the DIVA test is above 99.9%, and that the 95% credible interval for the posterior of the estimates of specificity produced by the trial should be above 99.85%.

In order to provide a context for the results of the sample size calculations, one can consider a hypothetical scenario where a perfect test is available for use on all animals in the study. Table 2.1 shows the relationship between sample size and width of the confidence interval for four different specificities, assuming that the imperfect DIVA test could be tested on gold standard negative animals.

Table 2.1: *This table indicates the relationship between sample size and precision for evaluating specificity of a diagnostic test with a gold standard approach. Numbers are taken from the appendix written by Innocent, McKendrick, and Rydevik of the Triveritas Ltd for a consortium (2015) report to Defra and the Welsh Government. These are based on an exact formula for sample size calculations of binomial distributions (Armitage et al. (2008), p117), and gives the lowest sample size for which there is an 80% probability that a random sample would produce a confidence interval of equal to or less than the required width.*

true sensitivity	width of credible interval	sample size
70%	+/-5%	353
70%	+/-1%	8230
75%	+/-5%	320
75%	+/-1%	7382
99.5%	+/-0.5%	1226
99.5%	+/-0.2%	5974
99.85%	+/-0.5%	696
99.85%	+/-0.2%	2508
99.99%	+/-0.3%	1162
99.99%	+/-0.2%	2034
99.99%	+/-0.1%	6150
99.99%	+/-0.05%	20320

While in reality there are no true gold standard tests available for bovine tuberculosis, post mortem identification of lesions with subsequent identification of the causal agent is considered to have 100% specificity. One approach would therefore use the results from post-mortem examinations to identity assuredly positive animals. Another approach to ensuring that the test is evaluated on guaran-

teed negative animals, instead of using a gold standard test for classification, is to apply the test to animals that are highly unlikely to have been exposed to bTB. One example would be a low-risk-classified farm in Scotland (which is classified as officially bTB-free since 2009). While these approaches are too expensive to carry out on a full scale, we can use both approaches on a smaller scale and use the results to inform our priors for the latent class analysis.
Figure 2.1 shows the effect of sample size on the credible interval of the estimated specificity of one of the DIVA tests. Each box represents the results from fitting the latent-class model to 100 different randomly generated data sets with a varying number of tested animals, assuming that the true specificity of the test was 0.999, and that in addition to the main study, 30 animals that were gold standard positive and 100 animals that were gold standard negative were tested to derive priors for the specificity and sensitivity of the DIVA test (see methods for details). We would desire that the study has at least 80% power, i.e. at least 80% chance that the posterior estimate would show the result desired if the hypothesis that is being tested is true. In the case of the current study, this translates to being able to have at least an 80% chance to prove that the DIVA tests are good enough (i.e. better than 99.85% specificity). For the figure, this translates to the lower end of the box being above the 99.85% line. Clearly, even with 100 000 animals, the study does not reach sufficient power.
Figure 2.2 shows how distribution of the lower bound of the credible interval change under different assumptions of true specificity and the number of gold standard tested animals used to inform the prior distribution for specificity and sensitivity. This figure clearly indicates that the most crucial parameter is the true specificity, which has to be much higher than the threshold efficacy in order to produce data that can reliably demonstrate that the specificity is, in-

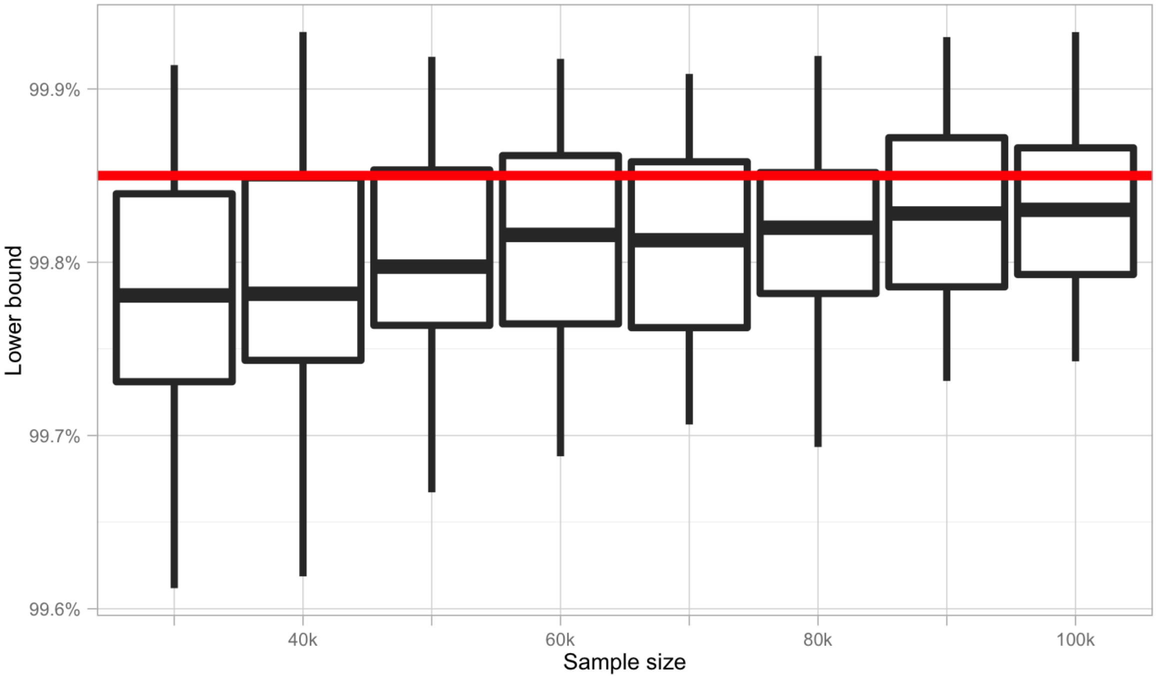

Figure 2.1: *The above box plots describes the distribution of the lower bounds of the credible intervals for the estimated specificity of the DIVA tests. Each box represents the results from fitting the latent-class model to 100 different randomly generated data sets with a varying number of tested animals, assuming that the true specificity of the test was 0.999, and that in addition to the main study, 30 animals that were gold standard positive and 100 animals that were gold standard negative were tested to give a baseline indication of the sensitivity and specificity (analysed as described in the methods). The box plot parameters were modified from a standard box plot to display 80% and 95% quantiles. Thus, the box represents the central region in which 60% of estimate lower 95% bounds of the estimated specificity $\hat{Sp}$ from 100 simulations falls. The whiskers indicate the extreme 95% range of the lower 95% credible interval, and the middle mark indicate the median. The red horizontal line indicate the 99.85% threshold above which a DIVA test would be cost effective when implemented - the lower edge of a box above this line would thus indicate a power of more than 80% for demonstrating that the test used is cost effective, based on data with the number of samples indicated.*

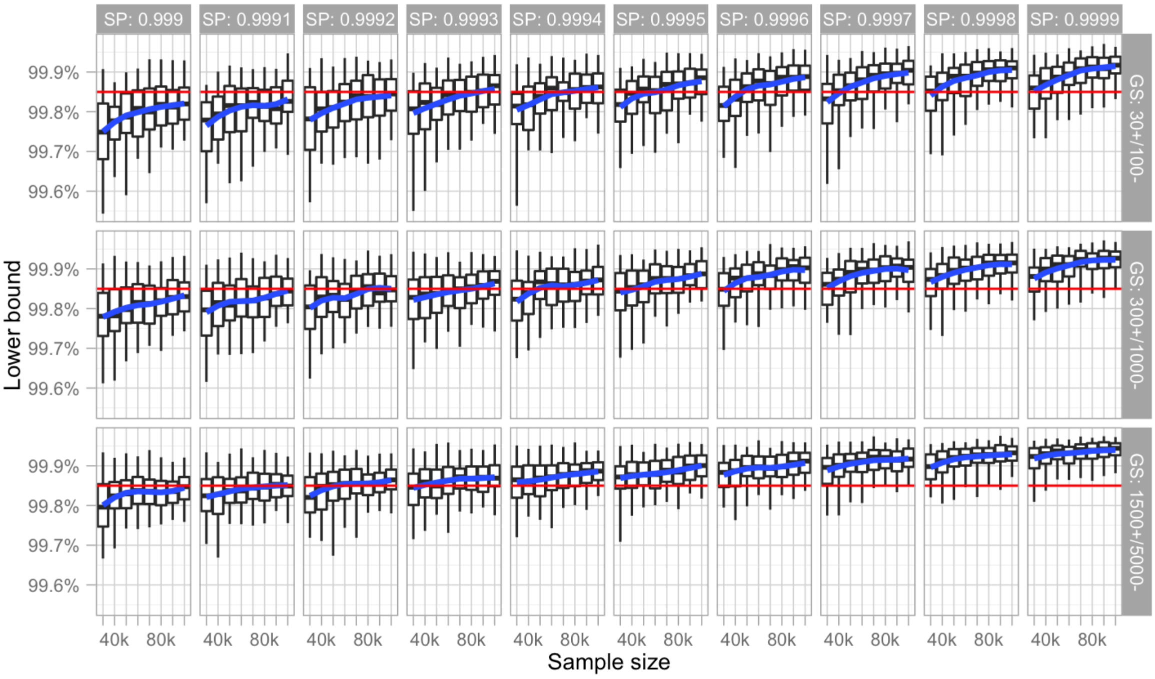

Figure 2.2: *This figure shows how the credible interval is affected by different conditions, and which conditions are required to reach 80% power. As in figure 1, the lower edge of the boxes signify the threshold for which 80% power is reached. The red line indicate the 99.85% specificity threshold - Columns indicate varying specificity; rows indicate varying sizes of a gold standard pilot study. The blue line indicate the trend of the medians as a function of sample size.*

deed above the threshold. From the results shown in the figure, it is possible to say that if the true specificity is high enough, above 0.9995, sufficient power can be reached with a combination of a pilot study using 1500 gold standard positive and 5000 gold standard negative animals; and a study on production animals including a total of 50 000 animals. If the DIVA test has near perfect specificity , >0.9999, and only on very rare occasions give false positive results, then sufficient power can be reached with a pilot study on 300 positive and 1000 negative animals, and a study on production animals including 30 000 individuals.

As described in the methods section above, in addition to estimating sensitivity and specificity for the two DIVA tests, the vaccine efficacy parameter is also estimated. For the middle-of-the-road scenario with an assumed true specificity of .9999, a sample size of 30 000 animals, 300 positive and 1000 negative gold-standard tested animals, and a true vaccine efficacy of 0.6, the mean posterior estimate of the vaccine efficacy had a root mean square error (RMSE) of 4.7 percentage units based on 100 different simulated data sets, a relative error of about 8% . The mean width of the 95% credible interval was 20.6 percentage units. The true sensitivities and specificity of the diva tests have little influence on these estimates (with an assumed true specificity of 0.999 , the RMSE was 4.3 percentage units, a 7% relative error, and the mean width was 21.8 percentage units); similarly for the number of gold standard tested animals used. However, the true vaccine efficacy has a strong influence on the width of the credible intervals. Figure 2.3 shows how the 95% posterior credible interval varies for different sample sizes and different true levels of vaccine efficacy.

As the true vaccine efficacy increases, so does the absolute size of the confidence interval. To estimate a 90% vaccine efficacy to within +/- 10 percentage units, a sample size of at least 70 000 animals

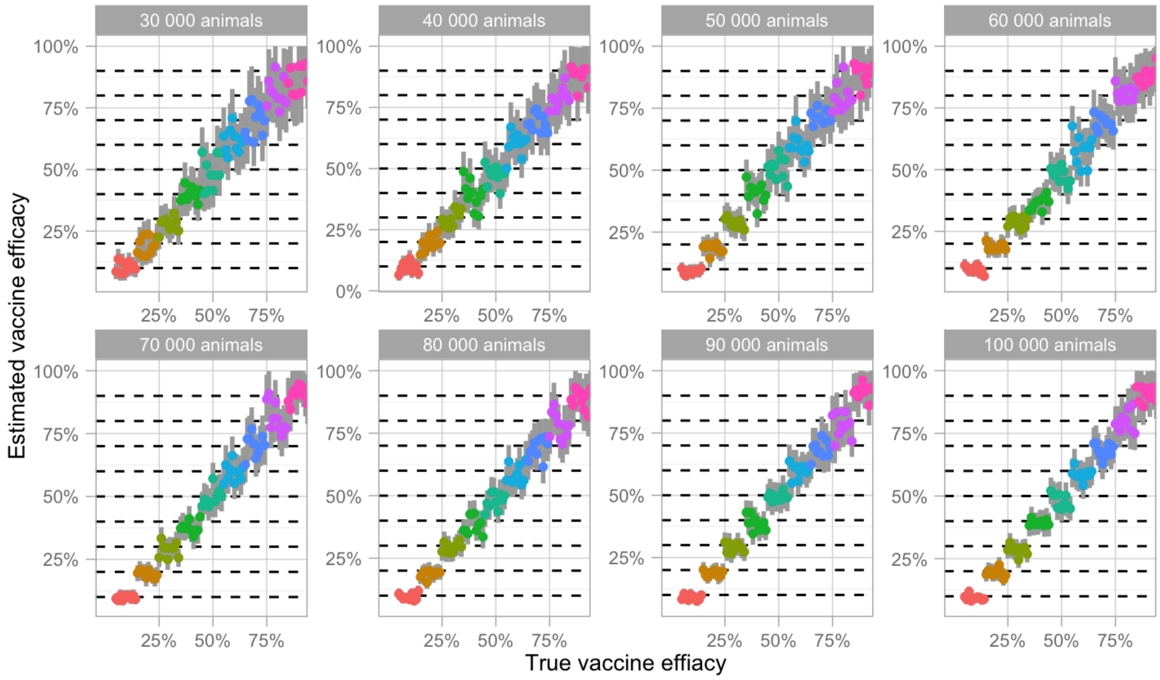

Figure 2.3: *Each vertical line represents the 95% posterior credible interval of the inferred vaccine efficacy parameter for different sizes of the pilot study, and for different true values of the parameter. The true specificity of the DIVA test was fixed at 0.9999, and the number of gold standard tests were fixed at 300 positive and 1000 negative animals. For each combination of sample size and true vaccine efficacy, ten different data sets were simulated. The horizontal axes are thus on on a categorical scale that indicate the level of true vaccine efficacy, ranging from 10% to 90%. The grey lines indicate the 95% posterior credible interval of the efficacy parameter for each data set, with the dots indicating the posterior mean, coloured according to the level of assumed true vaccine efficacy. The dashed horizontal lines indicate vaccine efficacy levels in 10%-unit intervals*

would be needed, while a 50% vaccine efficacy could be estimated to within 10 percentage units with 50 000 animals. For less effective vaccines, 30 000 animals would be more than enough to estimate efficacies of less than 30% to within 10 percentage units. The increasing size of the confidence interval is likely due to higher vaccine efficacy leading to fewer positive animals, making the evaluation of test sensitivity and specificity more uncertain, and thus as a consequence all parameters of the posterior.

2.4 DISCUSSION

This chapter has expanded on previous work on the Hui-Walter paradigm by demonstrating how to estimate the sensitivity and specificity of two tests with different properties in vaccinated and unvaccinated populations. In addition, it demonstrates that this can be done for a vaccine with unknown efficacy, using two populations with different levels of disease, each population divided into vaccinated and unvaccinated subpopulations.

Results from the sample size evaluation highlight that the ability to prove that the DIVA tests have a specificity high enough to pass the cost-effectiveness threshold of 99.85% established by Conlan et al. (2015), is highly dependent on the actual specificity of the tests. It is a challenging criteria that will require either a very large trial, or that the DIVA test is better than the required threshold by a large margin, or some combination of the two. A sample size of 50 000 animals in the trial would only be likely to clear the 99.85% threshold if the true specificity is above 99.99%, with sample sizes of 500 000 animals required if the test is "only" 99.9% sensitive, equivalent to 5% of the UK cattle population which would be a very large trial. Moreover, without the addition of animals with known true infection status the required sample sizes are even larger.

Note that these study sizes were calculated under the assumption that it is needed to prove "beyond reasonable doubt" that the vaccine is cost effective. It could be argued that the decision of whether to vaccinate or not should be decided based on which course of action is "most likely" to be economically beneficial; there is a cost to vaccinating, but there is likewise a cost to not vaccinate. By assuming that evidence "beyond reasonable doubt" is needed, the implicit assumption is that the status quo is preferable in the absence of considerable evidence to the contrary. If, instead, a "most likely" approach to the decision were taken, this would allow each option to be chosen based on their own merit. In technical terms, such an approach would be translated to requiring that at least 50% of the mass of the posterior probability distribution fall above the 99.85% threshold. If such a decision criteria would be seen as sufficient, the required study size to reach 80% power could be reduced considerably (preliminary investigations indicate that the study size could potentially be reduced by a factor between 5 and 10, assuming other study parameters remain the same).

The Hui-Walter paradigm is well established for estimating the properties of diagnostic tests in the absence of a gold standard. By using vaccinated and non-vaccinated individuals as two distinct populations, it becomes possible to estimate the efficacy of a vaccine in addition to the properties of the diagnostic tests. To our knowledge, no previous paper has made use of Hui-Walter type models in this way before. This could be useful in situations where novel vaccines and diagnostic tests are used to control a disease, where the vaccine effect has been tested in controlled studies, but the field efficacy of the tools is unknown. Such a situation is likely to occur for rapidly developing outbreaks, such as the recent Ebola outbreak (WHO Ebola Response Team, 2014), where both vaccines and new diagnostic tools are developed in response to the occurrence of the outbreak. While the resulting efficacy estimates are not as precise as those from a study designed for efficacy evaluation, the simulation results presented above indicate that they would still be sufficient to evaluate usefulness to within +/- 10 percentage-units.

An important assumption made in the use of the Hui-Walter paradigm is the assumption of independence between the results of the two different diagnostic tests, conditional on the true status on the animal. The conditional independence assumption is necessary in the case of two tests and two populations, or as in our case, between four subpopulations with four unknown sensitivity/specificity parameters. However, if the tests in question both measure the same type of response to infection, this assumption can be questioned.

In Toft et al. (2007), it is suggested that if the assumption of conditional independence is unreasonable, one can consider the use of three different tests. In the DIVA situation, this would be difficult, but a possible strategy is to conduct post-mortem evaluations on animals which have been indicated as positive. In such a situation, it is enough if one test is conditionally independent from the others.

Clegg et al. (2011) used a three-test latent class approach to evaluate the performance of the interferon-γ test, the SICCT test, and a multiplex antibody assay for detecting bovine Tuberculosis (all tests described in OIE (2009)), assuming that the SICCT test was conditionally independent from the two others. Unfortunately, in our situation the SICCT test would show vaccinated animals as positive; however, should another DIVA test be developed based on another principle from the two tests used in this study, a three-test approach could be used to estimate possible dependencies between the existing tests.

An additional consideration is the time between testing animals. While the analysis in this chapter assumes that all animals are infected randomly and independently from each other, and that there is a dichotomy between ″infected″ and ″non-infected″ animals, it is likely that test sensitivity is dependent on the time since infection. The disease progression after infection with *M.bovis* can take up to a year (Domingo et al., 2014), and can be usefully considered to pass through several distinct stages. A diagnostic test is likely to have different sensitivity at each stage in the disease. Therefore, if an animal is infected and tests negative, it is likely that subsequent tests close enough in time will also be negative, and the nominal sensitivity would therefore overestimate the herd-level sensitivity. A possible way to circumvent this would be to model the measured test response as a continuous variable instead of using the infected/non-infected classification. This could be done using either a mixture model approach, or by incorporating information on how the test response would depend on the time since infection.

The work in this chapter has shown that with the introduction of a vaccine for a disease, it is possible to estimate the vaccine efficacy using two tests with unknown sensitivity and specificity, demonstrating that tests that have not yet been assessed against a gold standard can still be used in a situation where a vaccine is expected to interfere with the test results. It has further been shown that such tests can be used even if the vaccine efficacy is unknown. In this situation, a trial of diagnostic tests would also produce estimates of a vaccine efficacy. Finally, in regards to the DIVA tests for bovine TB, the results indicate that attempting to establish confidence intervals for the sensitivity and specificity of new diagnostic tests is likely to only provide partial evidence for demonstrating that a test has reached a particular specificity threshold, when the required specificity is close to 100%. In such situations, the evidence from field studies of test properties needs to be combined with laboratory studies, the fundamental science underlying the tests and other sources of information to determine whether the true specificity is sufficiently high for the intend ended use.

2.5 FROM BINARY TO CONTINUOUS TESTS

In this chapter, the assumption has been that diagnostic tests produce binary results, and implicitly that each individual is either infected or noninfected. Often, a binary diagnostic test is generated by taking a continuous measurement of some quantity, such as the level of antibodies in the blood, and checking to see whether it is above a pre-defined cutoff value. By converting a continuous measurement to a true/false value, we are in essence throwing away information (Fig 2.4). However, knowing the quantitative value of the test measurement can tell us more than a binary test result can. A quantitative test result can allow us to estimate the overall distribution of test values in the population, evaluate the variability of the binary classification for a given test value, and increase the precision of population incidence or prevalence estimates.

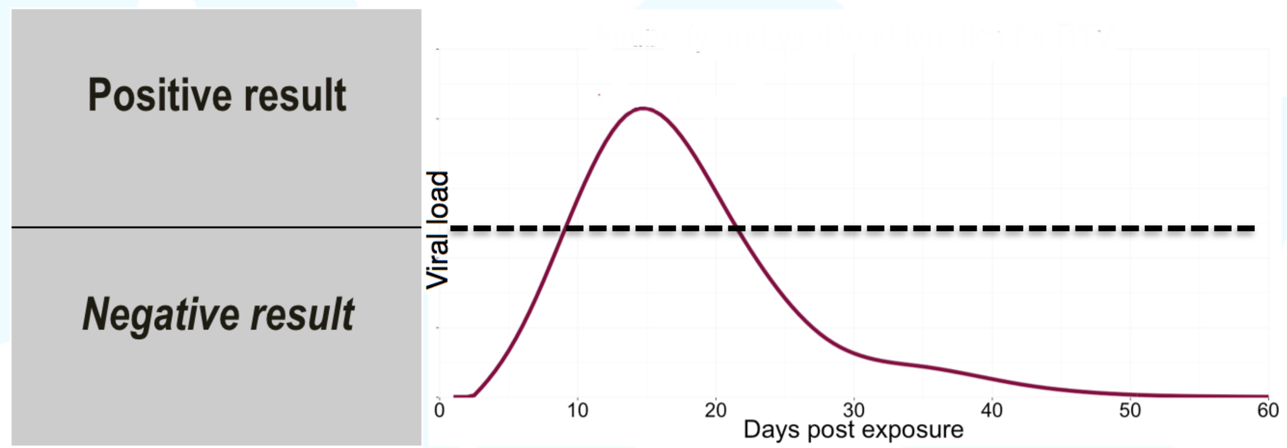

Figure 2.4: *This figure shows a typical curve for the development of viral load test measurements over time . A positive or negative test result can be generated by setting a cut-off value (black line). Above the line, the test is "positive"; below the line, the test is "negative".*

The binary classification is useful and straightforward. In clinical practice, such an approach makes perfect sense in order to decide on appropriate treatments. In modelling, the simplicity of the binary classification allow for tractable expressions for the disease dynamics. This is embodied by the basic SIR model where individuals move from uninfected/susceptible, to infected, to non-infected and recovered/removed. This type of model has a long history as a tool for understanding the dynamics of infectious diseases (Hethcote, 2007).

However, in several cases, this categorization is too simplistic. As a simple example, consider chickenpox, which lie dormant after the disease has run its course, kept in check by antibodies. If these wane, the virus may re-emerge as shingles. For bovine tuberculosis, as discussed in the previous section, the disease progresses via several stages, and many individuals carry the mycobacterium indefinitely without suffering ill health. Such processes are important aspects of disease dynamics and host responses, but the binary paradigm lacks the expressive power to capture such nuances. In mathematical modelling of disease dynamics, an approach is to include more stages representing recovered-but-immune, asymptomatic, asymptomatic but infectious, etc. In statistical inference, the diagnostic test can be modelled as a categorical response variable, or as a continuous response following some kinetic curve.

The following chapters in this thesis continue to model situations where each individual has been tested using multiple diagnostic tests. However, as we will see, instead of binary tests, tests with a continuous, quantitative response are considered. In this way, much richer types of analyses can be conducted. Chapter 3 describe how historic linear incidence trends can be estimated from cross-sectional data. Chapter 4 describes how it can be applied in an epidemic setting. These two chapters serve to demonstrate the potential analyses that can be done when going beyond the classical terminology of binary diagnostic testing of disease. Finally, chapter 5 suggests an approach to estimating test kinetics of multiple given observed data, which can be considered an extension of the estimation of sensitivity and specificity described in this chapter.

CHAPTER 3: HINDCASTING TRENDS OF INFECTION FROM MULTIPLE TEST DATA

Abstract
This chapter describes a novel statistical approach to combine data from multiple diagnostic tests with different temporal characteristics to hindcast the historical unobserved trend of an endemic infectious disease. Assuming a cross-sectional sample of individuals infected with a pathogen, a Bayesian MCMC approach is used for estimating time of exposure and the overall epidemic trend in the population prior to the time of sampling. It is demonstrated how to utilize this approach to distinguish between decreasing and non-decreasing trends. Further, the chapter describes results of applying this for idealised pairs of diagnostic tests, based on different host-pathogen dynamics. Finally, we discuss the benefits of this novel methodology for the management of infectious diseases, and for evaluation of policy interventions.

3.1 INTRODUCTION

Pathogens are one of the major contributors to the burden of disease in humans (Lopez et al., 2006), have a substantial economic impact on the livestock industry (Stott, 2003), and can be a serious threat to conservation and management of wildlife populations (Daszak et al., 2000). A crucial component of efforts to control endemic disease is the use of infectious disease surveillance for tracking trends and evaluating the effect of control measures. The current state of human disease surveillance has been characterized as deficient in terms of both coverage and reporting speed (Butler, 2006). The more complex settings typical of livestock and particularly wildlife systems tend to result in available surveillance data being sparser still for animal disease (Mörner et al., 2002; Perez et al., 2011; The Royal Society, 2002)

The structure of a functioning disease surveillance system is complex, with a string of tasks that need to be accomplished before a case is recorded in a database and becomes available to epidemiologists and policy makers. However, a crucial part is the use of diagnostic tests to identify and confirm the type of pathogen that caused infection. This and the following chapter will argue that combining two or more quantitative diagnostic tests with trajectories of the development of average test measurements following infection provide substantial additional information that can be used to estimate historic patterns of infection. Current analyses typically treat diagnostic tests as binary classifiers of infected/non-infected individuals. However, the behaviour of diagnostic tests are more complex as they typically return a result in terms of a non-binary response level. Moreover, the expectation of this test response varies as a function of time since infection. To make use of such data and realise these benefits, a novel statistical approach is introduced for recovering population-level trends of infection even from only cross-sectional data by combining knowledge of the dynamic characteristics of multiple diagnostic tests to infer the timing of infection events for individuals. The process of recovering such trends will be referred to as "hindcasting", following terminology established in other papers (Banakar et al., 2011; Kleczkowski and Gilligan, 2007; Wethey and Woodin, 2008) for reconstructing historical trends from currently available data. This chapter will focus on the potential use of hindcasting in the case of endemic diseases, while chapter four describes the potential for hindcasting in an epidemic setting.

Changes in the epidemiology and/or incidence of endemic pathogens are ideally tracked through the use of routine, ongoing surveillance. However, in a number of situations and for a number of pathogens, such ongoing surveillance is either non-existent or limited in its ability to provide a full, unbiased view. For some diseases, the epidemiology is known and the disease is considered important, but surveillance relies on diagnostic measures which are either expensive and underutilized, or lacking in sensitivity and/or specificity. One such example is the disease scrapie in sheep. Scrapie is a prion-spread disease with a very long incubation period, and difficult-to-detect symptoms. In the USA, scrapie has decreased from a 0.2% prevalence to less than 0.05% between 2003 and 2009 thanks to introduced policy measures (United States Department of Agriculture, 2010). However, there are substantial biases in reported prevalence numbers, raising the need for additional surveillance measures (Del Rio Vilas and Pfeiffer, 2010). Another pathogen, endemic in most of Europe, is *Mycobacterium avium* subsp. *Paratuberculosis*, also known as "Johne's disease". Paratuberculosis infections are asymptomatic for a long period of time, only detectable after some period and with the use of specifically targeted tests (OIE, 2014). Reported prevalences across Europe vary widely, from 0.1% to 20%, largely owing to the difficulty of diagnosis (Nielsen and Toft, 2009). With these kinds of so-called iceberg disease systems (see Section 1.2), where routine surveillance only captures a small proportion of actual cases, there is a strong need for alternative strategies that can ensure that the trend measured by routine surveillance systems is representative of the full epidemiology of the targeted disease system.

There are a number of endemic diseases considered to be of low importance and therefore not targeted by surveillance. When such a disease suddenly gains importance, because of increasing prevalence induced by changes such as mutation of the pathogen, or due to realizations of the extent of its economic impact, the ability to rapidly gain an understanding of the historic trends would be extremely useful in prioritizing and targeting interventions. This can be the case even for high profile pathogens such as the H5N1 flu virus, where the threat of silent spread in poultry flocks is a serious concern (Savill et al., 2006). Some pathogens have a very high incidence of undiagnosed infection, where the pathogen circulates widely in the population causing non-specific disease. Salmonella (Simonsen et al., 2011) and Pertussis (Hallander et al., 2009) are two examples of human diseases where the true extent of infections have been unknown until fairly recently. Sexually transmitted infections are also often under diagnosed because of social stigma associated with testing. Chlamydia is a disease with a significant disease burden in most parts of the world (WHO, 2012a), and where the prevalence in women is much better known, and often reported to be higher, than in men for whom the testing rate is much lower(see e.g. the introduction of Götz (2005)).

For many endemic diseases, policies are put in place to reduce incidence or eradicate the disease - either locally, as with bovine viral diarrhea(BVD) in Scandinavia and Scotland (Ståhl and Alenius, 2012); or globally, as happened with Rinderpest in Cattle (FAO, 2013) and smallpox in humans (WHO and Global Commission for the Certification of Smallpox, 1980). Measuring the impact of implementation of such policies is needed to ensure that eradication efforts are on the right track. High costs restrict implementation of longitudinal surveillance programmes whereas cross-sectional studies of disease are more common. Therefore, methodology that could infer temporal trends from cross sectional data would be extremely beneficial. The application of the hindcasting techniques introduced here could be used to extend the utility of such cross-sectional studies to fulfil some of the objectives of an ongoing surveillance system (see Section 1.2 for a discussion of such objectives).

Several papers have recovered limited historical characteristics of the spread of pathogens from cross-sectional data using a single diagnostic test, typically an antibody test. For example, Giorgi et al. estimated the time of the start of a local HIV outbreak under assumptions of exponential growth of viral load (Giorgi et al., 2010). Others have exploited information on diagnostic test kinetics, i.e., the pattern of diagnostic test values during the course of infection, to estimate average incidence rates. Example includes the use of antibody test kinetics to estimate sero-incidence rates for influenza (Baguelin et al., 2011), salmonella in cattle (Nielsen et al., 2011) and salmonella in humans (Simonsen et al., 2008). One challenge in these kind of studies is that the relationship between the magnitude of signals from diagnostic tests and time since infection is usually not monotonic; the signals tends to increase and then decrease. This means that the inverse problem of estimating time since infection given a test value is non-unique and although this can be framed as a statistical problem the resulting inference is highly uncertain (Giorgi et al., 2010; Simonsen et al., 2009), limiting what can be estimated from test data. However, there are often several diagnostic tests available that target different aspects of the multi-faceted dynamic interaction between host and pathogen (Casadevall and Pirofski, 2001), and would thus exhibit different test kinetics. That is, the profile of test responses, as a function of time since infection, will differ depending the underlying diagnostic used. This means that, in principle, we can generate a unique signal for a given time since infection by combining results of several diagnostic tests that respond on different time scales. Here, this fact is exploited to develop a more robust statistical approach for analyzing cross-sectional field data from two or more diagnostic tests. Empirical infection models that characterize test kinetics are used to infer the time since infection for each individual. While there is large uncertainty in the estimated infection time for each individual, the combined estimates from multiple individuals describe the overall population-level distribution of infection times, which can be used to estimate the overall trend of incidence. In an endemic setting, trends of infection are often gradual, and can be approximated by a constant change per time unit (month, year, decade). The chosen approach in this chapter was thus to posit that the incidence follows a linear trend with some slope. Section 3.2 develops the statistical framework for hindcasting in general, while section 3.3 details the mathematical consequences of assuming a constant linear trend with reinfections on inference of the trend from cross-sectional data. Section 3.4 details the choice and implementation of test kinetics. Section 3.5 details results from applying the framework to data simulated under a range of different scenarios. Finally, section 3.6 discuss the implications of the results and the hindcasting framework.

3.2 STATISTICAL FRAMEWORK

The statistical framework used for hindcasting in this thesis assumes test data y_{nk} from multiple disease diagnostics indexed by $k = 1, \ldots, K$ on individuals $i = 1, \ldots, N$. Each individual is assumed to have been tested at some time t_i, after having been exposed to the pathogen at some earlier time e_i. It is further assumed that these individuals are chosen in an unbiased, random manner from a larger population. Each diagnostic test is assumed to return a value measured on a continuous scale, which might, for example be the highest dilution at which antibodies are detected in a serological test. Without loss of generality, these values are assumed to be rescaled to the unit interval [0,1].

Initial exposure to a pathogen is the start of a complex dynamical process within the host. Such internal host-pathogen interactions can be conceptualised as a multivariate process that depends on the time since initial exposure. Each diagnostic test is assumed to target the state of a different component of this process so that each test k carried out at time t_i on individual i can be modelled as a latent variable $l_{ik}(t_i, e_i) = l_{ik}(d_i)$, with each test having differing but correlated response patterns over the time since initial exposure $d_i = t_i - e_i$. these latent variables are modelled using results from experimental infection studies for a given host-pathogen system, where the length of time since initial infection d_i is known.

The known data, across all individuals in the sample, comprises a set of test results denoted by $Y = \{y_{ik}\}$ with sampling times $T = \{t_i\}$. The aim is to infer the unknown set of exposure times $E = \{e_i\}$, using information on the behaviour of the latent processes $L = L(T, E) = \{l_{ik}(t_i, e_i)\}$ generating the test results. In the hindcasting model used here, L represents the expected value of the test results given e_i and t_i. Note that when describing these sets the limits of each index $k = 1, \ldots, K$ and $n = 1, \ldots, N$ are implicit.

Assume that the sampling times T and the observed test values Y,are known whereas the quantities L and E are assumed to be subject to uncertainty and variation. There are thus three components to the statistical model: a latent process model $P(L|T, E, \theta_L)$ describing uncertainty and variation in the host-pathogen interaction process within the host in terms of the time since initial exposure; a testing or observation model $P(Y|L, \theta_Y)$ describing the distribution of results from tests carried out on the hosts conditional on the internal latent process; and an epidemic trend model $P(E|T, \theta_e)$, describing the historical development of the epidemic in terms of the distribution of exposure times in the sampled host population, at the time of sampling. The specific implementations of each of these components in the linear trend setting is described in the next two sections. Combining the three parts of the model, the full data likelihood given an observed data set $\{Y, T\}$ is written as $P(Y, E, L|T, \theta) = P(Y|L, \theta_Y)P(L|T, E, \theta_L)P(E|T, \theta_E)$, where $\theta = \{\theta_Y, \theta_L, \theta_E\}$ Thus the likelihood combines models for testing with those for within and between host pathogen interactions.

According to Bayes' theorem, the so-called posterior distribution for the unknown parameters is proportional to the data likelihood and prior $P(\theta)$. Using the parameters of interest θ, the latent process L, the exposure times E, given the observed test data Y and sampling times T, the posterior distribution can be described by the equation

$$P(L, E, \theta|Y, T) = (P(Y, E, L|T, \theta)P(\theta))/(P(Y, T))$$

Within the Bayesian framework all inference is based on the posterior. The prior $P(\theta)$ can result from previous measurements or expert opinion, and represents knowledge about the values of the parameters before any of the data used in the likelihood is observed. In what follows, the simplifying assumption will be made that the latent process L is modelled by a known deterministic function of T and E. This means that the term $P(L|T, E, \theta_L)$ drops out of the likelihood which then simplifies to

$$P(Y, E|T, \theta) =$$

$$P(Y|L(T, E), \theta_Y)P(E|\theta_E)$$

, and the posterior becomes

$$P(E, \theta|Y, T) = \frac{P(Y|E, T, \theta_Y)P(E|\theta_E)P(\theta_Y)P(\theta_E)}{P(Y, T)}$$

Note that under this notation any parameters defining the deterministic latent process $L(T, E) = l_n k(t_n, e_n)$ are suppressed since they are not inferred i.e. $\theta = \{\theta_Y, \theta_E\}$.

In both cases above the normalization factor $P(Y, T)$ is typically unknown and computationally expensive to calculate. However, standard Markov Chain Monte Carlo (MCMC) methods circumvent this problem and are able to generate samples from the posterior even though the normalization is unknown (see Section 1.5.2).

3.3 PARAMETRIZATION OF A LINEAR TREND OF INCIDENCE

3.3.1 *Distribution of times since infection (tsi) under linear trend*

The hindcasting framework estimates the historic trend of infection $f(t)$ in the population from cross-sectional data. The trend $f(t)$ describes the incidence over time if the cases are reported continuously. If, instead, infected cases are sampled at a single point in time, this sample will consist of individuals that have been infected some time in the past. The objective is then to estimate the distribution of these times since infection("tsi"), based on collected test measurements. The probability density function distribution is hereafter denoted $f_{tsi}(t)$. In order to simplify calculations, the time is measured backwards; the time of the cross-sectional sample is defined as $t = 0$, and an infection that occurred 10 time units ago is denoted $t = 10$.
It should be noted that because we only have one observation per individual, it is only possible to estimate one time since infection, and the population level distribution of times since infection is in effect the distribution of times since *last* infection. Thus, this modelling approach implicitly assumes that any reinfection of an individual resets the "clock" of the infectious disease dynamic to zero.The most basic scenario used for hindcasting a disease trend represents an endemic disease, with cases occurring at a constant rate. Formally, this scenario can be defined by assuming that the entire population is exposed to a constant force of infection λ. For a randomly observed

infected individual, it can then be shown (Muench, 1934) that the time since infection f_{tsi} is distributed according to an exponential distribution with rate parameter λ, $f_{tsi}(t) = \lambda e^{-\lambda t}$.

The assumption of "last infection resets the clock" implies that later infections hide earlier infections, which leads to an apparent increasing incidence, captured by the above expression. If, instead, it is assumed that the first infection confers immunity, as so only the first exposure to the pathogen is observed, this will mean that *later* infections are blocked or invisible, leading to an apparent decrease in incidence. This can be described by changing the sign of the exponent in the above expression so that the equation becomes $f^{\star}_{tsi}(t) = \lambda e^{\lambda t}$. However, note it is now necessary to explicitly include an upper limit on time (such as the year of birth of each individual), as $f^{\star}_{tsi}(t)$ will otherwise have an infinite integral and not be a probability distribution.

This basic scenario assuming that only the last infection is visible was modified to a scenario in which the force of infection changes over time according to a linear trend. In this, the force of infection λ at a time t is equal to $\lambda(t) = \alpha + \beta t$ as mentioned above. Because of the linearity of the trend, the probability of having been infected during a time period from 0 to T is equivalent to the probability of having been infected under a constant trend of the mean incidence over the period, $\hat{\lambda}(t) = \alpha + \beta T/2$ (See Figure 3.1).

In the constant case, the probability of having been infected during the time period T can be written as $\int_{t=0}^{T} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda T}$. By analogy to the constant case, we can thus write the probability of having been infected by time t assuming a linear trend as $P(I < t) = 1 - e^{-(\alpha + \beta * t/2) * t}$. By taking the derivative of this, the probability density function for the times since infection can be calculated as $f_{tsi}(t) = d(P(i < t))/dt = (\alpha + \beta t) e^{-(\alpha + \beta t/2)t}$.

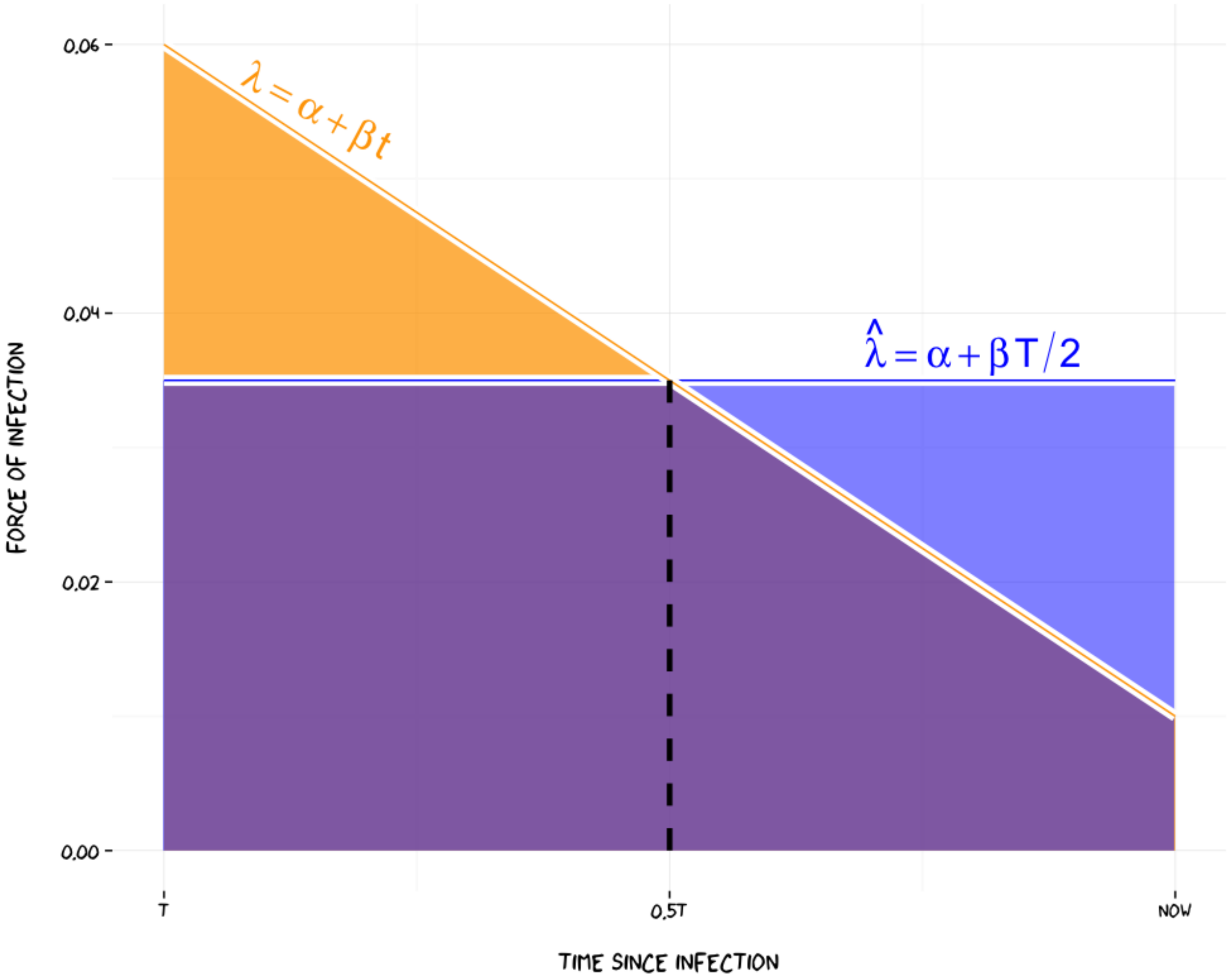

Figure 3.1: *The cumulative force of infection an individual is exposed to over a time period from 0 to T for a linear trend $\lambda(t) = \alpha + \beta t$ (shown by the total area shaded yellow plus that shaded purple), is the same as the cumulative force of infection for a constant force of infection at the intensity equal to that of the linear trend at $T/2$; $\lambda(\hat{t}) = \alpha + \beta T/2$ (shown in blue and purple).*

Furthermore, In the implementation this distribution was assumed to be censored at some time point in the past C, and it was further assumed that it was possible, a priori, to distinguish individuals that had been infected at some point during this time period, from naive individuals. When implementing such a censoring, the equation above needs to be modified by an additional scaling factor $1/(1 - e^{-(\alpha+\beta C/2)C})$, equal to one over the integral of $f_{tsi}(t)$ over the time span $(0, C)$. The full equation used to represent the distribution of exposure times was thus

$$f_{tsi} = (\alpha + \beta t)e^{-(\alpha+\beta t/2)t}/(1 - e^{-(\alpha+\beta C/2)C})$$

As the model is implemented in the Bayesian framework, priors for both the prevalence (α) and trend (β) parameters needs to be specified. In order to provide a prior for the incidence, information about the population size needs to be incorporated. This was done by noting that the number of positive and negative individuals in a population can be approximately described by a binomial model, parameterised by the probability of infection p. Then denote by N_+ the number of positive individuals known to have been infected between the time of censoring C and the time of sampling (defined as $t = 0$), from a population of size N. With an uninformative $Beta(1, 1)$ prior for the probability of infection p the distribution of the probability of infection given the number of positives and negatives observed is $p \sim Beta(N_+ + 1, N - N_+ + 1)$. Under the assumption of a linear trend, the mean incidence over the time period C is equal to the incidence at time $C/2$, $\hat{\lambda} = \alpha + \beta \times C/2$. The proportion p of observed positive individuals are exactly one minus those that had not been infected during any of the time periods up until the time of censoring C. From this observation, the mean incidence $\bar{\lambda}$ per time unit can be derived from the proportion of positive individuals over the time period C by the relationship:

$$p = 1 - (1 - \bar{\lambda})^C \rightarrow 1 - \bar{\lambda} = (1 - p)^{1/C} \rightarrow \bar{\lambda} = 1 - (1 - p)^{1/C}$$

For the trend parameter β, note that if the linear trend model is assumed to hold over the time period C, then the incidence is not allowed to become negative over this time. Using this, the restriction for the trend becomes:

$$\bar{\lambda} \pm \beta \times C/2 > 0 \rightarrow$$

$$\bar{\lambda} > \beta \times C/2 > -\bar{\lambda} \rightarrow$$

$$\bar{\lambda} 2/C > \beta > -\bar{\lambda} \times 2/C$$

Following this, the trend was assigned a uniform prior, $\beta \sim U(-\bar{\lambda} \times 2/C, \bar{\lambda} \times 2/C)$. The intercept parameter α was then simply calculated from trend and $\bar{\lambda}$ via $\alpha = \bar{\lambda} - \beta \times C/2$.

3.3.2 *Properties of the linear-trend-induced distibution of times since infection*

The properties of the distribution $f_{tsi}(t)$ of times since infection are somewhat counterintuitive. Figure 3.2 shows its shape for decreasing ($\beta = 0.05$), constant ($\beta = 0$), and increasing ($\beta = -0.05$) parameter values, holding α constant to 0.1. The first thing to note is that because we are looking backwards in time, coefficients have opposite sign to what at first might seem intuitive . $\beta = 0.05$ denotes that the incidence rate has been decreasing by 0.05 per time unit, whereas $\beta = -0.05$ denotes that the incidence rate is increasing. The second thing to note is that all three curves have a similar upwards slope. The further back in time we look, the less likely we are to find a case that occurred at that time. This is the consequence of the implicit assumption mentioned above that only the time of latest infection for each individual is taken into account, which means that

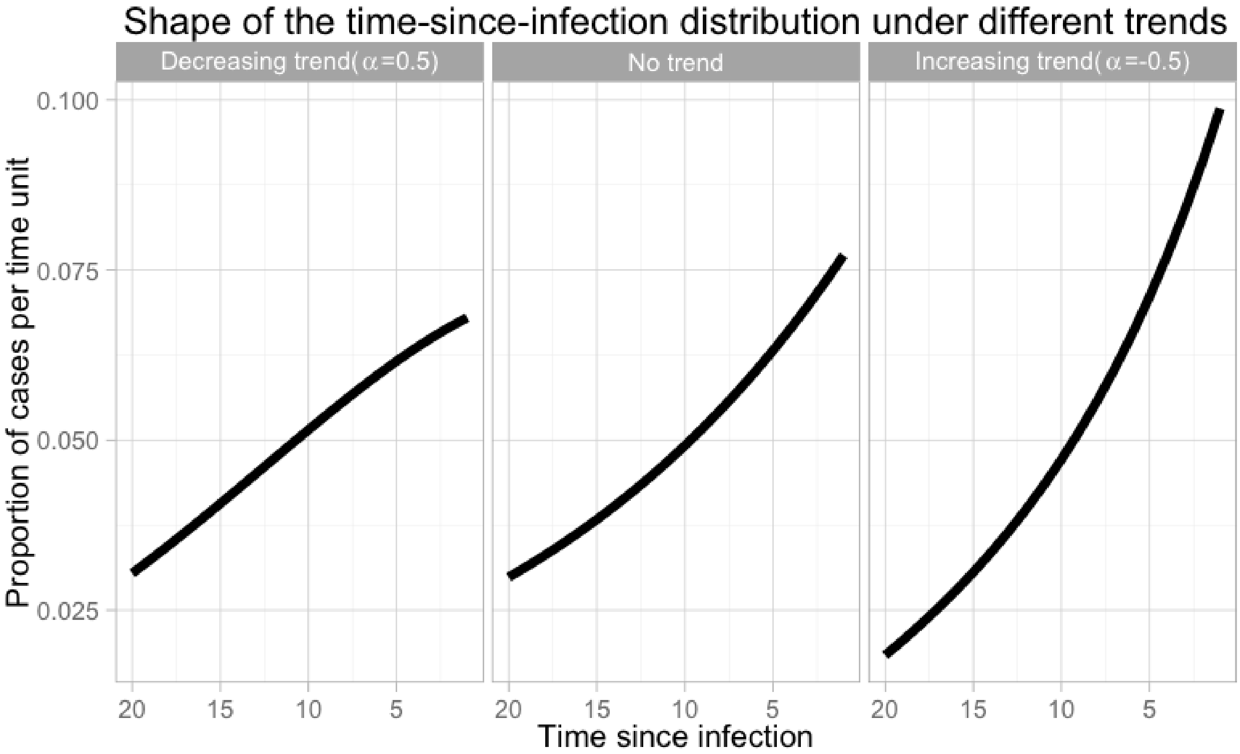

Figure 3.2: *Time since infection vs the value of $f_{tsi}(t)$, when assuming a linear trend of incidence that is either decreasing, constant, or increasing.*

more recent infections can hide infections that occurred further back in time. However, it will be demonstrated that it is still possible to estimate the incidence trend from the parameters of the exponential distribution.

The aim of this chapter is to recover the population-level trend of incidence from test measurements taken from individuals that have been infected at a some point in the past in a population where the time-since-infection (tsi) distribution is assumed to follow the equation under a linear trend defined above. In order to study the properties of this inference problem, a first approach is to investigate the simplified situation where the times of infection are known and generated using f_{tsi}.

Including the priors described in the previous section, the expression for the log posterior of f_{tsi} given observations of time since infections X (denoted by LP_f), becomes

$$LP_f(\alpha, \beta|X) = \log(U(\alpha| - 2 \times \bar{\lambda}/C, 2 \times \bar{\lambda}/C)) +$$

$$\log Beta(\bar{\lambda}|N + 1, N - N_+ + 1)) +$$

$$\sum_{\forall i} log[(\lambda + \beta \times X_i) \frac{e^{-(\lambda + \beta \times X_i/2)X_i}}{(1 - e^{-(\lambda + \beta \times C./2) \times C))}} \times I(X < C)]$$

Figure 3.3 shows the resulting log-likelihood surface of LP_f. Times of infections $X = \{x_i\}$ were simulated from the probability distribution of f_{tsi}, and the value of the log posterior LP_f given the simulated data was calculated over a grid of values for α and β. Note that this is assuming that the times of infection were exactly known. . There are two things to note in this image: the first one is that the region of highest likelihood is that surrounding the black line. This black line is the line for which the combination of α and β results in the same average incidence $\bar{\lambda}$, which indicates that the Beta prior on $\bar{\lambda}$ has a strong influence on the curvature of the log-likelihood surface. Another thing to note is that the estimated log-likelihood

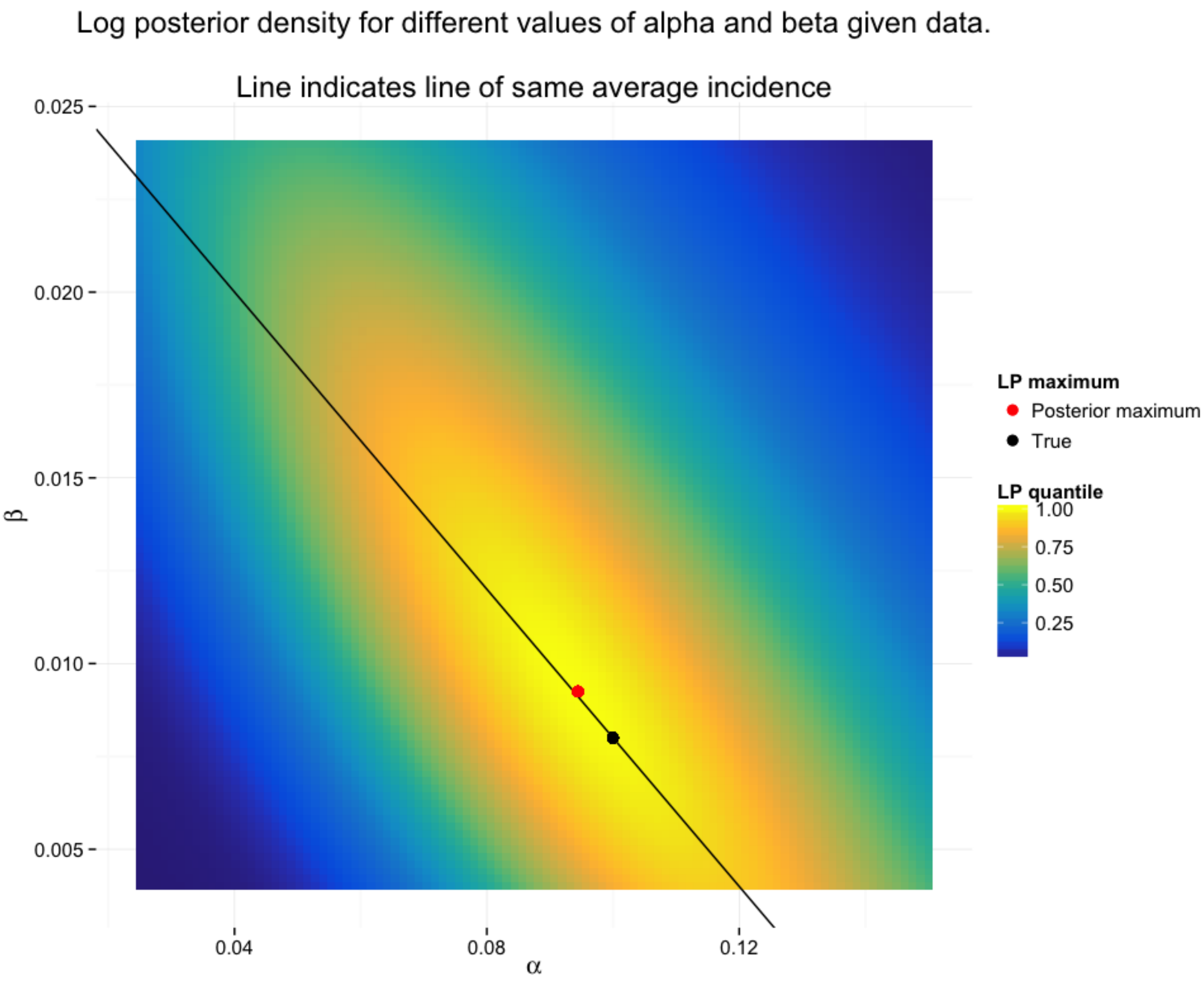

Figure 3.3: *Density surface of the log posterior distribution LP_f over α (x-axis) and β(y-axis), conditional on a collection of 1000 times since infection generated from the probability distribution defined by* $f_{tsi} = (\alpha + \beta T)e^{-(\alpha+\beta T/2)T}/(1 - e^{-(\alpha+\beta C/2)C})$. *Yellow indicates a density value in the highest quantile, blue a density value in the lowest quantile. Note that this assumes that times since infection are known exactly; the full inference procedure includes estimating these times since infection from test data.*

changes little along the line of equal mean incidence c, but drops off quickly with increasing perpendicular distance to the line. There is little distinguishing the region surrounding the true value (noted by the black dot), from the rest of the likelihood along the line of constant incidence. This indicates that while it will be relatively easy to recover the value of $\bar{\lambda}$, finding the correct combination of α and β is more challenging.

3.4 test kinetics

The hindcasting framework incorporates knowledge of the kinetics of diagnostic test responses after infection in the form of a latent process model $P(L|T, E, \theta_L)$. This latent process model describes (stochastic or deterministic) aspects of the dynamic process that develop following the introduction of a pathogen to a host. This process is complex, with a wide variety of factors. However, as pointed out by Pugliese and Gandolfi (2008), for many applications it is enough to describe this dynamic as a two-part interaction, with one variable representing the overall level of immune response, and a second variable representing the total pathogen burden. Formalising the interaction of these two variables over time by simple paired differential equations, it is possible to capture many of the qualitative patterns of interest for disease modelling, as will be demonstrated here, this approach proves useful as the basis for statistical inference.

It will be assumed that the kinetics of antibody test response after infection reflects an underlying development of the host immune response. Similarly, when looking at the kinetics of e.g. a quantitative PCR test, the development of the measured quantity of nucleic acid will be assumed to reflect the pathogen burden in the host. In other words, it is assumed that observed test measurements are proportional to host immune response and pathogen load, respectively. Under these assumptions, a paired differential equation approach can be used to generate realistic paired test kinetics reflecting underlying host-pathogen interactions.

For the purpose of evaluating the framework, example test kinetics were generated using a Lotka-Volterra predator prey type model. This type of model is usually defined in terms of the growth rates of two populations, a "predator" population, and a "prey" population, where the "predator" eats eats the "prey". For more details on the properties of such models, see e.g. Wangersky (1978). In our case, the pathogen fills the role of the "prey" which is being hunted by the immune response, our predator. Denoting pathogen levels at time t by $Na(t)$ and immune response levels at time t by $Ab(t)$ the model can formally be written as:

$$dNa/dt = (b_{Na} - h \times Ab(t)) \times Na(t)$$

$$dAb(t)/dt = b_{Ab} \times Na(t) - d_{Ab}Ab(t)$$

In these equations, b_{Na} can be interpreted as the growth rate of the pathogen in the host in the absence of an immune response, and h can be interpreted as the proportion of pathogens that dies per time unit for each unit level of immune response (i.e. the predation rate). b_{Ab} is the the unit level of increase in immune response generated per time unit for each single pathogen organism present in the host, and d_{Ab} indicates the rate of decline of the immune response per time unit in the absence of stimulation by presence of the pathogen. The equations can be solved for a given set of parameter values using any generic linear ordinary differential equation (ODE) solver to generate a bivariate function $LV(t) = (Ab(t), Na(t))$ describing the mean trajectory of antibodies and pathogen load over time (see figure 3.4). In practice, the values of $LV(t)$ were pre-calculated for a range of time points and stored in a lookup-table.For the purpose of evaluating the hindcasting framework, $LV(t)$ was assumed to be be known and used to define the latent process model.

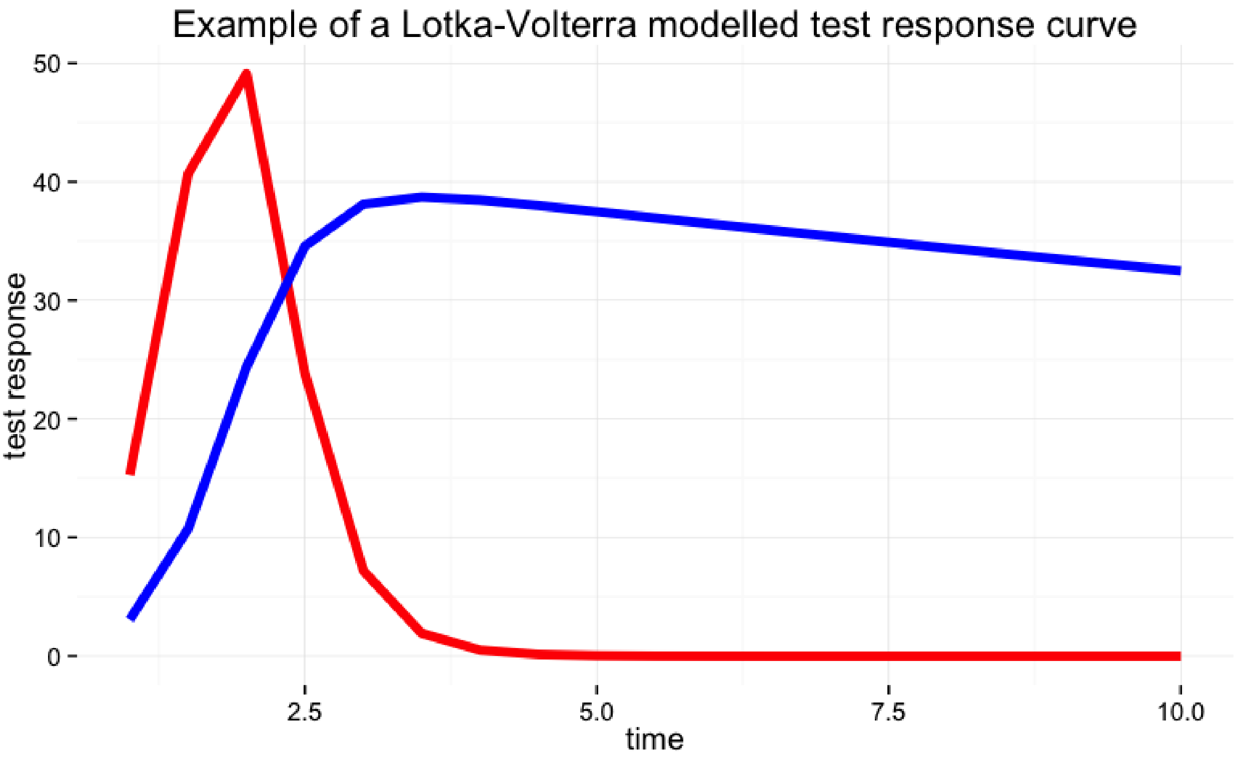

Figure 3.4: *Graph of a typical Lotka-Volterra curve. X-axis indicates time, y-axis indicates "population size"/antibody level/pathogen load. Red line indicates the trajectory for the prey (pathogen), blue indicates the trajectory of the predator (antibody level).*

Given a particular time since infection t, and using $LV(t)$ as the latent process model $LV(t)$, the observation model $P(Y|L(t), \theta_Y)$ can be defined by assuming that observed test measurements are then log-normally distributed around $LV(t)$. Specifically, it is assumed that test measurements come from a bivariate lognormal distribution around $LV(E)$, where E are the times of exposure, with zero correlation and independent standard deviation for the two tests:

$$\log N(Y|\mu = \log(LV(E)), \Sigma = \left(\begin{smallmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{smallmatrix}\right))$$

The exact combination of tests used could have an impact on the performance of the hindcasting procedure. By varying the parameters of the Lotka-Volterra equations, three different combinations of tests were generated, representing three disease types, or canonical patterns of host-pathogen interaction in terms of test responses as a function of time since exposure. Table 3.1 shows the parameter values used for each of these disease types; however it is the relative response times of the two tests which determine the disease type. Figure 3.5 shows development over time together with phase plots for the three different canonical patterns. The set of parameters labelled a type 1 disease correspond to an incubating pathogen and capture the type of interaction one would see for a pathogen which has an incubation period during which it is reaching full strength, followed by an immune response and a decline of pathogen levels, until the host is completely cleared. This is modelled by assuming a high growth rate for the pathogen, a moderately high number of antibodies generated per pathogen, antibodies being efficient at killing the pathogens, and a slow die off of antibodies. This results in an initial high pathogen growth, until the antibodies have caught up, bringing the pathogen load under control. Influenza is a classic example of a disease where incubation plays a major role in its epidemiology and pathogenesis (Carrat et al., 2008).

The set of parameters labelled a type 2 disease correspond to a fast-acting pathogen and assume that the growth phase of the pathogen has already been completed at t=1. This is modelled by having a high starting value for the nucleic acid values. It then models the dynamic with a slow pathogen growth, the antibody kill rate being ten times the pathogen growth, and a low growth and die-off rate for the antibodies, so that it takes some time for the pathogen load to be brought down to zero. This pattern models rapid-acting diseases such as norovirus infection (Lessler et al., 2009).

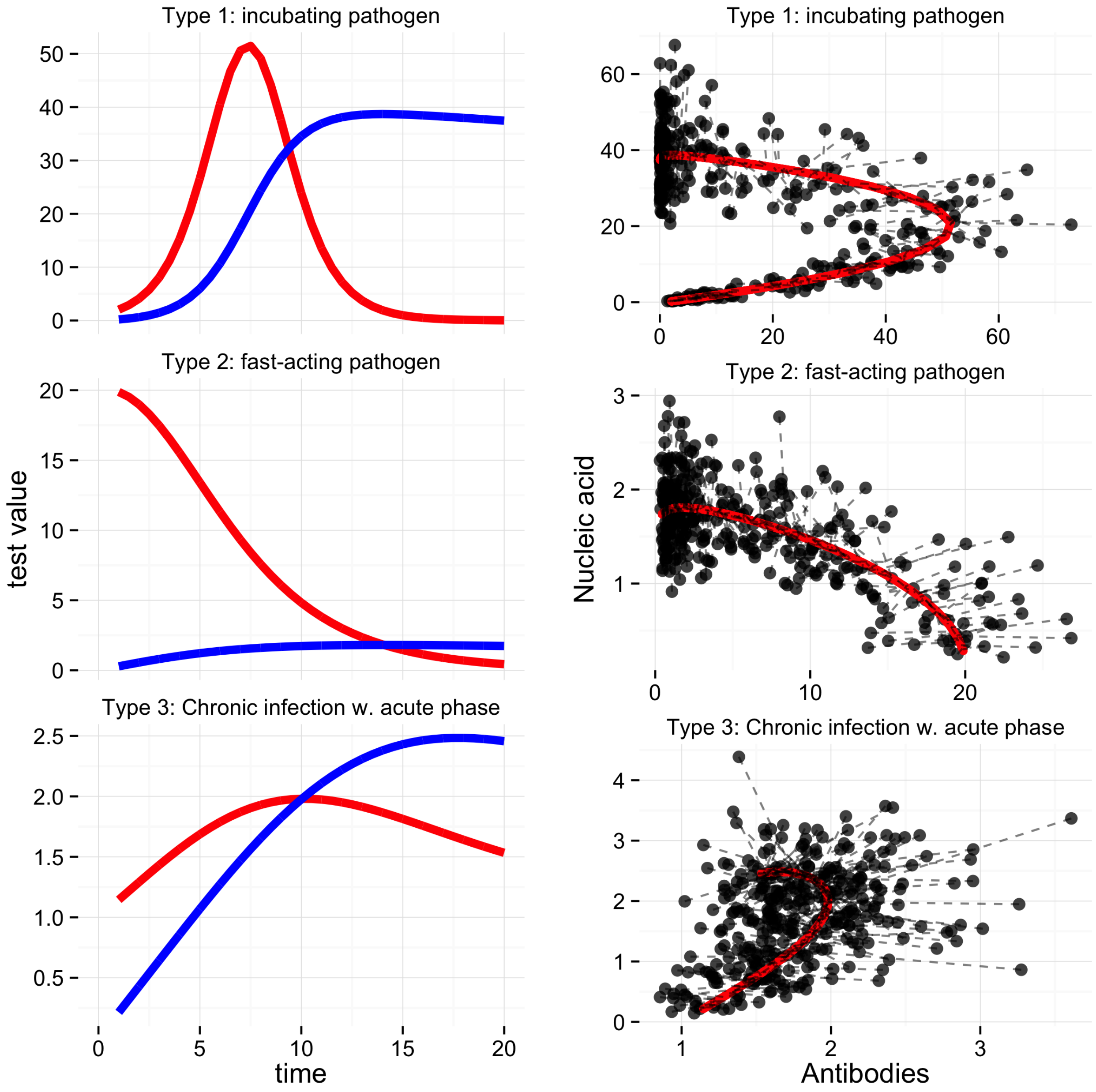

Figure 3.5: *Graphs of the three types of LV dynamics used, development over time and phase plots. Left hand side shows kinetic curve over time (red indicates pathogen load, blue indicates antibody response). Right hand side shows phase plots in terms of locations over time in the space of Antibody response (x-axis) versus pathogen load (y-axis). The graphs on the right hand side also display 1000 example test results, generated by assuming 25% lognormal noise around the mean curves. Dashed line indicates where on the the kinetic curve each point originate from.*

The set of parameters labelled a type 3 disease, models a "chronic infection with acute phase", and assumes that the growth rate of the pathogen is equal to the die-off rate of the antibody, an antibody growth of ~1 antibody/pathogen and time unit, but that each antibody is relatively ineffective in killing pathogens. In this way, the pathogen load and the antibody levels reach a slowly declining equilibrium after an initial growth phase, resulting in high levels of both antibodies and pathogen load remaining for some time after the pathogen has peaked. Scrapie and Tuberculosis are two diseases that follows this pattern.

Table 3.1: *Table of parameter values for different types of LV curves.*

Pathogen type	Prey growth	Pred kill %	Pred. growth	prey death	Starting state	Prey peak time
Type 1	3	0.15	0.6	0.3	na=1,ab=0	2
Type 2	0.06	0.6	0.06	0.06	na=20,ab=0	1
Type 3	0.6	0.3	0.9	0.6	na=1,ab=0	2.5

The time scale over which these kinetics develop varies depending on the type of pathogen modelled. As mentioned earlier, scrapie (which would be a type 3 disease) is a a very slow-developing disease, where the expected development of the pathogen level would be on the order of several years (United States Department of Agriculture, 2010). Similarly, Paratuberculosis (also type 3) can take up to a year of in-host replication before it starts showing symptoms (OIE, 2014). On the other hand, bluetongue virus (a type 1 disease, with a 10 day incubation period) in sheep has a time period of 2-7 weeks from infection, through incubation and showing of symptoms, to potential clearing of the disease (Sperlova and Zendulkova, 2011). In humans, Chlamydia (also resembling a type 1 disease) takes around a month to show symptoms, but the bacteria can then remain persistent at a low level for months or years after that, causing damage to the patient (Hogan et al., 2004). Norovirus infection (a type 2 disease) has a very quick course of action with only about 4 days from infection to clearance (Patel et al., 2009).

In general, we can only expect to estimate times since infection accurately while the infection process is ongoing, i.e up to the time it takes for the infection dynamic to go from exposure to clearing the infection and removing all antibodies. Fortunately, antibodies often remain for extended periods of time; however, once the pathogen is cleared, the benefits of multiple tests are lost. If the dynamic reaches a more-or-less steady state with only minor changes after $T = 5$ (whatever the unit of T is), then we can only reasonably expect to distinguish times since infection up until time 5, and by extension only expect to hindcast population-level dynamics of the disease that occurred within that time frame. The important characteristic of multiple diagnostic tests for the purpose of hindcasting is that the different tests have different dynamics over time. Looking at the test kinetic phase plots in Figure 3.5 (right column), we can see that the bivariate trajectories give more information than if we were to project them onto either axis (i.e. take only a single measure of infection). With differing test kinetics it is, in principle, possible to combine the test results into as close-to-unique signatures of times since infection as possible. The greater the difference between test trajectories, the more precision is gained from combining them. Even with two tests with identical kinetic trajectories, the combined measurement will reduce measurement error and increase precision.

However, with increasing separation of timescales, the ability to distinguish early from late infections becomes both more robust, and more sustainable under higher levels of measurement error.
From these observations, and looking at the three categories of paired kinetics displayed above, we would expect combined test data from a type 1 disease to give precise levels of estimate under moderately high levels of noise up until $T = 5$, since both test kinetics are changing and providing information up that point. A type 2 disease reaches a slow-changing state after $T = 7$ for both antibody levels and pathogen load, and would thus have difficulty providing information beyond this time horizon. The final type 3 disease exhibits a well-defined separation between the two curves and a strong interaction occurring up until $T = 15$, indicating that such an infection may provide useful information at least up until that time.

3.5 IMPLEMENTATION

3.5.1 *Describing endemic trends using the hindcasting framework*

The test kinetics and the distribution of times since infection are combined to make use of the hindcasting framework. Recall that in general, the posterior probability $P(E, \theta | Y, T)$ of a set of exposure times E and model parameters θ given observed data Y and observation times T, can be written as a combination of a deterministic process $L(T, E)$ of expected test results at a given time point , an observation process $P(Y | L(T, E), \theta_Y)$, and a distribution of exposure times $P(E | \theta_E)$, forming the expression

$$P(E, \theta | Y, T) = P(Y | L(T, E), \theta_Y) P(E | \theta_E) P(\theta) / (P(Y, T))$$

$$\propto P(Y | L(T, E), \theta_Y) P(E | \theta_E) P(\theta)$$

In the linear trend scenario, the distribution of exposure times is defined as $P(E|\theta_E) = f_{tsi}(E|\alpha, \beta) = (\alpha + \beta E)e^{-(\alpha+\beta E/2)E}/(1 - e^{-(\alpha+\beta C/2)C})$, with the priors for α and β defined in section 3.4. The deterministic function $L(T, E)$ describing expected test levels was set as the solution $LV(t)$ to the Lotka-Volterra equations defined in section 3.5, with values over time included in the JAGS model as a lookup table.

Finally, the observation process $P(Y|L(T, E), \theta_Y)$ giving the distribution of observed test data was set as a bivariate lognormal distribution around $LV(t)$ with zero correlation and independent standard deviation for the two tests:

$$P(Y|L(T, E), \theta_Y) = \log N(Y|\mu = \log(LV(E)), \Sigma = \left(\begin{smallmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{smallmatrix}\right))$$

Defining suitable priors for the lognormal distribution proved challenging - in the traditional parametrisation of the lognormal distribution, standard deviation σ is defined on the log scale. However, in terms of interpretability, it is easier to work on the observed scale, using $(\sigma_1^\star, \sigma_2^\star) = \exp((\sigma_1, \sigma_2))$. In this way, $\sigma^\star > 1$, and interpretable as the multiplicative variation around the mean; e.g. $\sigma^\star = 1.5$ implies a relative standard deviation of 1.5, and ~68% of the mass of the distribution fall within $\text{mean}/1.5, \text{mean} * 1.5$. However, when putting a prior on $\sigma^\star$, and generating posterior samples, the MCMC chains proved to be mixing very slowly. This was likely due to the sampler having a fixed step length despite the fact that a step change of size δ when $\sigma^\star$ is close to one has a much larger effect on the likelihood than the same δ step when $\sigma^\star$ is large. To counter this slow mixing, the prior was instead set for $\sigma^{\star\star} = e^{\sigma^\star - 1}$ and given an exponential distribution $P(\sigma^{\star\star}) = \lambda e^{-\lambda(\sigma^{\star\star}-1)}$, shifted to account for $\sigma^{\star\star}$ always being above 1 (since $\sigma^\star - 1 > 0$). The full expression of the prior is thus $P(sigma^{\star\star} = x) = \lambda e^{-\lambda(x-1)}$. After experimentation, λ was chosen to be 20, which gave a posterior where the majority of the mass was under $\sigma^{\star} < 1.5$, and only an infinitesimal mass above $\sigma^{\star} > 2.0$.

Given the priors and the model for times since exposure, the full expression for the posterior distribution of $P(E, \theta|Y, T)$ thus becomes

$$
\begin{aligned}
&P(E,\theta|Y,T) = P(Y|L(T,E),\theta_Y)P(E|\theta_E)P(\theta) \propto \\
&\prod_{\forall i}[\log N(Y_i|\mu = \log(LV(E_i)), \Sigma = (\sigma_1^{\star}, \sigma_2^{\star}))]\times \\
&\prod_{\forall i}[(\lambda + \beta \times E_i)\frac{e^{-(\lambda+\beta\times E_i/2)E_i}}{(1 - e^{-(\lambda+\beta\times C/2)\times C))}} \times I(E_i < C)]\times \\
&\lambda e^{-\lambda(e^{\sigma_1^{\star}-1}-1)} \times \lambda e^{-\lambda(e^{\sigma_2^{\star}-1}-1)}\times \\
&U(\alpha| - 2 \times \bar{\lambda}/C, 2 \times \bar{\lambda}/C)\times \\
&\beta(\bar{\lambda}|N+1, N - N_{+} + 1))
\end{aligned}
$$

3.5.2 *Simulated data for framework evaluation*

In order to determine the feasibility of the approach laid out in the previous section, the ability of the hindcasting framework to detect the direction of change in the incidence of infections in a population was evaluated. The impact and interactions of a number of different factors on the performance of the hindcasting procedure were investigated by generating data under a sequence of scenarios that is described below. For all the scenarios used to evaluate the performance of the hindcasting framework, data were generated by sampling the specified number of times since infection from the distribution f_{tsi} of times since infections under a linear trend. Given these sampled times of infection, sample data was generated using a lognormal observation error around the expected mean test values defined by specified test kinetics associated with the scenario. The multiplicative standard deviation for the lognormal observation term $P(Y|L(T,E),\theta_Y)$ was set to a relative variation of 1.25 around the mean kinetic curve. The overall population size that was sampled was chosen so that the expected number of positive samples would be constant across the scenarios which have different levels of incidence. By fixing the number of positive samples, rather than sample sizes of positive and negative individuals, the effect of differing trends and levels of incidence could be evaluated without being confounded with sample size effects.

The initial exploration of decreasing, constant, and increasing trends, was conducted using a single data set for each type of trend. Subsequent scenarios were evaluated repeatedly, each time with independently generated identically distributed data sets. The total number of different scenarios was 193, with each scenario being run 11 times. The JAGS code for evaluating these scenarios was run on Amazon's cloud computing service EC2 (Juve et al., 2009). Running all 2213 scenarios took 72 hours using a 32 core computation-optimised linux instance.

Scenarios with increasing, constant, and decreasing trends

In a first step, the framework was applied to three data sets generated under different scenarios: a scenario with decreasing incidence trend (50% decrease in the level of incidence over the time span), a scenario with constant incidence trend, and a scenario with increasing incidence trend (50% increase in the level of incidence over the timespan). For all of these initial scenarios, an incubating disease (type 1) was assumed, a sample size of 5000, an incidence at the time of sampling of 0.05 per individual and time unit, and a time span of 20 time units. For all three trends, it was assumed that two diagnostic tests were used.

Scenarios with different test kinetics and trends of different magnitude

In the second set of scenarios, the effect of trend and sample size on the ability of the hindcasting framework to recover trends was considered. Data sets were generated with different trends (50% increasing or decreasing, 25% increasing or decreasing, or constant), different assumed sample sizes (100, 250, 500, 1000 and 2000 samples), and different test kinetics (using disease types 1, 2, and 3 as defined in section 3.4). For all scenarios, it was assumed that the incidence at the time of sampling was 0.1 per individual and time unit, and that two diagnostic tests were used.

Scenarios comparing different test kinetics and number of tests

In the third set of scenarios, the trend was fixed to a 50% increasing trend, with an incidence of 0.1 per individual and time unit. Instead, it was assumed that either two diagnostic tests, only an antibody based test, or only a Nucleic acid (or antigen) based test were used for testing collected samples. Sample sizes again varied between 100, 250, 500, 1000 or 2000 samples.

Scenarios modelled after real-world settings

Finally, the performance of the framework was also evaluated for three case studies based on plausible parameter values, to produce a first indication of the real-world usefulness of the hindcasting framework. These three examples were modelled after scrapie in sheep, Chlamydia in humans, and squirrelpox in squirrels. Scrapie in the US has been subject to an intense control effort since around 2002. Between 2003 and 2009, the estimated prevalence decreased from 0.2% to 0.05%, a reduction of 75%,equivalent to a 12.5% reduction per year over 6 years (United States Department of Agriculture, 2010, p7). In terms of the Lotka-Volterra dynamics discussed earlier, scrapie would correspond to a slow-growing pathogen with a longterm chronic infection, the third of the three generic disease types. The development of scrapie in the US was thus modelled using a starting incidence of 0.2%, a 12.5% reduction per year over 6 years, using disease type 3 acting on a timescale where peak pathogen burden is reached after 3 years, the typical time of onset of clinical signs for scrapie (United States Department of Agriculture (2010), p9).

Chlamydia incidence in Sweden decreased between 1990 and 1995, after which the trend reversed and increased until 2007. In 2006, a mutated strain of *Chlamydia trachomatis* that was not detected by the standard tests started spreading in some Swedish counties (Herrmann et al., 2008). By the time this strain had been discovered and tests adjusted, a number of cases that would otherwise have been detected continued to carry the infection and infect others. By 2007, the number of cases reported yearly had increased from ~32500 in 2004 to 47500, a 50% increase. Based on these events, a scenario was modelled by assuming a 0.02% incidence (the approximate incidence among 20-24 year olds, the highest-risk group). Chlamydia was considered a type 1 disease where the disease dynamic plays out over approximately 2 years (to model both the potential for persistent infections and the duration of antibodies), and a trend of 50% increase over these two years.

Squirrelpox is a viral disease that plays a major role in the decline of European red squirrel (*Sciurus vulgaris*) populations in the UK. The eastern grey squirrel (*Sciurus carolinensis*), which are asymptomatic carriers of squirrelpox, was imported to the UK in the late 1800s (Stritch et al., 2015). Grey squirrels have since spread the disease to red squirrels, for which squirrelpox is a deadly infection. Squirrelpox infections have recently been reported in Ireland, and a study there indicated a 34% seroprevalence; though it is unknown how long the virus has been circulating. This setting was modelled by assuming a 30% incidence at the time of sampling, a fast-acting type 2 disease where the dynamic process lasts over three years (the life expectancy of red squirrels). Since the trend of squirrelpox is unknown, a range of different slopes were examined.

The scrapie scenario is used to investigate the number of cases required to use the hindcasting framework for evaluating the effect of control measures in a low-incidence, livestock context. The Chlamydia scenario is used to investigate the number of cases required to recover increasing trends in a medium-incidence human context. The squirrelpox scenario investigates the precision with which one can estimate differing trends in a high-incidence, wildlife context where the number of available samples are limited. For these three case studies, ten different datasets was generated for each set sample size, and the hindcasting framework was applied to each data set.

3.5.3 *Sampling from the posterior using JAGS*

For conducting inference of the endemic trend, the posterior distribution of parameters described in the previous section is evaluated, conditional on observed test data and knowledge of expected test kinetics.

A high level language for hierarchical Bayesian models known as JAGS (Plummer, 2003) was used to implement the statistical framework and evaluate the posterior distribution using the Metropolis-Hastings algorithm combined with Gibbs sampling (see section 1.5.2 for a more detailed discussion). The code was called from within R using the *rjags* package (Plummer, 2014). Initial starting values for the parameters in the MCMC were chosen at random from the prior distributions, independently sampled for each chain. Samples were then taken from the joint posterior distribution of times since infection across all individuals, and parameters of the trend of incidence. As noted in the introduction (section 1.5.4) a key question with the implementation of MCMC algorithms is that of convergence and mixing. The reliability of our sampling tools were assessed using

Gelman-Rubin (GR) statistics for the posteriors as well as visual inspection of trace plots. Figure 3.6 shows typical trace plots from three scenarios (decreasing, constant and increasing trend) the last 1000 draws for each chain (thinned so that every 10th draw is shown) of the population level trend parameter β, the intercept parameter α, and the mean incidence $\hat{\lambda}$, from five different chains after all chains have been run for a 1500 iteration burn-in.
For the full range of scenarios, it was not feasible to inspect trace plots. Instead, a Gelman-Rubin statistic above 1.15 was used to filter out those runs that had not converged (~5% of the total runs). The results from these runs were excluded from subsequent analysis.

3.6 RESULTS

3.6.1 *Scenarios with increasing, constant, and decreasing trends*

For the first set of scenarios, the posterior distribution was evaluated using the JAGS implementation of the framework as described earlier, and the results studied in-depth. Figure 3.7 shows the posterior distributions for the incidence parameter α, the slope parameter β, and the mean incidence $\hat{\lambda}$ using data from the three scenarios.
Figure 3.7 clearly indicates that the posterior distributions of the trend parameter differ significantly between the increasing, the constant and the increasing trend scenarios, and correctly identifies the direction of change, with the 95 % credible intervals excluding the zero for the increasing and decreasing scenario. In each scenario the posteriors associated with the five realisations (denoted by the five colours) overlap, indicating that the MCMC chain has converged. It is however interesting to note that the estimates are all biased compared to the true value, although the true value lies within the 95 % credible interval.

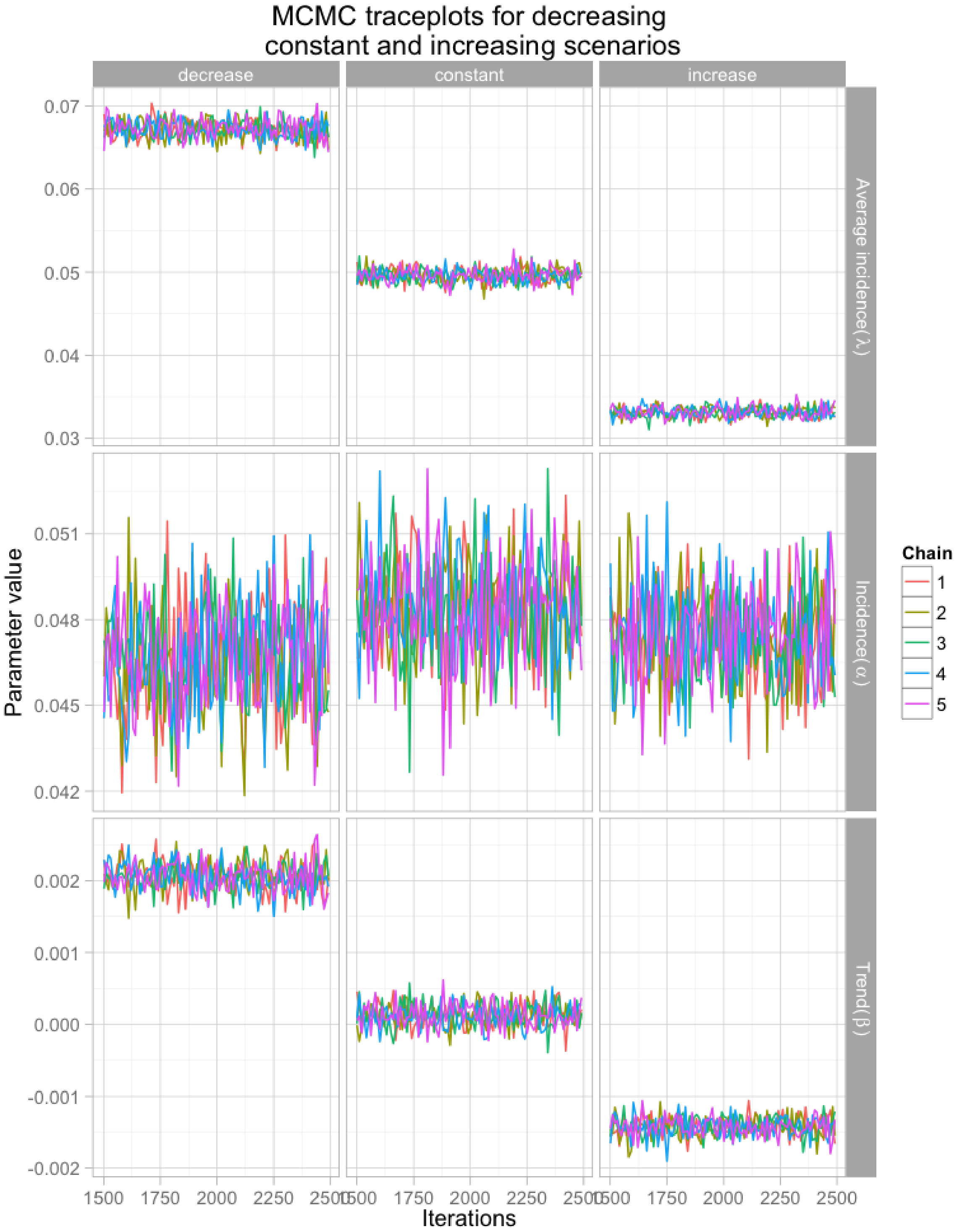

Figure 3.6: *Traceplots of the the MCMC samples of the α(incidence at the time of sampling) and β(trend) parameters for three different scenarios, with five chains each. Colours indicate the respective chains.*

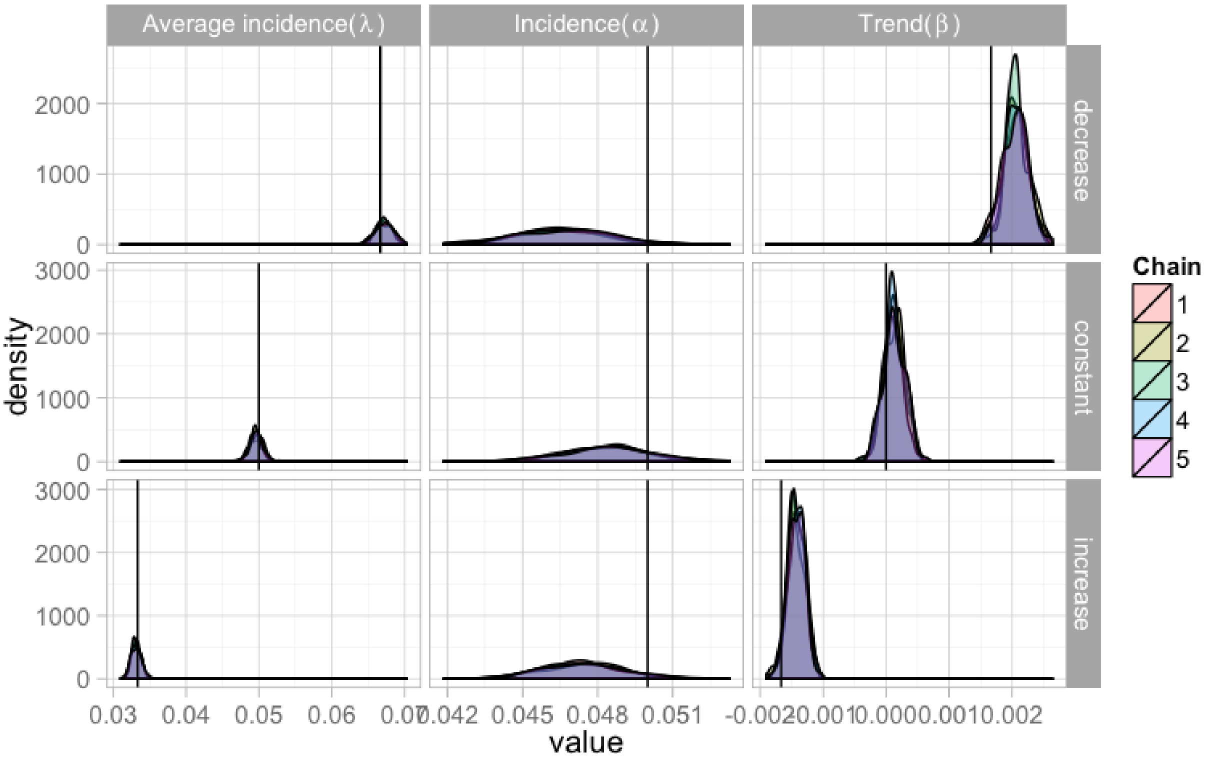

Figure 3.7: *Density plots for the MCMC samples of α and β from the posterior, fitted to data generated assuming populations with decreasing(top row),constant (middle row), and increasing (bottom row), and of incidence. Each colour for a density distribution indicates the results from one of the five chains that was generated from each MCMC run. The vertical line indicate the true parameter values.*

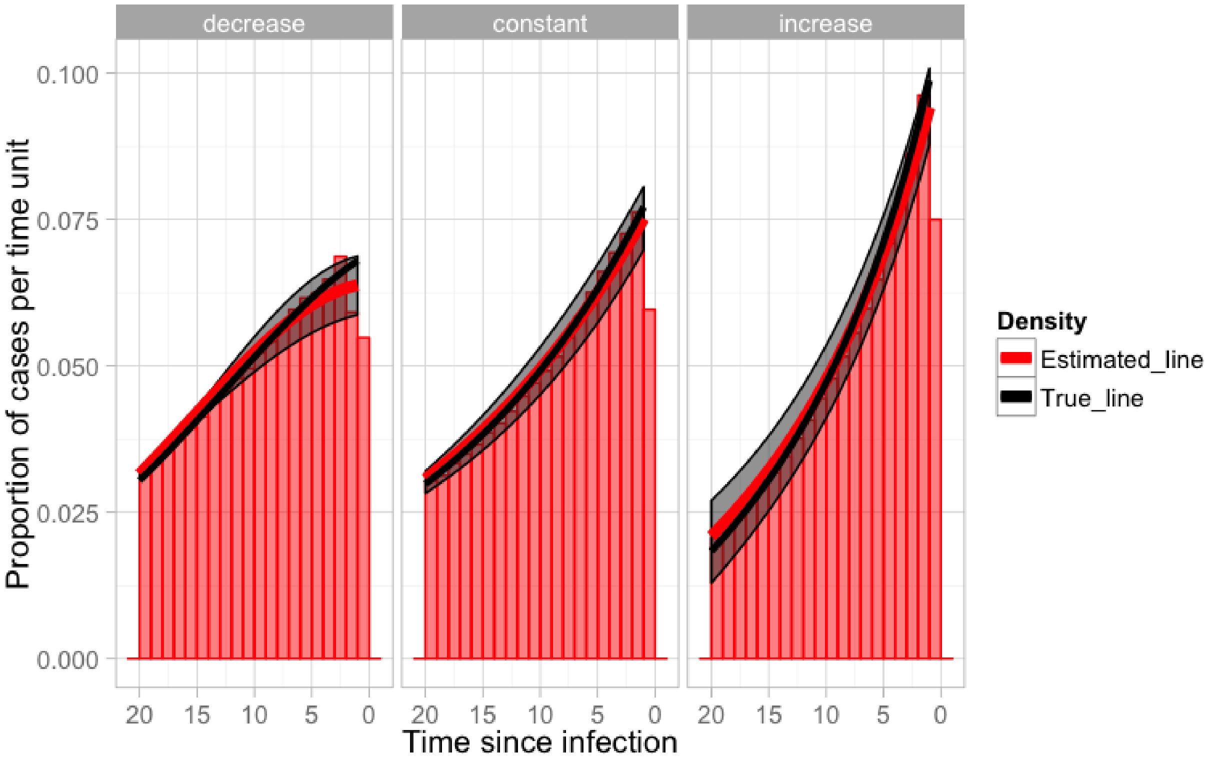

Figure 3.8: *Histogram of estimated times since infection, together with the trend line based on estimated parameter values of $f_{tsi}(t)$ (red) and the trend line based on true parameter values for $f_{tsi}(t)$(black). Bars indicate the proportion of mean posterior estimates of infection times for individuals falling into that time point. Grey shading indicates the 95% posterior credible interval for the trend.*

The distribution of estimated infection times across the population of samples is plotted in Figure 3.8. This figure displays the histogram of the mean posterior estimates of the infection times, overlaid with a red line indicating the probability distribution of times-since-infection (TSI) defined by the mean posterior estimates of the trend and intercept parameters and the definition of f_{tsi} from section 3.3. The black line indicates the TSI distribution defined by the true parameters and that was used to generate the data. The true trend curve and the estimated trend curves follow each other reasonably well, and the histogram of estimated times since infections also seem to follow the true trend. However, there seem to be small bias in very recent infections - the final bar of the histogram is lower than preceding ones, and the trend (in particular the decreasing trend) seem to be dragged down by this. This seems to indicate that the source of the bias is a minor tendency for the hindcasting framework to estimate very recent infections as having been infected for a longer time period.

As mentioned in section 3.3.3, because of the way reinfections hide preceding infections, the trend lines are roughly exponentially increasing for all three scenarios. The good fit of the estimated trend can be better seen by viewing them in a traditional manner as indicating the incidence of cases over time (see figure 3.9). This graph shows the linear trend lines defined by the estimated parameters, together with histograms of the number of cases that were infected in each time period. Note that the uncertainty of the estimated incidence increase as you go back in time.

3.6.2 *Scenarios with different test kinetics and trends of different magnitude*

Following evaluation of the individual runs, performance was further evaluated by generating data with different slopes of the inci-

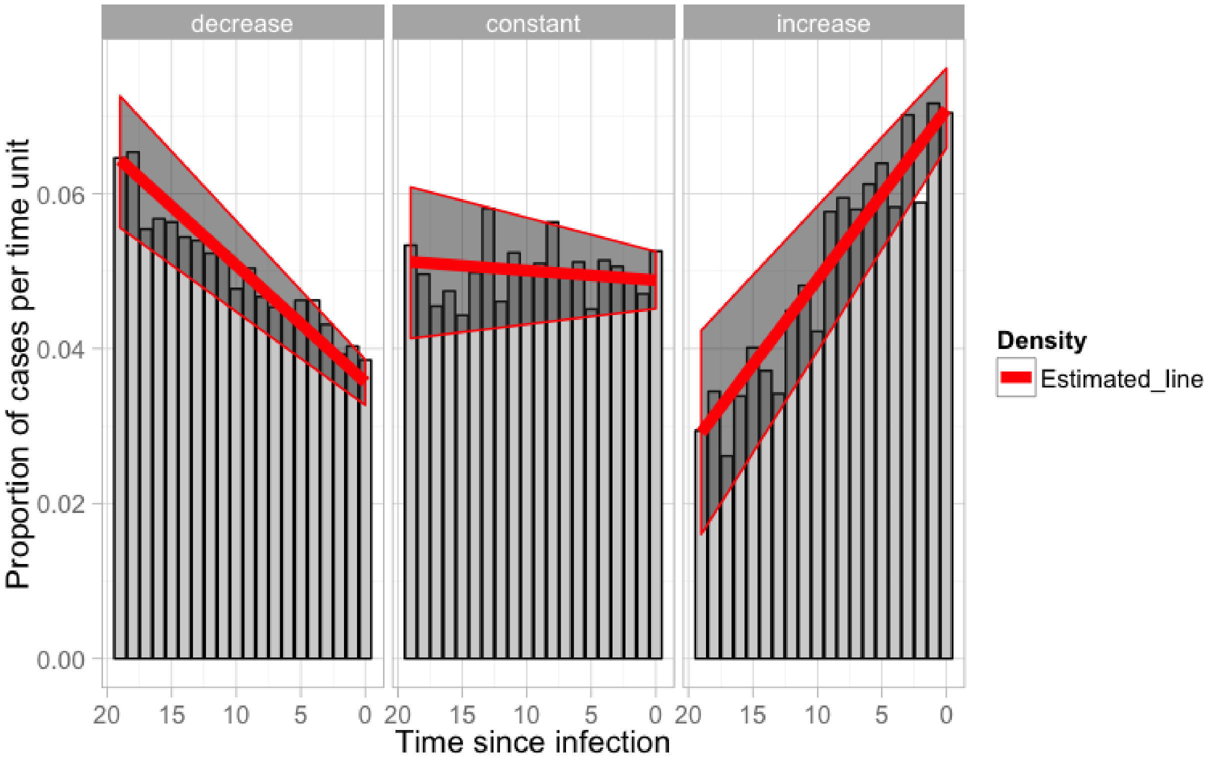

Figure 3.9: *Histograms of actual times of infection generated under the different scenarios, (counting multiple infections per individual separately) together with estimated trend line. Grey shading indicate 95% posterior credible intervals of the linear trend*

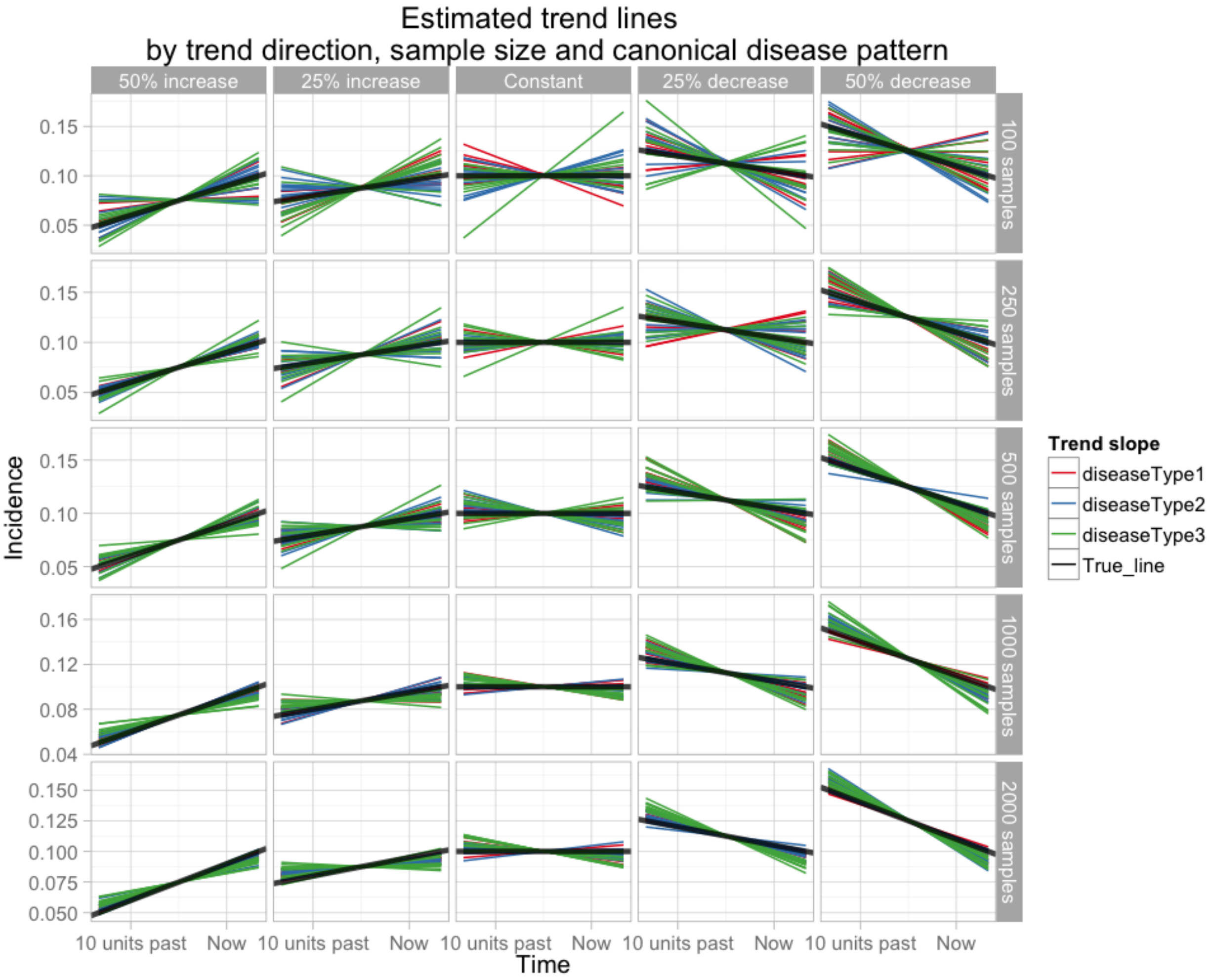

Figure 3.10: *Estimated trend lines (thin lines) for different disease types (indicated by colour), using different combinations of number of sampled positive individuals and strength of slope. The thick black lines indicate the true trend line*

dence trend, with different numbers of collected positive samples, and different disease type. Figure 3.10 shows the mean posterior estimated trends obtained by applying the hindcasting framework to generated data sets for each combination of slope, sample size and disease types. These plots indicate that the mean of the posterior distribution of the trend parameter usually correspond well with the true trend lines (indicated in black), if the sample size was 500 positive individuals or more. For trends of +/-25% change over the time period and samples sizes of 250 or less, the estimated trend was no longer reliable, and could even have the wrong direction of slope. There was no obvious relationship between disease type and performance.

It is also of interest to quantify the level of uncertainty in the estimated trend. Figure 3.11 show posterior credible intervals of the trend parameter in the various scenarios. Here it can be clearly seen that with a sample size of 2000 positive individuals, all posterior credible intervals exclude zero for the scenarios with 50% increase, 25% decrease, and 50% decrease over the time period of interest. With 1000 samples, the posterior credible intervals exclude zero for the +/-50% change scenarios. With 500 samples, the maximum posterior estimates are still close to the true values for most scenarios, but the credible intervals start overlapping zero. At 250 samples or less, the posterior estimates of all scenarios become unreliable. Again, no obvious pattern between different disease types can be seen.

Figure 3.12 shows the reliability of the posterior estimates, summarised in terms of the average proportion of the mass of the posterior distribution that lies on the correct side of zero. This quantity can be interpreted as the posterior probability for the correct sign of the slope. . Note that these proportions are somewhat unreliable, as we only have 11 replicates for each point. Nevertheless, some in-

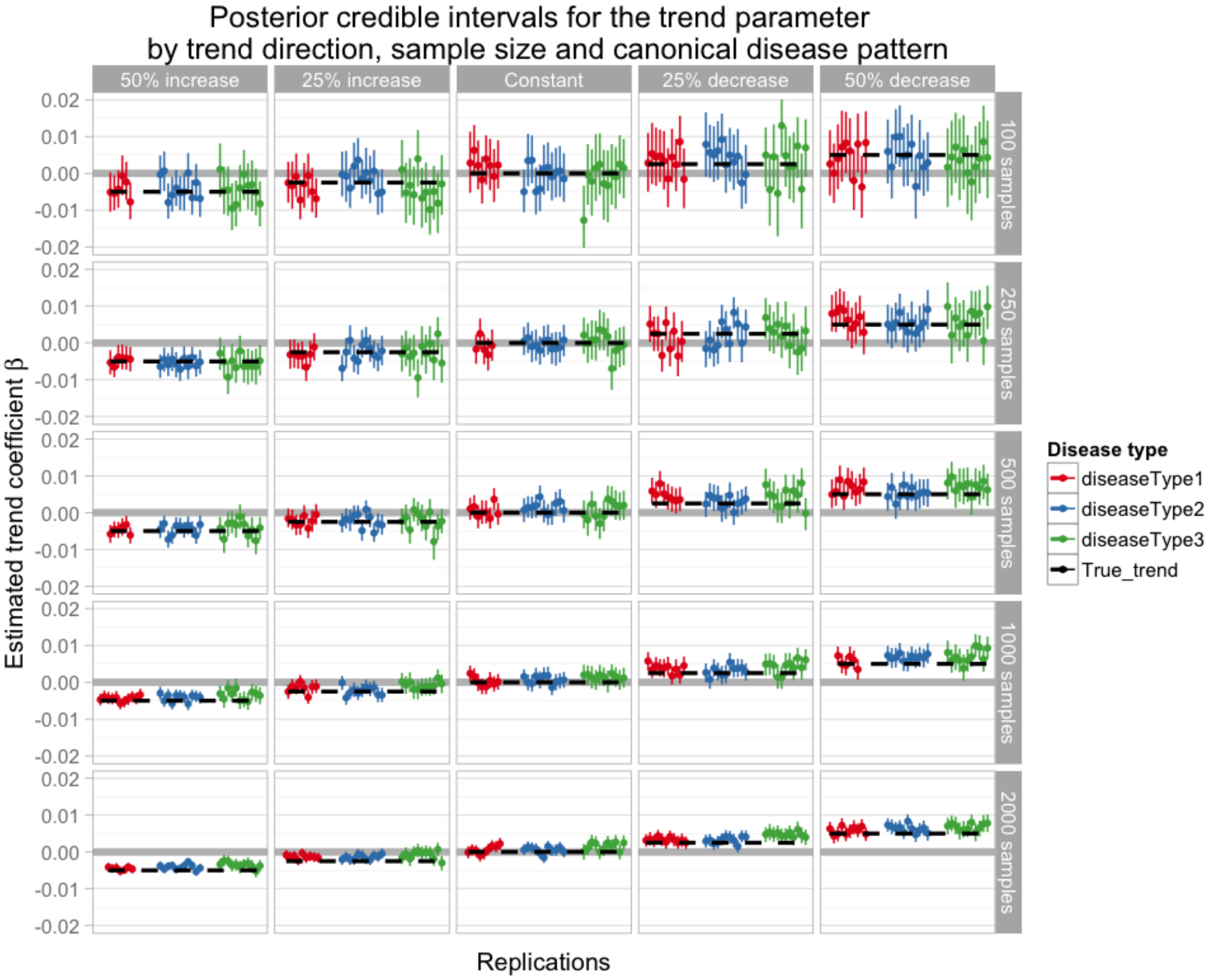

Figure 3.11: *Posterior Credible intervals for the slope parameter of the trend from scenarios with different sample size and slope relative to an incidence at the time of sample of 0.1 per time unit and individual. Thick grey line indicates zero slope, dashed black line indicate true slope. Dots indicate mean posterior estimate of the slope; lines indicate 95% posterior credible interval. Color indicate different disease type*

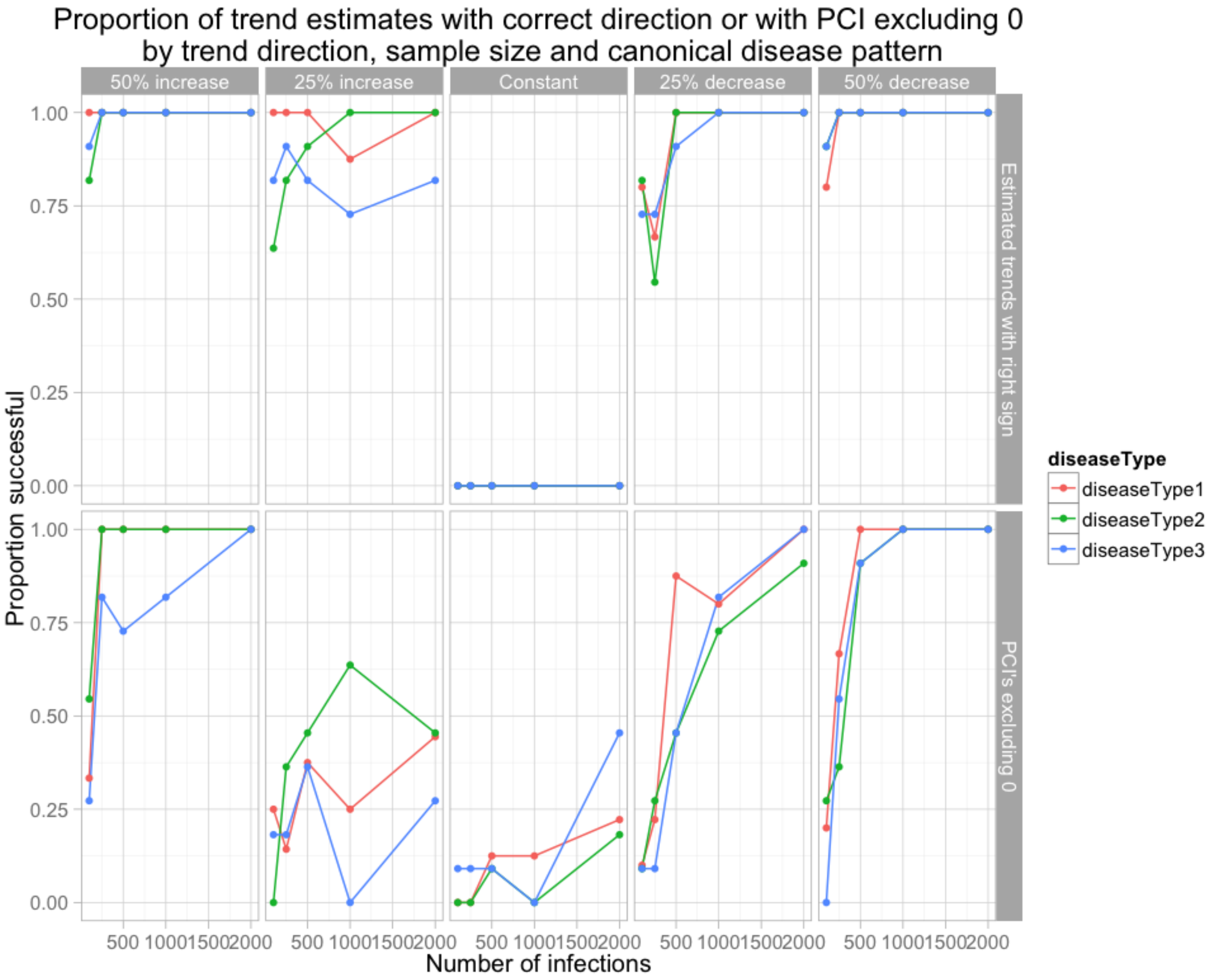

Figure 3.12: *Average posterior probability for the true sign of the trend parameter . The posterior probability is defined as the proportion of the 95% posterior credible interval that indicate correct sign for the trend parameter. The dots indicate the posterior probability averaged over the 11 replicates for each combination. Colour indicate different disease types.*

teresting patterns can be seen. A proportion of 75% of runs excluding zero seems to be reached at 1000 samples for the 50% increase, 25% and 50% decrease scenarios; for the +/-50% change scenarios, 500 samples seem to have been enough to clearly distinguish the trend from a constant one. Also, as long as the number of samples are 250 or above, the hindcasting procedure can distinguish between an increasing and a decreasing trend, as indicated by the upper row, though it seems more difficult to correctly identify an increasing trend than a decreasing one. As in the earlier graphs, no clear patterns between different disease types could be identified.

3.6.3 *Scenarios comparing different test kinetics and number of tests*

The effect of combining diagnostic tests for hindcasting, compared to using only single tests, was evaluated in scenarios assuming a 50% increasing trend (see figure 3.13). Whereas the combination of tests performed well for sample sizes above 250, the single-test scenarios showed some peculiarities. Scenarios with a single antibody test and disease type 3 produced exceptionally bad estimates of the trend no matter the sample size. On the other hand, scenarios with either disease type 1 or 2 fell down when using only a nucleic-acid based test. Equally surprising is that for disease type 1 and 2, the performance was almost identical with a single antibody test, as with a combined test; similarly with disease type 1 and nucleic acid test. This indicates that the interaction of types of tests used, the trend of disease in the population, and the type of pathogen considered is not completely straight-forward and warrant further research.

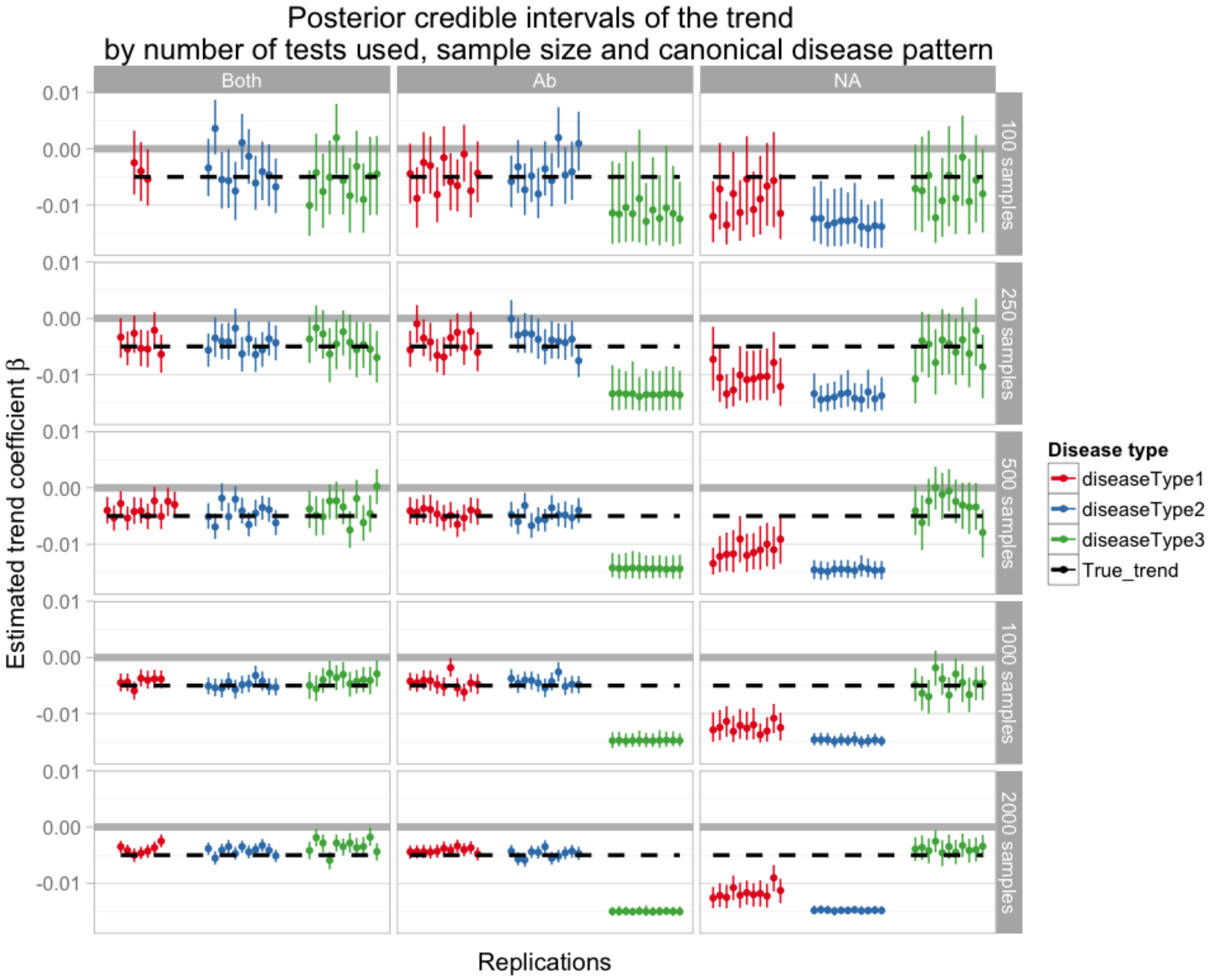

Figure 3.13: *Posterior Credible intervals for the slope parameter of the trend from scenarios with different sample size and number of tests used. Thick grey line indicates zero slope, dashed black line indicate true slope. Dots indicate mean posterior estimate of the slope; lines indicate 95% posterior credible interval. Colour indicate type of disease. Column indicate whether only antibody tests was used, only nucleic acid based tests, or both types of tests combined*

3.6.4 *Scenarios modelled after real-world settings*

The resulting estimates of the three different cases-studies using real-world plausible parameters can be seen in figure 3.14.
The scrapie-based scenario indicates that for these parameters, it would be possible to prove a disease reduction with as few as 100 samples, seeing as all but one of the posterior credible intervals exclude zero. With 1000 samples, the posterior credible interval indicates the correct slope to within a factor of 2.
For the chlamydia-based scenario, 500 cross-sectional samples would be needed to prove that the trend is increasing; with 2000 cross-sectional samples, the correct slope could be estimated to within a factor of 2.
Finally, with the squirrelpox scenario, it was assumed that only a limited number of animals would be able to be caught. Assuming that 250 animals were caught, the true trend would have to be +/-50% change in incidence over the three years, in order to conclusively demonstrate an increasing or decreasing trend.

3.7 DISCUSSION

This chapter has introduced and tested a novel technique for hindcasting the history of exposure to disease in a population using only cross-sectional data combined with information on pathogen test kinetics. The results demonstrate that this procedure enables estimation of changes in disease incidence over time. The results also demonstrated how this approach is able to distinguish between an increasing trend and a stable, or decreasing trend, as well as produce posterior estimates quantifying this disease trend. This goes beyond previous sero-incidence studies which estimated the average incidence in a population, without attempting to estimate tempo-

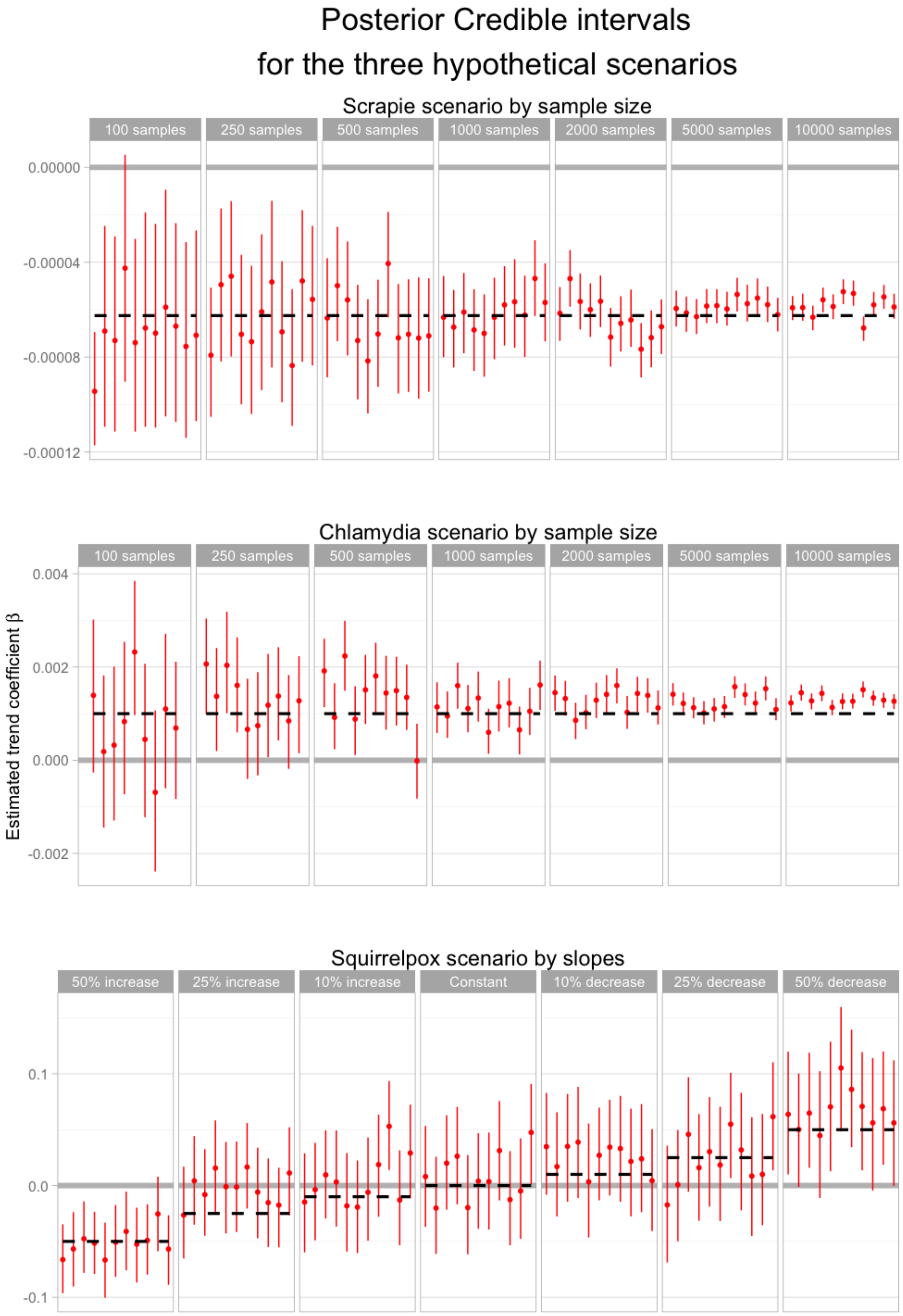

Figure 3.14: *Posterior Credible intervals for the slope parameter of the trend from three different scenarios with real-world plausible parameters. For the scrapie and Chlamydia scenarios, the trend was fixed and the sample size varied. For the Squirrelpox scenario, sample size was fixed at 250 animals, and the trend varied. Dots indicate mean posterior estimate of the slope; lines indicate 95% posterior credible interval. Black dashed lines indicate true trend, and light-grey line indicate zero.*

ral trends in prevalence (Baguelin et al., 2011; Nielsen et al., 2011; Simonsen et al., 2008).

The use of Lotka-Volterra (LV) equations to describe the pathogen-host dynamic, and thus the joint development over time of an antibody and an antigen-based test, made it possible to consider several different archetypes for the pathogen-host dynamic. Hindcasting was found to be possible for all of the different archetypes examined. Further, different archetypes proved to result in very similar overall hindcasting performance, with the exception of robustness to changes in number of tests used. The results on number of tests versus performance in section 3.6.3 seem to indicate that using a combination of diagnostic tests is more robust across the board than using a single diagnostic test; and that it may be possible to use a single diagnostic test for hindcasting, but it may also fail completely. This warrants further research into the specific requirements of disease kinetics for hindcasting.

The results obtained from evaluating scenarios with parameters close to observed epidemiological patterns in scrapie, Chlamydia, and Squirrelpox, indicate that useful precision levels can be reached with realistic sample sizes. The scrapie scenario indicated that on the order of 100-1000 samples could provide useful information. Given the size of the sheep industry, and that schemes for the surveillance of Scrapie already exists, such studies could feasibly be conducted (under the assumption that suitable test diagnostics could be found). The chlamydia scenario indicated that at least 500, and preferably more than 2000, crossectional samples would need to be taken. This would be a large study, but in settings where chlamydia surveillance is otherwise lacking, could potentially still be motivated by the overall public health costs of Chlamydia. Finally, in the case of squirrelpox, it was indicated that at least 250 animals would need to be caught and tested. Thisis a very large study in wildlife settings;

further, the trend would only rarely be as large as 50%. Hindcasting would thus only be usable for squirrelpox in certain special cases, such as a setting where it had only recently been introduced to a naive population, and it was needed to demonstrate that the introduction had happened.

The results described in this chapter indicate that the hindcasting procedure can provide reliable estimates of epidemic trends under a range of conditions. However, as indicated in section 3.6.1 and by figure 3.7, there seem to be a consistent, small bias when estimating the incidence and slope parameters. The source of this bias has yet to be fully explained. The prior distributions used in the hindcasting framework are weak, and in the case of the slope and incidence parameters, unbiased. Further, the density surface shown in figure 3.3, indicate that conditional on knowing the true time of infection, the maximum likelihood estimates for slope and incidence are unbiased. Together, this means that the source of the bias lie in the implementation of the MCMC algorithm and/or in the estimation of the times since infection. Indeed, figure 3.8 indicate that the proportion of very early infection tends to be underestimated (which would lead to the estimated trend lines being pulled toward zero, causing the described bias in the trend parameters). Further research would be required to understand why this is the case. Early infections are close to the edge of the support for the distribution of times of infection, and the lognormal distribution used for these has a discontinuity at $t = 0$ - both discontinuities and sampling at the edge of support could be creating problems for the MCMC sampler. A solution might be to choose a prior density that does not asymptotically vanish at $t = 0$, but it is not obvious what distribution(s) would be suitable. It could also be the case that the optimal step size is different for very early infections and for later infections. Since the lognormal distribution is implemented with tail switching, this could then mean that samples from early times are disproportionally likely to be rejected, biasing the resulting posterior samples. Drawing samples from a transformed variable might in that case be helpful but again, it is not obvious which transformations would be suitable.

In real world applications, the kinetics used to inform the hindcasting technique would likely be derived from other published data, such as experimental infection studies. In such cases, the LV calculations could be replaced with a simple lookup table for the expected mean response of the test at a given point in time, combined with information on the variability of the test. Alternatively models, such as the LV equations, fitted to available test kinetic data could be used.

The natural pairing of tests to be modelled with the LV approach is a nucleic acid test for genetic material from the pathogen (e.g. a realtime PCR test) or an antigen ELISA, combined with a test measuring the antibody test response, such as a quantitative ELISA test. However, any combination of two or more tests commonly used for pathogen diagnostics could be used, though a LV approach would no longer be suitable. Other examples are a pairing of a culture-based test combined with IGG antibodies, or even the severity of symptoms measured on an ordinal scale combined with viral load measurements. Thus, a wide range of diagnostic measures could potentially be used within the hindcasting framework presented here.

The results from this and the next chapter provide strong arguments in favour of recording raw test results together with the resulting diagnosis, and for utilising more than one diagnostic test whenever feasible. Thus, when setting up surveillance systems, it should be emphasised that the results of all diagnostic tests used should be recorded in the database. Such a database should also detail the quantitative level of evidence (i.e. the test level) in addition to the regular binary ”infected/non-infected” result. The cost of conducting and recording the result of two or more diagnostic tests should be considered in relation to the benefits. For example in terms of feedback to farmers and policymakers on the impact of control measures and for detecting any potential costly changes in the prevalence. It should also be noted that the methods introduced here enable such benefits to be derived from cross-sectional data and therefore the additional costs described above should be compared with the costs of running longitudinal studies.

An important extension to the work presented here is to consider more complex changes in pathogen incidence than simple linear trends. In principle, since the hindcasting procedure provides approximate times of exposure any model that describes the pattern of times of exposure could be considered. The linear trends described here are primarily suitable for endemic diseases. Therefore, in the following chapter, development of the hindcasting technique is continued by considering outbreaks of epidemic diseases.

CHAPTER 4: THE APPLICATION OF HINDCASTING IN EPIDEMIC SCENARIOS

The following chapter has been submitted as a paper to PloS Computational Biology on the 18th of June 2015, and is reproduced here in the format of that journal. The original title of the paper was *Using combined diagnostic test results to hindcast trends of infection from cross-sectional data*
The paper describes an expansion of the hindcasting framework described in chapter 3 to the hindcasting of disease outbreaks. In such settings the timescale of interest is shorter than for endemic diseases, and so the issue of reinfections can be ignored. Since the disease incidence change more rapidly, the assumption of a linear trend is exchanged for that of a lognormal trend. The hindcasting framework assuming lognormal trends is then applied to two case studies, based on one outbreak of bluetongue and one outbreak of whooping cough, to estimate the incidence trend in both increasing epidemics, as well as epidemics past their peak. In this way, it is shown that hindcasting can be used to determine the stage of an outbreak at the time of sampling, thus informing potential outbreak responses.

4.1 AUTHOR LIST

Gustaf Rydevik[123], Giles T. Innocent[1], Glenn Marion[1], Ross S. Davidson[2], Piran C. L. White[3], Charalambos Billinis[45], Paul Barrow[6], Peter P. C. Mertens[7], Dolores Gavier-Widén[8], Michael R. Hutchings[2]
[1]Biomathematics and Statistics Scotland (BIOSS), Edinburgh, UK

[2]SRUC, Edinburgh, UK

[3]Environment Department, University of York, York, UK

[4]Laboratory of Microbiology and Parasitology, Faculty of Veterinary Medicine, University of Thessaly, Karditsa, Greece

[5]Department of Biomedicine, Institute for Research and Technology of Thessaly, Larissa, Greece

[6]School of Veterinary Medicine and Science, University of Nottingham, Loughborough, UK

[7]The Vector-Borne Viral Diseases Programme, The Pirbright Institute, Pirbright, United Kingdom

[8]National Veterinary Institute (SVA), Uppsala, Sweden

Gustaf Rydevik Gustaf.rydevik@roslin.ed.ac.uk

Giles T. Innocent giles.innocent@bioss.ac.uk

Glenn Marion glenn.marion@bioss.ac.uk

Ross S. Davidson ross.davidson@sruc.ac.uk

Piran C. L. White piran.white@york.ac.uk

Charalambos Billinis billinis@vet.uth.gr

Paul Barrow paul.barrow@nottingham.ac.uk

Peter P. C. Mertens peter.mertens@pirbright.ac.uk

Dolores Gavier-Widén dolores.gavier-widen@sva.se

Michael R. Hutchings mike.hutchings@sruc.ac.uk

4.2 ABSTRACT

Infectious disease surveillance is key to limiting the consequences from infectious pathogens and maintaining animal and public health. Following the detection of a disease outbreak a response in proportion to the severity of the outbreak is required. In order to assess this severity, it is critical to obtain accurate information concerning the origin of the outbreak and its forward trajectory. However, there is often considerable uncertainty about the outbreak's history prior to first detection, which may lead to over- or under-reaction.

Data on the infectious status of individuals is accessible from a widening range of diagnostic tests that typically have different temporal characteristic, e.g. in terms of when peak test response occurs relative to time of exposure. We have developed a statistical framework that combines data from multiple diagnostic tests and is able to hindcast (infer historical trend of) an infectious disease epidemic prior to the time of detection.

Assuming diagnostic test data from a cross-sectional sample of individuals infected with a pathogen during an outbreak, we use a Bayesian Markov Chain Monte Carlo (MCMC) approach to estimate time of exposure and the overall epidemic trend in the population prior to the time of sampling. We evaluate the performance of this statistical framework on simulated data based on two historical outbreaks: a bluetongue outbreak in cattle, and a whooping cough outbreak in humans. The results show that hindcasting the outbreaks can provide accurate estimates of epidemic trends, whether an outbreak is increasing or past its peak. We conclude that it is possible to recover epidemic trends of both human and animal pathogens from cross-sectional data collected at a single point in time.

4.3 introduction

Infectious disease surveillance is the first line of detection and defence against infectious pathogens and therefore crucial to maintaining animal and public health. However, the current state of disease surveillance has been characterised as deficient in terms of both coverage and reporting speed for both humans (Butler, 2006) and animals (Mörner et al., 2002; The Royal Society, 2002). The challenge is to use the data generated by this often sparse and biased surveillance, to decide on an appropriate response to disease outbreaks. Any response needs to balance the social and economic consequences of the adopted control strategy with the social and economic risks posed by the outbreak (WHO, 2012b). In the case of the pandemic H1N1 flu in 2009, early analyses mistakenly assumed the epidemic had been only recently introduced, causing substantial overestimates of the basic reproduction ratio (Mercer et al., 2011) and case fatality rates (Echevarría-Zuno et al., 2009) that suggested a far greater risk to human life than was actually the case, leading to a more robust response than was necessary (Leung and Nicoll, 2010). The more complex settings typical of livestock and particularly wildlife systems tend to result in the available surveillance data being sparser still for animal diseases (Perez et al., 2011). In the UK, the absence of routine surveillance for Salmonella in poultry in the mid 1980s meant that the emergence of Salmonella Enteritidis PT4 was not recognised until it had become a major public health and political problem by 1988 (Rodrigue et al., 1990). Early identification of this epidemic caused by the new strain would have enabled faster intervention..

Using the data available when an epidemic is first detected to estimate its development at earlier times would help inform early decisions of the potential risks posed by an outbreak, leading to a more proportionate response than would be the case from waiting for the epidemic trends to be revealed by subsequent real time monitoring. In the current study, we introduce a novel statistical approach to recover population-level trends of exposure from only cross-sectional data by combining knowledge of the dynamic characteristics of multiple diagnostic tests to infer the timing of exposure events for individuals. Here we refer to the process of recovering such trends as "hindcasting", following terminology established in other papers (Banakar et al., 2011; Kleczkowski and Gilligan, 2007; Wethey and

Woodin, 2008) for reconstructing historical trends from currently available data.

Several papers have recovered limited historical characteristics of epidemics from cross-sectional data using a single diagnostic test, e.g. an antibody test. For example, Giorgi et al. estimated the time of the start of an HIV outbreak under assumptions of exponential growth of viral load (Giorgi et al., 2010). Others have exploited information on diagnostic test kinetics, i.e., the pattern of diagnostic test values during the course of infection, to estimate average incidence rates. Examples include the use of antibody test kinetics to estimate sero-incidence rates for influenza (Baguelin et al., 2011), *Salmonella* in cattle (Nielsen et al., 2011) and *Salmonella* in humans (Simonsen et al., 2008). One challenge in these kind of studies is that the relationship between the magnitude of signals from diagnostic tests and time since exposure is usually not monotonic; they tend to increase and then decrease. This means that the inverse problem of estimating time since exposure given a test value is non-unique and although this can be framed as a statistical problem the resulting inference is highly uncertain (Giorgi et al., 2010; Simonsen et al., 2009), limiting what can be estimated from test data. However, there are often several diagnostic tests available that target different aspects of the multi-faceted dynamic interaction between host and pathogen, and thus exhibit different test kinetics (Casadevall and Pirofski, 2000). That is, the profile of test responses, as a function of time since exposure, will differ depending on the underlying diagnostic used and the immunopathogenesis of the disease. Thus, in principle we can generate a unique signal for a given time since exposure by combining results of diagnostic tests that respond on different time scales. Here, we exploit this fact to develop a more robust statistical approach for analysing cross-sectional field data from multiple diagnostic tests. To do so we make use of empirical infection models that characterise test kinetics to infer the time since exposure for each individual. While there is a considerable uncertainty in the estimated exposure time for each individual, the combined estimates from multiple individuals can be used to describe the overall population-level distribution of infection times and estimate the shape of the overall epidemic trend with a high level of confidence. A detailed description of the hindcasting framework and case studies can be found in the methods section. We demonstrate the hindcasting of epidemic trends by applying the developed framework to case studies of real outbreaks of two contrasting diseases, whooping cough in humans and bluetongue in cattle (see Fig 4.1). For each disease, we investigate two scenarios representing detection during either the increasing or the decreasing phase of the epidemic. We conclude that when combined with two (or more) appropriate diagnostic tests (i.e. that differ in their temporal response following exposure, see Fig. 4.2) our methods allow historical epidemic trends to be recovered from cross sectional sample data. Moreover for the example diseases considered, suitable diagnostic tests already exist.

4.4 RESULTS

We applied the hindcasting framework (described in the methods section, below) to case studies based on a recorded outbreak of whooping cough in humans, and a bluetongue outbreak in cattle (see Fig 4.1). For each outbreak we assumed two scenarios, firstly where a cross-sectional sample was taken midway through the outbreak (increasing epidemic trend/early detection), and in a second scenario towards the end of the outbreak (decreasing epidemic trend/late detection). Based on published temporal characteristics of real diagnostics, test results were then simulated for these samples (see methods). For each disease (whooping cough and bluetongue) and each scenario (increasing and decreasing outbreaks) the hindcasting technique was applied to the corresponding test results to assess performance in recovering early increasing phases and late decreasing phases of outbreaks.

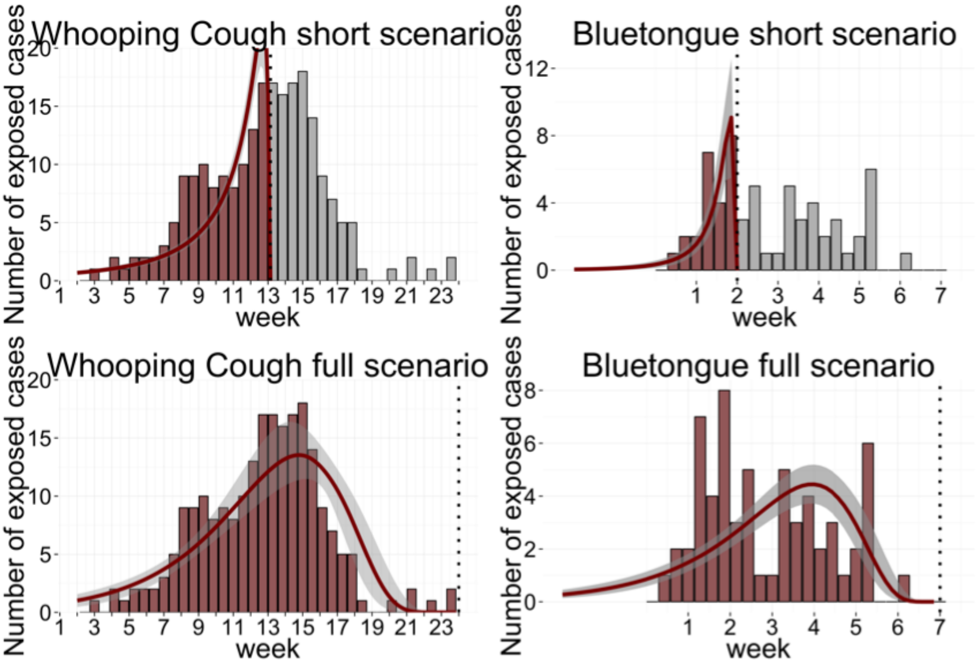

Figure 4.1: *Outbreak scenarios together with estimated epidemic curves. Top left: Testing 100 whooping cough cases at week 35 of the Wisconsin outbreak. Top Right: Testing 100 bluetongue cases at week 7 of the 2007 UK outbreak. Bottom left: Testing 25 cases at week 25 of the Wisconsin outbreak. Bottom right: Testing 30 cases at week 3 of the 2007 UK outbreak. In all scenarios, cases were sampled from the full population of cases shown in the outbreak data of Fig. 4.1 that had been exposed before the time of testing. Vertical dashed lines indicate time of cross-sectional sample. Red bars indicate cases included in the sampling frame for testing, grey bars indicate cases not included. Red lines indicate the mean posterior hindcast trend based on the cross-sectional test data. The grey transparent regions around the trends indicate the 95% posterior credible interval.*

We evaluated the robustness of the hindcasting framework by restricting the number of individuals sampled (i.e. in the outbreak data of Fig. 1 selecting only a subset of individuals infected at the sampling time), and by using only a single diagnostic test.

The results show this technique was able to estimate epidemic trends for both increasing and decreasing scenarios, in both whooping cough and bluetongue outbreaks (Fig 4.2). For the increasing whooping cough epidemic, when assuming a sample of all 122 cases that had occurred between the start of the epidemic up to week 25, the coefficient of determination (R^2) between underlying case counts (smoothed by a 7-day moving average) and the estimated epidemic trends was 0.74, with a 95% confidence interval of [0.69-0.78]. When sampling 230 cases from the full whooping cough epidemic up until week 36, after it had declined, the curve fit was somewhat better, with R^2 of 0.82[0.68-0.94].

We also looked at the bluetongue epidemic, which had substantially fewer cases. When assuming a sample of the 26 animals that had occurred during the increasing phase, during the first two weeks, the fitted curve was nearly perfect, with an R^2 of 0.9[0.86-0.92]). However, for the corresponding decreasing scenario, assuming a sample of the 61 animal cases that had occurred up to week seven, the hindcast trend could not fully capture the erratic nature of the underlying case count data, as indicated by R^2 values of 0.21[0.15-0.27]. Nonetheless, the trend did indicate an elevated incidence over the stretch of time when the majority of cases occurred, thus capturing the approximate time that had elapsed between the start of the epidemic and the time of sampling.

When reducing the sample size, the hindcasting technique was still able to recover both increasing and decreasing phases. The good fit was maintained with sample sizes as low as 20 individuals, with R^2 values for the whooping cough scenarios, of 0.77[0.27-0.83], for the increasing and 0.67[0.09-0.86] for the decreasing scenario. Similarly,

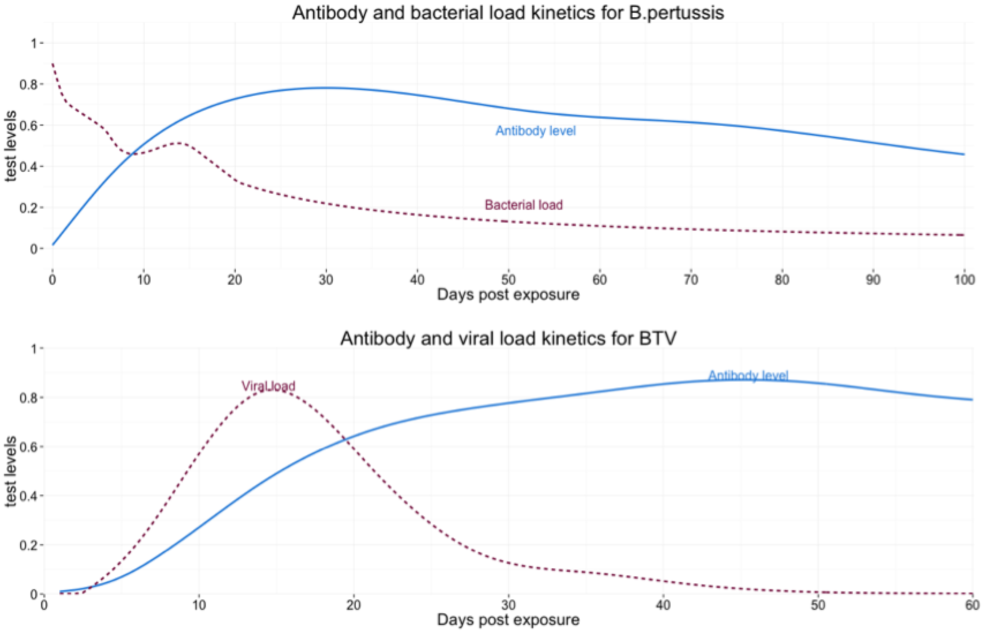

Figure 4.2: *Graphs of the kinetics of diagnostic tests used in the paper. Top: Diagnostic test kinetics for Whooping cough, with an antibody test(solid line) and a test measuring bacterial load (dashed line). Bottom: Diagnostic test kinetics for bluetongue, with an antibody test (solid line), and a test measuring viral load (dashed line). The graph is showing idealised test kinetics, based on published data on Pertussis (Bidet et al., 2008; Teunis et al., 2002) and Bluetongue tests (López-Olvera et al., 2010) (see methods for details).*

for the increasing bluetongue scenario, also assuming 20 samples, the R^2 was maintained at 0.91[0.87-0.93]).

The performance of the hindcasting technique was significantly hampered when using a single diagnostic test instead of two contrasting ones. For the increasing phase of the bluetongue and whooping cough outbreaks, the average performance was still acceptable when hindcasting trends based on a single antibody test from 20 individuals, but with substantially higher variability. The average R^2 was 0.83[0.04-0.86] for the increasing whooping cough, and 0.85[0.74-0.89] for the increasing bluetongue scenario. Hindcasting performed substantially worse when using only one test in the decreasing whooping cough scenario, with R^2 of 0.59[0.28-0.89] even when using the full set of cases, and failed completely when using just 20 cases (R^2 0.04[0-0.28]).

To get a better understanding of the effect of different tests on the performance of the hindcasting technique, we investigated how diagnostic test data and the combination of different tests affected the prediction of the time since infection. The images in Fig 4.3 show how the likelihood of estimated times of infection given observed test data varies as a function of actual time since exposure. Each pixel is coloured by generating 10 observations from the distribution of test measurements at a time since exposure given by the X axis, and calculating the likelihood for a time of exposure given by the Y axis, conditional on these observations. Areas in dark red indicate regions of higher likelihood. From the point of view of accurately recovering historic trends in the epidemic the ideal result would have the maximum likelihood values along the diagonal which would mean the likelihood of time since exposure was very firmly focussed on the true exposure time given observed data.

It was interesting to compare the kinetics of the whooping cough antibody test as seen in Fig 4.2, with the corresponding image (4.3a) of

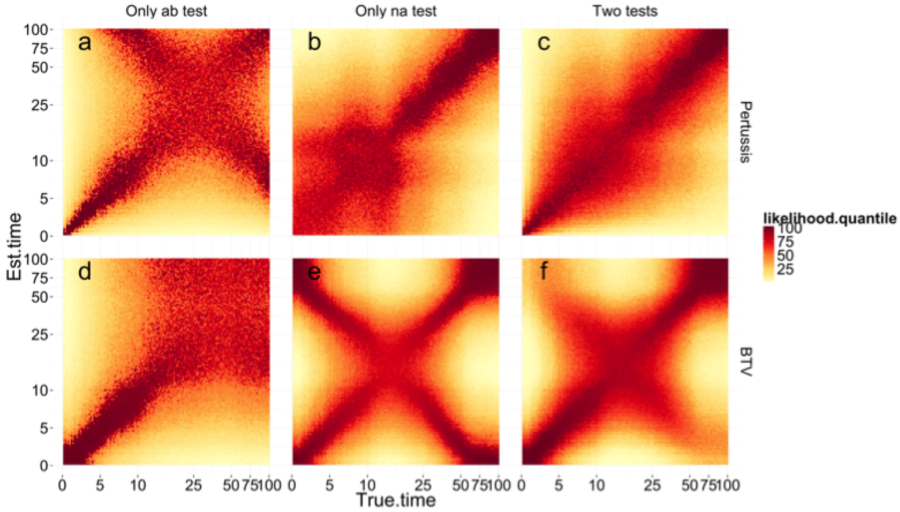

Figure 4.3: *The graphs show the log likelihood of inferred times of exposure as a function of true time since exposure, given test data generated assuming that the individual was exposed to whooping cough (top row) or bluetongue (bottom row) at the true time. Both the X and Y axes are on a log scale. Each pixel represents the value of the likelihood at a time of exposure given by the Y axis, given 10 test results, generated assuming a time since exposure given by the X axis.. The colour of the pixel indicate the likelihood for an estimated time, given the sample data, with dark red being most likely, and pale yellow being least likely. A clear, dark red diagonal indicates that the time since infection is easily recoverable, while a more diffuse diagonal indicates higher levels of uncertainty (see the results section for details). The first column shows results based only on data from the antibody test relevant to the disease in question, the middle column results based on an appropriate nucleic acid test, and the right hand column shows the results based on both tests (see text for details).*

the likelihood of exposure time, based on observing test data from this single diagnostic test. The times shown in both Figures 4.2 and 4.3 are times since exposure; short times since exposure represent more recent infections. The times of 20 days or less since exposure correspond to the phase of infection where the test response is increasing rapidly. Here, the probable infection times (red coloured pixels), given the data, are centred on the diagonal (i.e. the true exposure times) with a narrow band of high-probability red pixels. We can see that for times since exposure of greater than 20 days, when the kinetics of the antibody test are developing at a slower pace, the diagonal of red pixels becomes wider and more diffuse, indicating a greater variation around the true times since exposure. Furthermore, we can see that there are two different diagonals crossing at 25 days. This corresponds to the peak of the diagnostic response curve, with the two diagonals indicating the possibility that a given test result could have been the result of testing an individual during either the increasing or the decreasing phase of the response curve. In general, estimation of the time since exposure is more precise when the true time since exposure corresponds to phases where the response is changing rapidly, and is more difficult to infer when the test response levels out (4.3b and d). For diagnostic tests with a peaking response, estimating the time of infection can be precise but not unique, with two different regions of probable infection times for a given test response (4.3e). By combining early responding tests with later responders, it becomes possible to create a test signature that combines the best feature of both tests. The best combination of tests is the combination that provides unique and precise signature along the timeline of infection for an individual (4.3 c and f).

4.5 DISCUSSION

In this paper, we have shown that it is possible to recover epidemic trends of both human and animal pathogens from cross-sectional data collected at a single point in time. We were able to recover this temporal information using a novel statistical framework which combines paired diagnostic test measurements made on collected samples with the temporal kinetics of diagnostics test measurements over the course of infection.

The inferential framework introduced here enables utilisation of all the case data available up until identification of an outbreak. Here we focussed on purely cross-sectional samples but the methods are applicable to longitudinal data and data sets combing both longitudinal and cross sectional samples. We were able to estimate the trends of both increasing epidemics and decreasing epidemics, as well as estimate the approximate pace of increase or decrease. Such information would be valuable for tailoring appropriate management decisions immediately when an outbreak has been detected, without the need to observe subsequent spread to estimate the trend.

The implementation of the framework used in this paper combines surveillance data with information on the test kinetics using a simplified model. For example, individual variation in the test response is modelled as variation around a common mean test curve, rather than as variation in the shape of the curve itself. Variations in the two tests are considered independent, and the error distribution is assumed to be log normal. This limits the pattern and range of variation our model can capture, but facilitates model specification and estimation. More detailed modelling of the individual and population level processes in order to tailor the model to a particular disease is entirely consistent with the statistical framework introduced and would increase the real-world validity and predictive power beyond that shown in the applications presented here.

Likewise, we make use of the lognormal distribution as a parsimonious parametrization of the epidemic trend. This is suitable for epidemics where only a single peak is expected, allowing fast model fitting whilst capturing the time span and general direction of the trend. The trade off is that more complex aspects of trends in the epidemic cannot be captured. Moreover, the lognormal distribution requires the trend to decline to zero after any peak. Should either of these limitations pose a problem, more suitable models can be used, though such models are likely to come at higher computational cost.

The hindcasting framework introduced here estimate epidemic trends by combining observed data with information on how tests responses develop after exposure. Matthews and Woolhouse (2005) give an extensive overview of studies that incorporate different data sources to recover the underlying dynamics of disease spread (Haydon et al., 2003; Presanis et al., 2011), and argue that the future of disease analysis lies in models taking account of a wider range of inputs, such as diagnostic test performance, disease pathogenesis, or transmission mechanics, in addition to regular surveillance data. Our methodology improves on earlier studies incorporating test kinetics (Baguelin et al., 2011; Nielsen et al., 2011; Simonsen et al., 2008) in three ways: by incorporating information from more than one diagnostic test; by considering their joint kinetic pattern; and by modelling non-constant incidence.

Similar approaches could be used to model other aspects of the disease system such as population demography, contact networks, or spatial location, to estimate more complex aspects of disease dynamics e.g. not only temporal patterns of spread, but also, spatial spread, and the pattern of spread through the demographic structure of the population.

Hindcasting exploits knowledge of the host-pathogen interaction, and thus relies on previously conducted longitudinal studies of such interactions, and in particular on the test response after initial pathogen exposure. Our results demonstrate one of the many ways in which experimental infection studies can provide substantial additional benefits to disease control and research. Currently, only a fraction of pathogen tests have published information on how time since exposure affects test response; the method introduced here gives another reason why such studies on test kinetics are useful. We have described a new framework for hindcasting the temporal patterns of epidemics, using two example host-pathogen systems and the pairing of antibody tests with pathogen load. The framework demonstrates the potential to utilise the information inherent in the increasing variety of diagnostic tests. We were able to estimate both increasing and declining epidemic trends under the assumption that all individuals were being tested at a single point in time, implying its usefulness for cross-sectional surveillance data as well as in less restrictive settings. Recovering temporal incidence trends using multiple tests on cross-sectional field data has the potential to be of considerable value in the early phase of an outbreak and as a key determinant of introducing proportionate responses to newly detected disease threats.

4.6 METHODS

4.6.1 *Statistical framework*

Our method assumes test data y_{ik} from multiple disease diagnostics indexed by $k = 1, \ldots,$ K on individuals $i = 1, \ldots, N$. We assume that each individual is tested at some time t_i, after having been exposed to the pathogen at some earlier time e_i. We further assume that these individuals are chosen in an unbiased, random manner from a larger population. Each diagnostic test is assumed to return a value in the form of a continuous 'level', which might, for example be the highest dilution at which antibodies are detected in a serological test. Without loss of generality we assume that these levels are scaled to the unit interval [0,1].

Initial exposure to a pathogen is the start of a complex dynamical process within the host. We conceptualize such internal host-pathogen interactions as a multivariate process that depends on the time since initial exposure. Each diagnostic test is assumed to target the state of a different component of this process so that each test k carried out at time t_i on individual i can be modelled as a latent variable $l_{\mathrm{ik}}\left(t_i, e_i\right) = l_{\mathrm{ik}}\left(d_i\right)$, with each test having differing but correlated response patterns over the time since initial exposure $d_i = t_i - e_i$. We model these latent variables using results from experimental infection studies for a given host-pathogen system, where the length of time since initial exposure d_i is known.

The known data, across all individuals in the sample, comprises a set of test results denoted by $\mathbf{Y} = \{y_{\mathrm{ik}}\}$ with sampling times $\mathbf{T} = \{t_i\}$. Our aim is to infer the unknown set of exposure times $\mathbf{E} = \{e_i\}$, using information on the behaviour of the latent processes $\mathbf{L} = \mathbf{L}\left(\mathbf{T}, \mathbf{E}\right) = \{l_{\mathrm{ik}}(t_i, e_i)\}$ generating the test results. Note that when describing these sets the limits of each index $k = 1, \ldots, K$ and $i = 1, \ldots, N$ are implicit.

Under our statistical model we assume that the sampling times **T** are precisely known whereas the quantities **Y, L** and **E** are assumed to be subject to uncertainty and variation. There are thus three components to the statistical model: a latent process model $P(\mathbf{L}|\mathbf{T}, \mathbf{E}, \theta_L)$ describing uncertainty and variation in the host-pathogen interaction process within the host in terms of the time since initial exposure; a testing or observation model $P(\mathbf{Y}|\mathbf{L},\theta_Y)$ describing the distribution of results from tests carried out on the hosts conditional on the internal latent process; and an epidemic trend model $P(\mathbf{E}|\mathbf{T}, \theta_e)$, describing the historical development of the epidemic in terms of the distribution of exposure times in the sampled host population, at the time of sampling. We discuss specific implementations of each of these components in the examples described below.

Combining the three parts of the model, we write the full data likelihood given an observed data set $\{\mathbf{Y}, \mathbf{T}\}$ as

$P\left(\mathbf{Y}, \mathbf{E}, \mathbf{L}|\mathbf{T},\theta\right) = P\left(\mathbf{Y}|\mathbf{L}, \theta_Y\right) P(\mathbf{L}|\mathbf{T}, \mathbf{E},\theta_L)P(\mathbf{E}|\mathbf{T}, \theta_E)$,

where $\theta = \{\theta_Y, \theta_L, \theta_E\}$. Thus the likelihood combines models for testing with those for within and between host pathogen interactions. According to Bayes' theorem, the so-called posterior distribution for the unknown parameters is proportional to the data likelihood and prior $P(\theta)$. We can express this relationship for the parameters of interest, the latent process $\mathbf{L}$, the exposure times $\mathbf{E}$ and the parameters θ, given the observed test data $\mathbf{Y}$ and sampling times $\mathbf{T}$, by the equation

$$P\left(\mathbf{L}, \mathbf{E}, \theta|\mathbf{Y}, \mathbf{T}\right) = \frac{P\left(\mathbf{Y}, \mathbf{E}, \mathbf{L}|\mathbf{T},\theta\right) P(\theta)}{P(\mathbf{Y}, \mathbf{T})}$$

Within the Bayesian framework all inference is based on the posterior. The prior $P(\theta)$ can result from previous measurements or expert opinion, and represents knowledge about the values of parameters before we see the data used in the likelihood.

In what follows, we will make the simplifying assumption that the latent process $\mathbf{L}$ is modelled by a known deterministic function of $\mathbf{T}$ and $\mathbf{E}$, and represents the expected value of the test results given the times since exposure. This means that the term $P\left(\mathbf{L}|\mathbf{T}, \mathbf{E},\theta_\mathbf{L}\right)$ drops out of the likelihood which then simplifies to $P\left(\mathbf{Y}, \mathbf{E}|\mathbf{T},\theta\right) =$ $P\left(\mathbf{Y}|\mathbf{L}\left(\mathbf{T}, \mathbf{E}\right), \theta_Y\right) P\left(\mathbf{E}|\theta_\mathbf{E}\right)$, and the posterior becomes

$$P\left(\mathbf{E},\theta|\mathbf{Y},\mathbf{T}\right)=\frac{P\left(\mathbf{Y},\mathbf{E}|\mathbf{T},\theta\right)P\left(\theta\right)}{P(\mathbf{Y},\mathbf{T})}$$

Note that under this notation any parameters defining the deterministic latent process $\mathbf{L}\left(\mathbf{T},\mathbf{E}\right)=\{l_{\mathrm{nk}}(t_n,e_n)\}$ are suppressed since they are not inferred i.e. $\theta=\{\theta_Y,\theta_E\}$.

In both cases above the normalisation factor $P(\mathbf{Y},\mathbf{T})$ is typically unknown and computationally expensive to calculate. However, standard Markov Chain Monte Carlo (MCMC) methods circumvent this problem and are able to generate samples from the posterior even though the normalisation is unknown. The results presented in this paper are generated from an MCMC sampler implemented with a Metropolis-Hastings algorithm in JAGS (Plummer, 2003) using Gibbs sampling (Casella and George, 1992).

4.6.2 *Case studies*

Whooping cough is a human disease caused by the bacteria *Bordetella pertussis*, causing prolonged spasmodic coughing. Despite widespread vaccination coverage there has been a resurgence of cases in several countries; in the Netherlands there has been a steady increase in the incidence since 1996, and in California, USA in 2011, there was a widespread outbreak with 9000 cases and ten deaths (Winter et al., 2012). Possible reasons for such resurgence include decreasing vaccine coverage and/or antigenic drift (Greeff et al., 2010). Here we make use of data describing a countywide outbreak of Pertussis primarily among adolescents and adults in Fond du Lac County, Wisconsin, USA in 2003-2004, (Sotir et al., 2008). After an early cluster of cases in a high school in early May 2003, there was a large outbreak of Whooping Cough throughout the county starting from October. After some time, this outbreak was contained, and the final cases occurred in February 2004. The upper part of Fig 4.1

shows interpolated case counts per 48-hour period over the duration of the outbreak.

Bluetongue virus (BTV) is a midge-borne virus that can infect ruminants such as sheep, cattle, deer and camelids, causing bluetongue disease with symptoms such as internal haemorrhages, swelling of the tongue, lesions in the mouth, and in some species death (most notably in naïve sheep and white tailed deer). Bluetongue infections can have severe economic consequences for livestock farming, both due to loss of productivity, and because of the severe control measures needed to prevent spread (Alban et al., 2010). In 2006, BTV emerged throughout northern Europe, with recorded outbreaks in the Netherlands, Belgium, Germany, and Luxembourg. In 2007, the UK had its first recorded outbreak (DEFRA, 2008). The first infections occurred sometime in early August 2007 (DEFRA, 2008) when midges introduced the pathogen to the British Isles, but the first case was not detected until September. The lower part of Fig 1 shows the case count per day, with numbers interpolated from the published weekly data (DEFRA, 2008).

In order to assess our methodology we consider two scenarios for each pathogen outbreak. In the "increasing" scenarios we assume the epidemic is recognised early and explore test results from samples taken at a time early on in the outbreak (when the outbreak is increasing, see e.g. Fig 4.1). In contrast in the "decreasing" scenarios we use test results assumed to be obtained from individuals exposed during the entire outbreaks, with samples collected at a relatively late stage in the outbreak (i.e. when it is in decline). The goal was to see how well hindcasting could distinguish between such increasing scenarios, and scenarios where the epidemic had declined. We were also interested to see if it was possible to estimate the approximate time span of the epidemics, measured as the longest estimated exposure time among tested cases.

4.6.3 *Implementation of the Whooping Cough scenarios*

Outbreak data and detection scenarios

We based our whooping cough data set on the case count curve of the 2003 Wisconsin whooping cough outbreak. We used published bi-weekly case counts, and interpolated these using a LOESS (Cleveland and Devlin, 1988) approach to generate estimated 48 hour case counts.

We investigated two hypothetical scenarios for when the outbreak could have been first detected and cases tested. We simulated one scenario where we assumed that cases were sampled and the samples tested 25 weeks after the first observed case. At this time point, the first wave had passed, and the second sharp increase in incidence had been going on for about a month. 126 cases had been reported by this time in the actual outbreak. The second scenario assumed that testing of cases occurred at week 36, taking samples from the 230 cases from the full whooping cough epidemic up until that time. This time point marks the end of the epidemic, with no later cases reported.

Diagnostic test characteristics

The results of diagnostic testing are characterised in terms of an underlying mean trend and a model which accounts for variation around this reflected measurement error, and within and between individual variability in test response.

The sampled cases produced from the scenarios as described above were assigned simulated test results, based on the elapsed time between the time of exposure in the actual outbreak and the assumed time of sampling, using published kinetics of real-time PCR analysis and quantitative ELISA for Pertussis (Bidet et al., 2008; Teunis et al., 2002), to inform a latent process $P(\mathbf{L}|\mathbf{T},\mathbf{E},\theta_L)$. Specifically these were the kinetics of ELISA IgG pertussis antitoxin (Teunis et al., 2002) for antibody test response ab (d) as a function of time since exposure d, and real-time PCR measurement of persistence over time of *Bordetella pertussis* DNA in nasopharyngeal secretions (Bidet et al., 2008) (see Fig 4.2) for the pathogen load DNA (d) . As noted earlier formally, we defined the deterministic function $\mathbf{L}(d_i) = (D\text{NA}\,(d_i)\,, ab\,(d_i))$ by fitting interpolated curves to the published data on DNA and antibody levels using LOESS (Cleveland and Devlin, 1988).

The distribution $P\,(Y_i|\mathbf{L}\,(d_i))$ of test measurements was modelled as a lognormal distribution conditional on the state of the latent process: let $\mathbf{y_i} = (y_{\text{NA}}, y_{\text{ab}})_i$ represent a bivariate measurement of nucleic acid and antibody levels on individual i, and define the distribution $\text{P}\,(Y_i|\mathbf{L}\,(d_i)) = \mathbf{l}\mathcal{N}(\mathbf{L}\,(d_i)\,, \Sigma^2))$, where Σ^2 is a diagonal covariance matrix, reflecting the assumption of no correlation between test results when conditioned on the time since exposure. The variance for each test (i.e. the diagonal elements of) was assumed known. Antibodies as well as level of pathogens in a host often follow log-normal distributions, as has been rigorously argued (Koch, 1966); the suitability of using the lognormal distribution for modelling a wide range of biological phenomena has also been described more recently (Limpert et al., 2001).

Epidemic trend

A lognormal distribution was also used to parameterize the unknown distribution $P(\mathbf{E}|\mathbf{T}, \theta_e)$ of times of exposure given the time of testing, $P\,(\mathbf{E}|\mathbf{T},\ \theta_e = \{\mu, \sigma\}) = \mathbf{l}\mathcal{N}(\mu,\ \sigma)$. (See section 4.7 for technical details). In this case, we exploit the ability of the lognormal to model extreme skewness to capture both increasing and decreasing epidemics using only two parameters.

4.6.4 *Implementation of the Bluetongue scenarios*

Outbreak data and detection scenarios

As with Whooping Cough, we assumed two different hypothetical scenarios for when the outbreak was noticed and animals tested: one assuming that animals tested are sampled from the 26 exposed cases at week two, and the other assuming that animals tested were sampled from the 61 animals exposed by the end of week seven (Fig 2).

Diagnostic test characteristics

We modelled test behaviour based on published data (López-Olvera et al., 2010), and assumed lognormal distributions for the epidemic trend, as well as for the variance of the diagnostic tests (Fig 3). Specifically we based the behaviour of the latent process $P(\mathbf{L}|\mathbf{T}, \mathbf{E}, \theta_L)$ on a study of experimental infection of European red deer with BTV serotype 1 and 8 that described the dynamics of BTV serotype 1 viral load (vl) as measured with RT-PCR, and antibody levels (ab) as measured with ELISA (López-Olvera et al., 2010). As above we define the latent process describing antibody concentration and viral load as a deterministic bivariate function of the duration d elapsed since exposure as $\mathbf{L} = \{\mathbf{l}(d_i)\} \equiv \{vl(d_i), ab(d_i)\}$ which does not vary between individuals. We estimate $\mathbf{L}$ by fitting smooth and interpolated curves to the experimental study data on viral load and antibody levels independently and take the values of these curves at each exposure time d to define the values of the deterministic functions, $\text{vl}(d), ab(d)$. LOESS (Cleveland and Devlin, 1988) was used as a nonparametric fitting method.

Conditional on the time since exposure, the observed test values $\mathbf{y}_i = (y_{\text{vl}}, y_{\text{ab}})_i$ were modelled as a bivariate log-normal distribution with mean equal to the deterministic latent process $= \{\mathbf{l}(d_i)\} = (vl(d_i), ab(d_i))$. For individual i, this can be formally written as

$P\left(\mathbf{y}_i|\mathbf{l}\left(d_i\right)\right) = \mathbf{l}\mathcal{N}(\mathbf{l}\left(d_i\right), \Sigma^2)$, where $\mathbf{l}\mathcal{N}$ indicates a bivariate lognormal probability function, and Σ^2 is the covariance matrix. We assumed that the variation in observed antibody levels and viral loads to be independent so that the covariance matrix Σ^2 is diagonal, with variance components σ_1^2, σ_2^2. The variance for each test (i.e. the diagonal elements of) was assumed known.

4.6.5 *Epidemic trend*

The third and final part of the model, the distribution of times since exposure $(\mathbf{E}|\mathbf{T}, \theta_e)$, was modelled as a lognormal distribution

$$\mathrm{P}\left(\mathbf{E}|\mathbf{T},\ \theta_e = \{\mu, \sigma\}\right) = \mathbf{l}\mathcal{N}(\mu,\ \sigma)$$

(see section 4.7 for further details).

4.6.6 *Choice of priors*

We followed the recommendations of Gelman (2006) and used a combination of informative and weakly informative priors for the parameters. The means for the lognormal distribution describing the epidemic trends were themselves given lognormal priors. It was assumed that any prior information about the start of the epidemic would be correct to within an order of magnitude. This translated to prior means for the increasing whooping cough of 100 days, and prior means for the decreasing whooping cough scenario of 200 days. The corresponding values for BTV were 10 days and 100 days, respectively. The standard deviations for the prior distributions were chosen as $\log\left(\sqrt{3}/2\right)$, corresponding to 95% confidence intervals of $(mean * 1/3, mean * 3)$. This standard deviation was chosen to model a confidence that the "best guess" was within a factor 10 from the true values.

The standard deviation of the lognormal distributions for the epidemic trend was assumed to follow a vague folded t-distribution, with five degrees of freedom and standard deviation of log(100). The high standard deviation was chosen so as to allow even an epidemic trend that is currently increasing to have cases occurring several hundred days ago.

We also attempted to use uninformative priors, but these were found to lead to slow and /or failing convergence of the MCMC algorithm. Changing the specific values of the priors did not influence the posterior estimates noticeably.

4.7 SUPPLEMENTARY INFORMATION

4.7.1 *Evaluation of convergence*

When evaluating convergence of the MCMC runs used to estimate the parameters of the epidemic trends of outbreaks (see the results section in main body of the paper), the Gelman-Rubin (GR) statistic (Gelman and Rubin, 1992) was used as a first indicator. This widely used statistic (Kathryn et al., 1996) measures the ratio of the within-chain variability to the between-chain variability e.g. for multiple chains started from different initial conditions.

The convergence behaviour (Fig 4.4) was quantified for the MCMC algorithm for each of the 100 different data sets generated for each combination of scenario, type of disease diagnostic, and the number of individuals tested (see main text). For each of these datasets the MCMC samples were generated using a 10 000 iteration burn in, and subsequent capturing of the following 10 000 iterations. In each case 5 chains were run in parallel to ensure sufficient information on between-chain variability.

As indicated in Fig 4.4, for the whooping cough scenarios, the GR statistic was nearly always below the threshold of 1.1 recommended by Gelman et al. (2004). The convergence was less fast for bluetongue scenarios using paired disease diagnostics with only approximately 50% of the runs having converged by 20 000 iterations as measured by the threshold standard.
Fig 4.5 show examples of traceplots from scenarios with diagnostic test results from 100 individuals. The patterns of the bluetongue traceplots indicate that that chains are mixing slowly, which explains the low values of the GR statistic obtained. Fig 4.6 shows the estimated epidemic trends using mean parameter values from the final 5000 iterations of the MCMC algorithm, with separate trends for each of the 5 chains. These plots indicate that despite the slow mixing, the practical difference between chains was miniscule.

4.7.2 *Generating simulated data*

The bluetongue and whooping cough data sets we applied the hindcasting inference framework to were simulated in R using the following procedure:

- Choose an observation time T after the first recorded case of the epidemic we base our simulated data on.
- Record the observed exposure times $\mathbf{E} = \{e_i\}$ for all cases that had occurred in the real epidemic up until the time of observation T and evaluate the duration of infection/exposure for each individual $\{d_i = T - e_i\}$ at time T.
- Use the deterministic function $\mathbf{l}(d_\mathbf{i})$ based on the interpolated test trajectory to assign mean test values for each case given the durations of infection
- Set the variance Σ^2 of the log-normal distribution to correspond to published test variability.

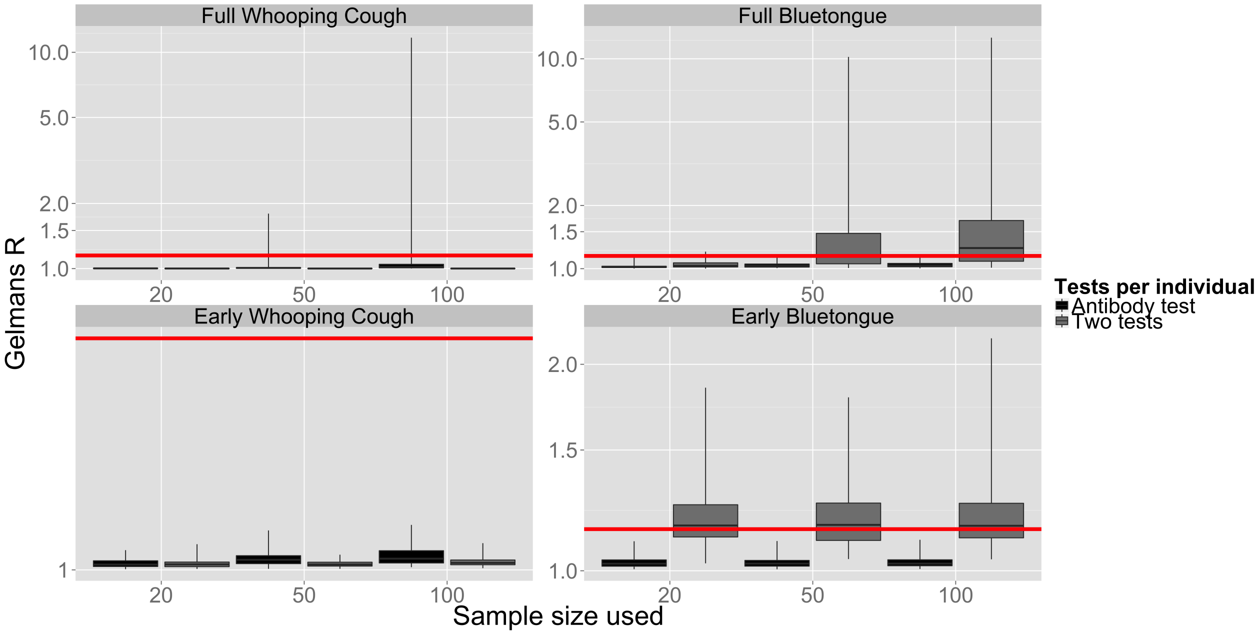

Figure 4.4: *Distribution of the GR statistic across the different scenarios. Each box represents the distribution of GR calculated on different sets MCMC samples, the samples generated by fitting the hindcasting model 100 times to different data sets. The runs represented by each box was all generated assuming the same combination of scenario, diagnostic tests used, and number of individuals sampled. The middle line of the boxes indicate the median, the top and bottom of the boxes indicate 25 and 75% quantile, respectively, and the thin black lines indicate the range of results. The horizontal red lines indicates a GR value of 1.1, which is considered to be an indicator that the chains have reached full convergence*

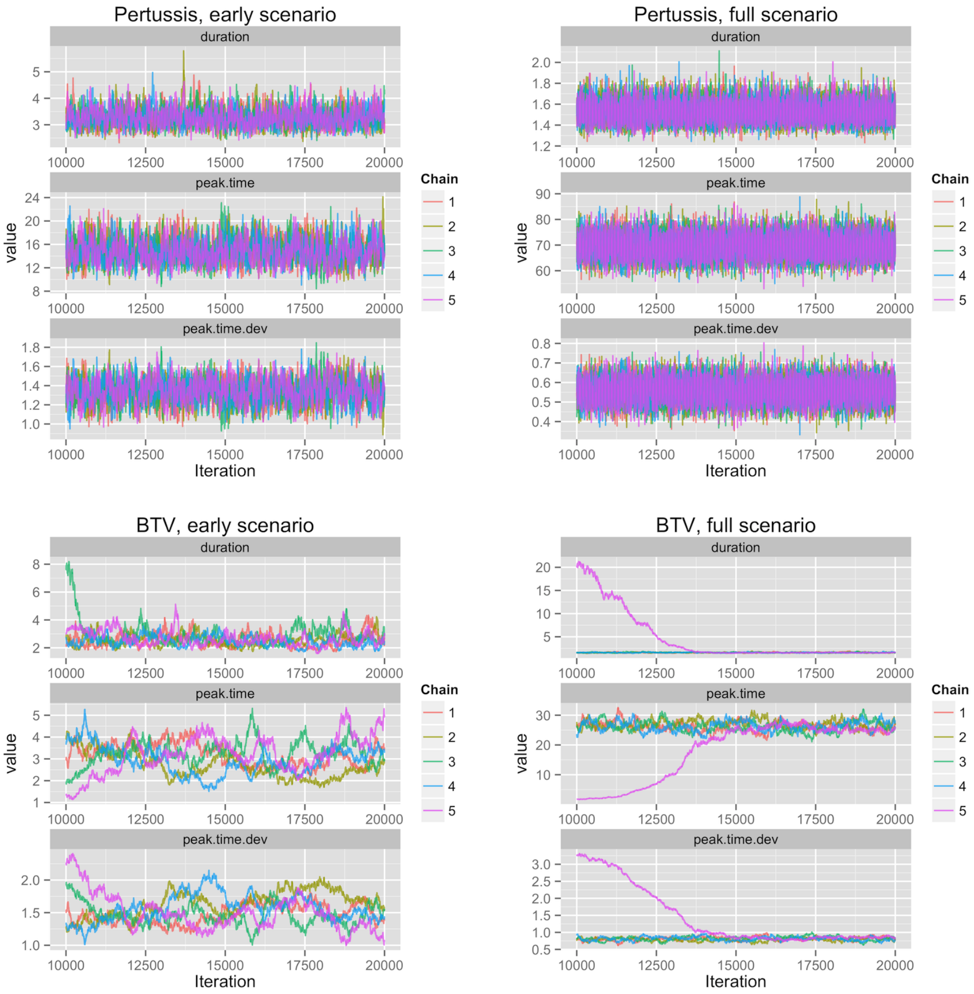

Figure 4.5: *Examples of MCMC traceplots, generated by fitting the hindcasting model to scenarios assuming that disease diagnostics were collected from 100 individuals. Each colour indicates a different chain (i.e. started with different initial conditions). "Peak.time" and"duration" are the parameters for the mean and variance of the lognormal distribution describing the epidemic trends. "Peak.time.dev" is the variance of the prior of peak.time.*

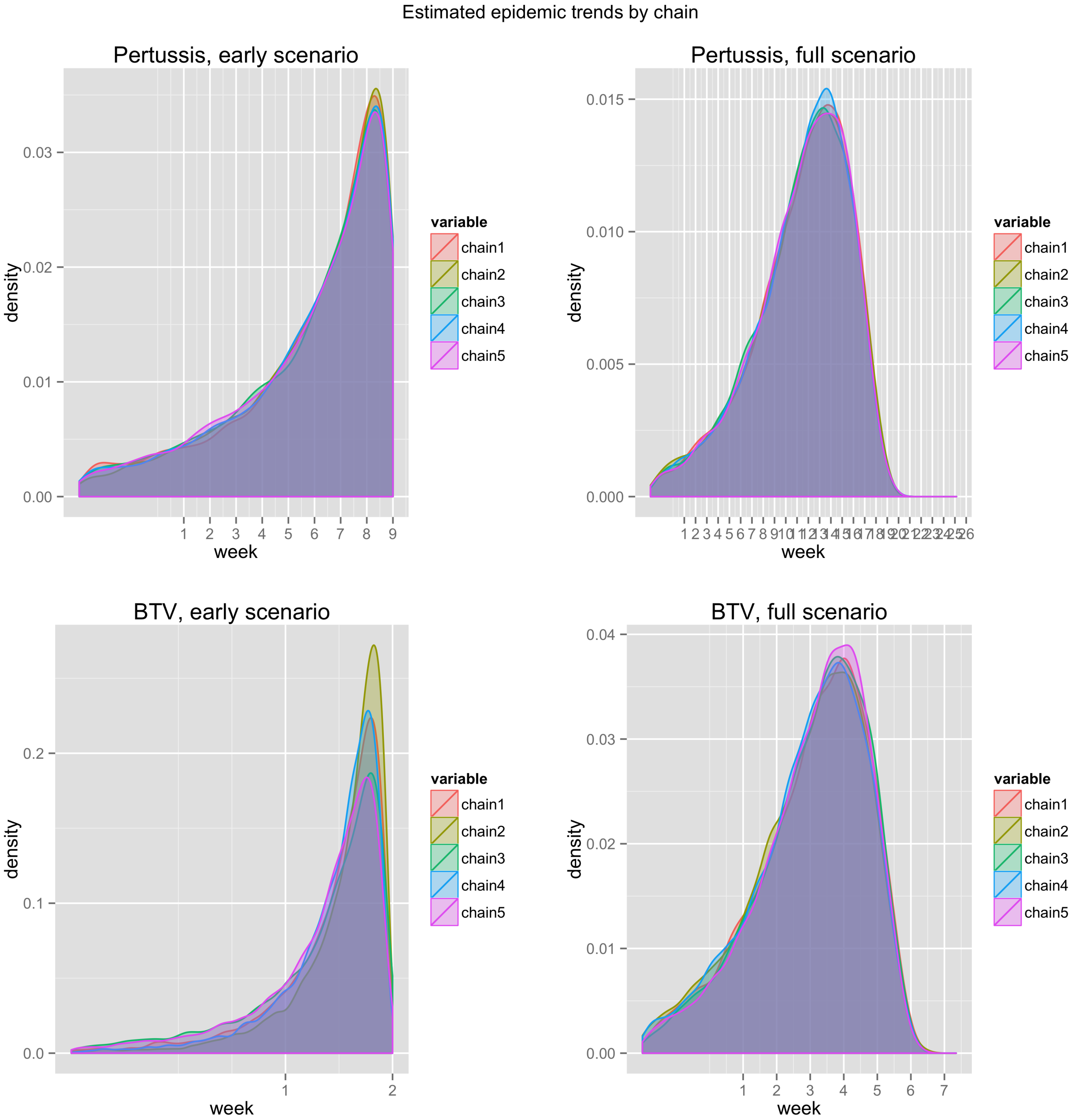

Figure 4.6: *Estimated epidemic trends by chain. The parameters of the lognormal distribution describing the epidemic trend was calculated from the mean of the last 5000 samples of "peak.time" and "duration" in the traceplots in Fig 4.5, above.*

- Generate test results from the log normal distribution with Σ^2 variance and mean test values generated in steps 3 and 4.

Test scores corresponding to viral load and antibody response are then simulated using the deterministic function $\mathbf{L} = \{\mathbf{l}(d_i)\} \equiv \{vl(d_i), ab(d_i)\}$ described in the methods section. Note that the infection times are not subsequently used in the inference procedure, but they do provide an opportunity to assess the inferences obtained.

The lognormal variation around this curve was set to be 57% following data on *B. Pertussis* antibody variability published by Versteegh et al. (2005). DNA measurements and antibody measurements were assumed to have the same level of noise.

Based on published data (Chatzinasiou et al., 2010) on variability of RT-PCR we set the log-normal variability to 27%. For simplicity, we assume equal variability for antibody measurements.

4.7.3 *An alternative formulation of the lognormal distribution to allow the MCMC sampler to efficiently sample multimodal likelihood surfaces*

As shown in figure 4.3 in the main text, the likelihood for times since exposure is often multimodal. Combining multiple tests can reduce this problem, but if the MCMC sampler is initiated in the wrong region, convergence can be an issue where the chain becomes 'stuck' around one mode with an extremely small probability of jumping to an alternative mode. By allowing the MCMC sampler to jump between disjoint regions of exposure times in one step, the different modes are no longer isolated from each other. The following describes how to implement such a solution in the case of the lognormal distribution used in the paper.

Assume that times of exposure $\mathbf{e}_i$ are lognormally distributed, $e_i \sim logN(\mu, \sigma)$ For the standard lognormal parametrization, the density for a particular time of exposure **e** given mean μ and variance σ is given by

$$p\left(\mathbf{e}|\mu,\sigma\right) = lN\left(\mathbf{e}|\mu,\sigma\right) = N\left(\log\left(\mathbf{e}\right)|\mu,\sigma\right) = \phi(\frac{\log\left(\mathbf{e}\right) - \mu}{\sigma})$$

, where $\phi\left(x\right) = N\left(x|\,\mu = 0,\ \sigma = 1\right)$.

Tautologically, we can rewrite **e** as

$$\mathbf{e} = e^{\log(e)} = e^{\mu+\log(\mathbf{e})-\mu} = e^{\mu+[(\log(\mathbf{e})-\mu)/\sigma]*\sigma} = \ e^{\mu+\Delta*S*\sigma}$$

where $S \in \{-1,1\}$, and $\Delta = \left|\frac{\log(\mathbf{e})-\mu}{\sigma}\right| \geq 0$.

Now, instead of assuming that **e** is coming from a lognormal distribution, we can

assume that S has a discrete probability function,

$$a)\ p\left(S = 1\right) = p\left(S = \ -1\ \right) = 0.5$$

, and that Δ has a folded standard normal distribution with a probability density function given by

$$b)\ p\left(\Delta,\ \Delta > 0\ \right) = \phi\left(\Delta\right) + \phi\left(-\Delta\right) = 2\phi\left(\Delta\right)$$

In this way, Δ can be interpreted as how far **e** is away from the mean of the lognormal distribution, measured in the number of standard deviations, and S indicates whether it is in the upper or lower quantile. Note that for fixed $\mu$and σ, each exposure time **e** can be written uniquely as a combination of Δ and S.

$$p\left(\mathbf{e}|\mu,\sigma\right) = p\left(\Delta,S|\mu,\sigma\right) = p\left(\Delta\right)*p\left(S\right)$$

From a) and b), we get that

$$p\left(\Delta\right) * p\left(S\right) = \; 2 * \phi\left(\Delta\right) * 0.5 = \phi(\Delta) = \phi(|\frac{\log\left(\mathbf{e}\right) - \mu}{\sigma}|)$$

Thus, (since the normal distribution is symmetric) this new formulation results in the same probability distribution for T as the lognormal distribution, and thus an equal contribution to the data likelihood. For the bluetongue and whooping cough examples, the full posterior likelihood is written as

$$L\left(\mathbf{E}, \theta|\mathbf{Y}, \mathbf{T}\right) \sim \left(\prod_{\forall i}\left(\; 1\mathcal{N}(\mathbf{Y}|\mathbf{L}\left(\mathbf{T} - e_i\right), \Sigma^2)\right)1\mathcal{N}(e_i\,|\mu,\; \sigma)\right) \text{Prior}\left(\mu\right) Prior(\sigma)$$

Using the new formulation, this becomes

$$p\left(\mathbf{E}, \theta|\mathbf{Y}, \mathbf{T}\right) \sim$$

$$\left(\prod_{\forall i}\left(\; 1\mathcal{N}\left(\mathbf{L}\left(T - e^{\mu + \Delta_i * S_i * \sigma}\right), \Sigma^2\right)\right)\phi\left(\Delta_i\right) * 2 * p(S_i)\right) \text{Prior}\left(\mu\right) Prior(\sigma) =$$

$$\left(\prod_{\forall i}\left(\; 1\mathcal{N}\left(\mathbf{L}\left(T - e^{\mu + \Delta_i * S_i * \sigma}\right), \Sigma^2\right)\right)\phi\left(\Delta_i\right)\right) \text{Prior}\left(\mu\right) Prior(\sigma)$$

In an MCMC setting, this formulation allows for generating a new proposal T' by jumping from $(\Delta, S^+) \rightarrow (\Delta, S^-)$, thus reducing the risk of getting stuck in local maximums of the likelihood. In effect, by decomposing **e** into two separate variables, we are adding an extra dimension that the MCMC sampler can jump through, bringing the separate modes closer together.

CHAPTER 5: THE CHALLENGES OF ESTIMATING TEST KINETICS

5.1 INTRODUCTION

In previous chapters, we have shown how to use hindcasting to understand the past spread of infectious diseases from cross-sectional studies, and thereby better inform current actions. This estimation of historic trends from recent samples is done by treating diagnostic test measurements as "clocks" that can indicate the time since pathogen exposure. The development of test measures as a function of time since exposure is often referred to as "test kinetics", and knowledge of these kinetics is crucial in order to be able to calibrate the clocks used for hindcasting. Understanding test kinetics is also important in other contexts such as producing accurate diagnoses (Pawlotsky, 2002); estimating the duration of host immunity (Hallander et al., 2005); and understanding the pathogenesis and dynamics of the within-host infectious processes (Kirschner and Linderman, 2009).

During development of the hindcasting framework, it was necessary to consider how and where to find information on test kinetics. However, a major bottleneck for current kinetic studies is the reliance on experimental or longitudinal follow-up data, which are difficult to collect. Consequently, published information on kinetics is scarce. This chapter represents a preliminary investigation of whether it is possible to estimate test kinetics from observational data in a way that makes them usable for modern statistical procedures such as hindcasting, or sero-incidence studies that rely on kinetics to inform inference.

Commonly used sources for data on test kinetics are experimental infection studies, longitudinal follow-up studies, and observational studies. Each study design bring with it a particular tradeoff between factors such as ease of implementation, study size, study cost, number of observations per time unit, and bias in the resulting estimates. In an experimental infection study, such as that by López-Olvera et al. (2010) used in Chapter 4 to model BTV kinetics, individuals are infected in a controlled manner with a particular dose of pathogens, and then tested using available diagnostics at regular intervals until the end of the experiment. Such studies require strict ethical consideration, and can require access to facilities with high biosafety level ratings. In some cases, such as with biosecurity level 4 (BSL4) pathogens, (US Department of Health and Human Services, 1999) only a handful of institutions world-wide have the suitable containment facilities to carry out experimental infections . Because of the costs of containment, experimental studies tend to only infect a small number of individuals. This means that estimated test kinetics can fail to be fully representative for the population as a whole. The question of how representative such studies are of the entire population is intensified by the fact that study criteria usually exclude individuals that suffer from sickness, malnutrition, and other conditions that are common in field situations (OIE (2013), p. 10). In addition, modes of primary and secondary disease contact in experimental conditions can be far removed from those encountered in field conditions.

Alternatively, naturally occurring infections can be tracked in longitudinal follow up studies, such as the study by Versteegh et al. (2005) that was used to model Pertussis antibody kinetics in Chapter 4. Infected cases are detected via routine or enhanced surveillance, and are then tested repeatedly for some length of time thereafter. On the one hand, such studies do not come with the same limitations as experimental studies: kinetics estimated based on naturally occurring cases are more likely to be representative of other cases occurring in the field; there are fewer ethical considerations; and a larger number of individuals can be included in the study for the same cost. On the other hand, identifying cases poses a challenge, especially for rare or emerging pathogens, as can identification of the initial infection time. Furthermore, repeated testing of cases can be difficult or impossible, in particular for wildlife pathogens in free-ranging animals.

The term "observational study" is here used for studies conducted on samples from infected individuals that has been discovered (observed) in the course of regular surveillance, and not necessarily comprising of a random sample from the entire population. This implies that samples are collected from naturally infected individuals, with no follow-up sampling. This is even more limiting than follow-up studies, since there is no information on how the test measurements of the individual develop over time. However, it is a logistically much easier design than a follow-up study and often, especially in the case of rare, less studied, or emerging pathogens, the only available source of kinetics data.

An ideal inference method would therefore be able to estimate test kinetics even from regular observational studies. Given the above considerations, it would need to be usable even when only a single observation per individual is available, and account for uncertainty in the time of infection. In situations with limited observations per individual, it becomes important to leverage as much as possible from previously obtained knowledge of the host-pathogen dynamics. Describing within-host disease processes with mathematical models can make it possible to fit the kinetic curve parsimoniously using a few key parameters. Given that such processes are driven, at a minimum, by two components, namely immune response and pathogen growth, this implies that incorporating multiple tests that correspond to each of these components into mathematical and/or statistical models would improve the estimate of the test kinetics. The work of Simonsen et al. (2009) is one of the more detailed published papers that estimates test kinetics from population data. They used a mathematical model of antibody kinetics that also implicitly describes antigen levels, and longitudinal follow-up data on *Salmonella* immune response to estimate the parameters of the model. However, the study was limited by not having access to antigen data, and the assumed model for antigen response was therefore quite simplistic.

In section 5.2, the model of Simonsen et al. is described and used as a stepping stone for considering the benefit of including multiple measurements of different aspects of the host-pathogen system, by explicitly incorporating antigen measurements. Section 5.3 then discusses the challenges for developing methods for estimating test kinetics from observational data to inform the growing number of studies and models that use test kinetics as clocks, such as hindcasting and sero-incidence studies (e.g. Teunis et al. (2012), Simonsen et al. (2009), Simonsen et al. (2011), Baguelin et al. (2011) and many more).

5.2 a suggested approach for improving the joint estimation of multiple test kinetics

5.2.1 *Reimplementing Simonsen etal (2009) with added nucleic acid measurements*

Simonsen et al. (2009) describe a study in two parts: part 1 consisted of a longitudinal study of *Salmonella* antibody kinetics where 302 patients with a culture-confirmed diagnosis of *Salmonella enteritidis* or *Salmonella typhimurium* infection were tested up to four times over an 18 month period following the detection of infection. Samples thus generated were tested for three different types of antibodies: IgG, IgA, and IgM. The test results were then used to estimate the kinetics of these antibodies after infection, assuming the reported dates of onset of symptoms were the actual times of infection. Part 2 of the study involved a cross-sectional sample of the population, where 1780 individuals were tested for antibodies against *Salmonella*. Using the kinetics of antibody levels following exposure estimated in part 1, the average time since exposure to *Salmonella* could be calculated and used to estimate the incidence of exposure, or sero-incidence, of *Salmonella* in the population.

In the Simonsen et al. study, only antibody measurements were collected, and therefore antigen levels were only included implicitly in the model as a driving force for the antibody response. However, the formulation they used allows us to include antigen measurements in the observation process with only a slight adjustment, and thus explore the potential benefits of including antigen measurements.

The model they used was reimplemented using the RStan (Stan Development Team, 2014) MCMC engine (see section 1.5.3). A model with only antibody measurements was compared with a model that assumed paired antibody and antigen measurements. For simplicity, it was assumed that only a single type of antibody was measured instead of the three types of antibodies considered in the original paper.

The modelling approach is based on describing the interaction of a particular antibody level (denoted by $IG(t)$) and pathogen load (denoted by $Z(t)$) by the following pair of equations:

$$\frac{dIG(t)}{dt} = S \times Z(t) - a \times (IG(t) - X^{\star})$$

and

$$\begin{cases} Z(t) = (1 - t/D) \text{ for } t < D \\ Z(t) = 0 \text{ for } t \geq D \end{cases}$$

These equations represent a system where pathogen load is at its maximum value at the time of infection, and then decreases linearly with time until the pathogen is cleared. The antibody levels increases with S units for each unit increase of pathogen load per time unit, and a proportion a of the difference between current antibody level and the steady state antibody level $X^{\star}$ is removed per time unit. Thus $X^{\star}$ represents the background antibody level that is present even in the absence of infection. These are first-order linear differential equations which allows them to be solved as a function of time since infection t. Doing so leads to the following explicit expressions for $IG(t)$, under the condition that $IG(0) = X^{\star}$:

$$IG(t) = X^{\star} + \frac{S + aS(D - t))}{Da^2} - \frac{S + aSD}{Da^2}e^{-at}$$

if $t < D$, and

$$IG(t) = X^{\star} + \frac{Se^{aD} - S - aSD}{Da^2}e^{-at}$$

otherwise.

The statistical model for observed test values is then constructed in the following manner: Assume that a set of test results $X = \{x_{ij}\}$ has been obtained by taking samples from individuals $i = 1, \ldots, N$, with each sample j having been taken at certain times since exposure e_{ij}. The model is hierarchical; the test response for each individual i is assumed to be governed by individual parameters θ_i, which in turn are sampled from a population-level lognormal distribution, $\theta_i \sim \log N(\Theta_2, \Sigma_\theta)$, where Θ_2 denotes the population means for the parameters, and Σ_θ denotes the covariance matrix between the parameters. To facilitate legibility, let $\Theta_1 = \{\theta_1, ...\theta_N\}$ denote the set of individual parameters i.e. across all individuals in the population. Given the parameters and an infection time for individual i, the test values $X = \{x_{ij}\}$ taken at times $j = 1, 2, \ldots, N$ are then assumed to come from a separate log-normal distribution. In the case of antibody-only tests, this can be written as

$$x_{ij}|e_{ij}, \theta_i \sim \log N(IG(e_{ij}|\theta_i), \Sigma_{IG})$$

where Σ_{IG} denotes the variability of the IG levels around the value of the mean test kinetics at time e_{ij}, assuming parameters θ_i for the test trajectory.

Given these elements, the posterior for the model given observed test values X can be written as a product of the likelihood for the population-distribution parameters Θ_2, the likelihood of the individual parameters Θ_1 given the population distribution, and the likelihood for the observed test values X given individual parameters, and the priors for the parameters. The posterior probability thus becomes

$$P(\Theta_1, \Theta_2, \Sigma_\theta|X) \propto \prod_{\forall i} \left[P(\theta_i|\Theta_2, \Sigma) \prod_{\forall j} \left(P(x_{ij}|\theta_i, e_{ij})\right)\right] Prior(\Theta_2) Prior(\Sigma_\Theta) Prior(\Sigma_{IG}) =$$

$$\prod_{\forall i} \left[\log N(\theta_i|\mu = \Theta_2, \Sigma_\Theta) \prod_{\forall j} \left(\log N(x_{ij}|\mu = IG(e_{ij}|\theta_i), \Sigma_{IG})\right)\right] \times$$

$$Prior(\Theta_2) Prior(\Sigma_\theta) Prior(\Sigma_{IG})$$

When including antigen measurements, $IG(t)$ is replaced with $F(t) = \{IG(t), Z(t)\}$, and $x_{ij} = (x_{ab}, x_{na})_{ij} \in X$ are then defined as vectors of antibody and antigen test levels, with values coming from a bivariate lognormal distribution:

$$x_{ij} = (x_{ab}, x_{na})_{ij} \sim \log N(\mu = \{IG(t), Z(t)\}, \Sigma_F)$$

where Σ_F is a 2x2 covariance matrix; it will be assumed that the off-diagonal elements are zero, i.e. that the two components of the lognormal distribution are conditionally independent. The individual variance of the two components may differ, however. Note that when modelling antibody and antigen measurements jointly, X becomes a two-column matrix.

All parameters were given vaguely informative priors (as discussed in section 1.5), that regularise the posterior distribution without requiring precise information about the parameters. The log of the population-level parameter means Θ_2 were given $Cauchy(0, 2)$ priors. The between-parameter covariance Σ_θ was defined as the product of a $LKJ(1.5)$-prior on the between-parameter correlation and a $Cauchy(0, 2)$ priors on the variance around each parameter mean. The notation LKJ refers a distribution on correlation matrices that was introduced in Lewandowski et al. (2009) ("LKJ" stands for the first letters of the surnames of the authors). The measurement variances Σ_F were given exponential priors with parameter $1/\log(1.05)$. In order to compare the inferences when collecting either a single or a double test measurement, RStan was used to implement both the model with only antibody levels, and the model with both antibody and antigen levels. These models were then applied to simulated data sets with known parameter values. For both models, individuals were assumed to be sampled at three different time points over the course of infection,with the sampling times being randomly sampled from a uniform $U(0, 300)$ distribution. After confirming that the sampler had converged using MCMC trace plots and the

Gelman-Rubin statistic, the resulting posterior estimates could then be compared to the true underlying parameter values. Fig 5.1 shows resulting estimates of the antibody and antigen kinetic curves, compared with the underlying true population-level kinetics, and the observed individual-level measurement trajectories. As can be seen, the estimated mean kinetics were nearly identical for the two-test and the one-test models, both for antibody kinetics and for antigen kinetics, and close to the population-level mean trajectories. However, in the one-test model, the variation of individual-level antigen kinetics is non-existent, while the estimated spread is closer to the actual spread in the two test model. Since the one-test model lacks any data on antigen levels, any estimated variance would be strongly influenced by the Cauchy prior on the log parameter scale, which places a large proportion of the prior density at low values and therefore has a shrinking effect. The fact that the two test model underestimates the actual variability can be explained by the fact that very few individuals have more than one measurement before the antigen has declined fully, making it difficult to estimate the exact slope of this decline.

Despite the these difficulties in estimating variability, these results seem to indicate that one could estimate antigen kinetics without measuring mean antigen kinetics. However, this is likely a consequence of the assumption that the antigen levels follow a simple linear decay with fixed initial value 1. For some situations, such as for pathogens with a very rapid initial growth, or where the timescale of interest is long relative to the increasing phase, this may well be an adequate description. In general, however, a more flexible description of the development of pathogen load is clearly desirable.

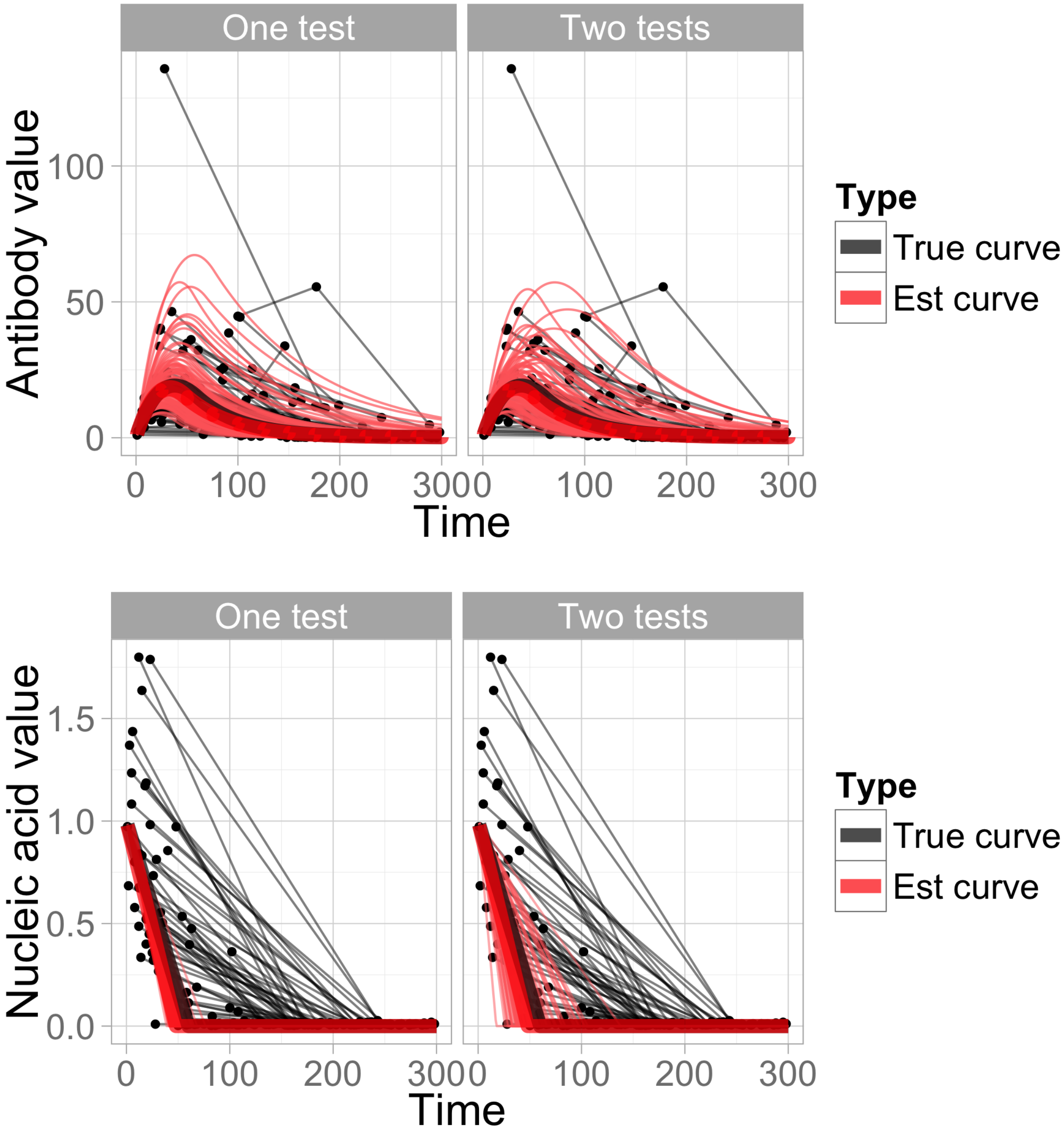

Figure 5.1: *Estimated antibody and antigen kinetics using the Simonsen et al. (2009) model on a generated data set. Results are from either a model including only antibody test("One test") or antibody and antigen test measurements ("Two tests"). Thin red lines indicate estimated test kinetics for the individual, and the black dots measured test values observed at these times; thin black lines connect the three samples taken from each individual. The thick red line indicates the estimated population-mean kinetics, while the thick black line (mostly hidden by the estimated red line) indicates the true population-mean kinetics.*

5.2.2 *Using piecewise linear curves to describe pathogen load kinetics*

Assuming a linear trend is a very simplified and unrealistic model for the antigen kinetics. This was not an issue for the Simonsen et al. paper as it was primarily focused on antibodies, and was interested in using the test kinetics on a time scale where antigen presence was low. However, it is a clear obstacle when the antigen levels are of interest as well. A major reason for assuming a linear trend is that this leads to mathematically tractable explicit solutions for the antibody curves, a requirement when implementing inference with the current generation of MCMC engines (see section 1.5.3). A more flexible approach would be to include differential equations, such as the Lotka-Volterra equations used in Chapter 3, directly in the statistical model, and use ODE-solvers to generate numerical solutions. This would remove the need for analytical solutions, and allow for a wide range of dynamics to be modelled. An ODE solver with limited capacity has recently been made available in RStan (with the v2.5.0 release on 20 October 2014). An attempt was made to use the RStan implementation, but unfortunately, it was not found to be sufficiently robust for modelling noisy and partially observed data.
An alternative approach to including more realistic pathogen load kinetics in observational studies would be to allow the kinetics to follow a piecewise linear curve. Here we show that such an approach is compatible with currently available statistical tools such as Rstan as it remains mathematically tractable, while allowing for greater flexibility in the kinds of patterns that could be captured (see Figure 5.2). Using the framework described in the previous section, this can be implemented by replacing the expression for $Z(t)$ with a piecewise linear curve, and solving the systems of equations in this more general case.

Recall that the explicit expressions for $IG(t)$ in the simplified Lotka Volterra model, under the condition that $IG(0) = X^\star$:

$$IG(t) = X^\star + \frac{S + aS(D-t))}{Da^2} - \frac{S + aSD}{Da^2} e^{-at}$$

if $t < D$, and

$$IG(t) = X^\star + \frac{Se^{ad} - S - aSd}{Da^2} e^{-at}$$

otherwise.

If the above is generalised by assuming that $Z(t)$ can be any linear trend, i.e. $Z(t) = \alpha t + \beta$, the expression for $IG(t)$ instead becomes:

$$Ig(t) = X^\star + \frac{\alpha a S t - \alpha S + \beta a S}{a^2} + ke^{-at}$$

where k is a constant chosen so as to match any boundary conditions.

Recall that the original formulation assumes that the pathogen load follows a simple linear decrease until it is cleared. If the pathogen load is instead allowed to have a piecewise linear trajectory, then the resulting differential equations are still solvable, but a wider range of kinetics can be captured. Specifically, assume that the pathogen load passes through four stages instead of two: an increasing phase, a steady phase, a declining phase, and a cleared phase. This can be parametrised by

$$Z(t) = \begin{cases} t/R, \text{ for } t < R \\ 1 \text{ for } R \leq t < t_D + R \\ 1 + (t_D + R - t)/D \text{ for } t_D + R \leq t < R + t_D + R \\ 0 \text{ for } t \geq R + t_D + D \end{cases}$$

R is then the initial rate of increase, t_D is the length of time that the pathogen load is constant, $t_D + R$ is the time it starts declining, and D is the rate of the decline.

Using these expressions for $Z(t)$, the explicit form for $IG(t)$ becomes

$$IG(t) = \begin{cases} X^\star + \frac{S(at-1)}{Ra^2} + k_1 e^{-at} \text{ for } t < R \\ X^\star + \frac{S}{a} + k_2 e^{-at} \text{ for } R \leq t < t_D + R \\ X^\star + \frac{aS(D+t_D+R-t)+S}{Da^2} + k_3 e^{-at} \text{ for } t_D + R \leq t < R + t_D + D \\ X^\star + k_4 e^{-at} \text{ for } t \geq R + t_D + D \end{cases}$$

In order to find the constant k_1, we note that $IG(t)$ must be continuous at (and thus equal at either side of) all change points, and that $IG(0) = X^\star$.

For k_1, this implies that

$$IG(0) = X^\star + \frac{S(a*0-1)}{Ra^2} + k_1 e^{-a*0} = X^\star \Rightarrow k_1 = \frac{S}{Ra^2}$$

Given k_1 and the change point at R, we can then calculate k_2 by the following:

$$IG(R) = X^\star + \frac{S(aR-1)}{Ra^2} + \frac{S}{Ra^2}e^{-aR} = X^\star + \frac{Sa}{a^2} + k_2 e^{-aR} \Rightarrow$$

$$k_2 e^{-aR} = \frac{S(aR-1)}{Ra^2} + \frac{S}{Ra^2}e^{-aR} - \frac{Sa}{a^2} \Rightarrow k_2 = \frac{S(aR-1)e^{aR}}{Ra^2} + \frac{S}{Ra^2} - \frac{Sae^{aR}}{a^2} =$$

$$= \frac{aRSe^{aR} - Se^{aR} + S - aRSe^{aR}}{Ra^2} \Rightarrow$$

$$k_2 = S \times \frac{1-e^{aR}}{Ra^2}$$

Again, given k_2, and the change point at $t = t_D + R$, we get k_3 from the following:

$$IG(t_D + R) = X^\star + \frac{Sa}{a^2} + S \times \frac{1-e^{aR}}{Ra^2}e^{-a(t_D+R)} =$$

$$= X^\star + \frac{aS(D + t_D + R - (t_D + R)) + S}{Da^2} + k_3 e^{-a(t_D+R)} \Rightarrow$$

$$k_3 = \left(\frac{Sa}{a^2} + S \times \frac{1 - e^{aR}}{Ra^2} e^{-a(t_D+R)} - \frac{S(Da+1)}{Da^2}\right) e^{a(t_D+R)} =$$

$$= S \times \left(\frac{RDae^{a(t_D+R)} + D(1 - e^{aR}) - R(Da+1)e^{a(t_D+R)}}{RDa^2}\right) =$$

$$= S \times \left(\frac{D(1 - e^{aR}) - R(Da - Da + 1)e^{a(t_D+R)}}{RDa^2}\right) \Rightarrow$$

$$k_3 = S \times \left(\frac{1 - e^{aR}}{Ra^2} - \frac{e^{a(t_D+R)}}{Da^2}\right)$$

Finally, using the value for k_3, k_4 is calculated by looking at the value for $IG(t)$ at $t_D + D + R$:

$$IG(t_D + D + R) = X^\star + \frac{aS(D + t_D + R - (t_D + D + R)) + S}{Da^2} +$$

$$S \times \left(\frac{1 - e^{aR}}{Ra^2} - \frac{e^{a(t_D+R)}}{Da^2}\right) e^{-a(t_D+D+R)} = X^\star + k_4 e^{-a(t_D+D+R)} \Rightarrow$$

$$k_4 = S \times \left(\frac{e^{a(t_D+D+R)} - e^{a(t_D+R)}}{Da^2} + \frac{1 - e^{aR}}{Ra^2}\right)$$

The expressions defined above combine to define joint antibody/antigen test kinetics, allowing for a piecewise linear development of the antigen kinetics, and consequently a more nuanced model for the antibody kinetics as well. Figure 5.2 shows an example of the trajectory of kinetic curves using this new parametrisation.

These expressions could directly replace simpler models for test kinetics used when fitting curves to directly or indirectly observed multiple test data. In the model described in 5.1.1, it is simply a matter of redefining $F(t) = \{IG(t), Z(t)\}$. However, when using MCMC-type algorithms for fitting piecewise-defined curves, care needs to be taken with the sampling algorithm and priors to ensure convergence.

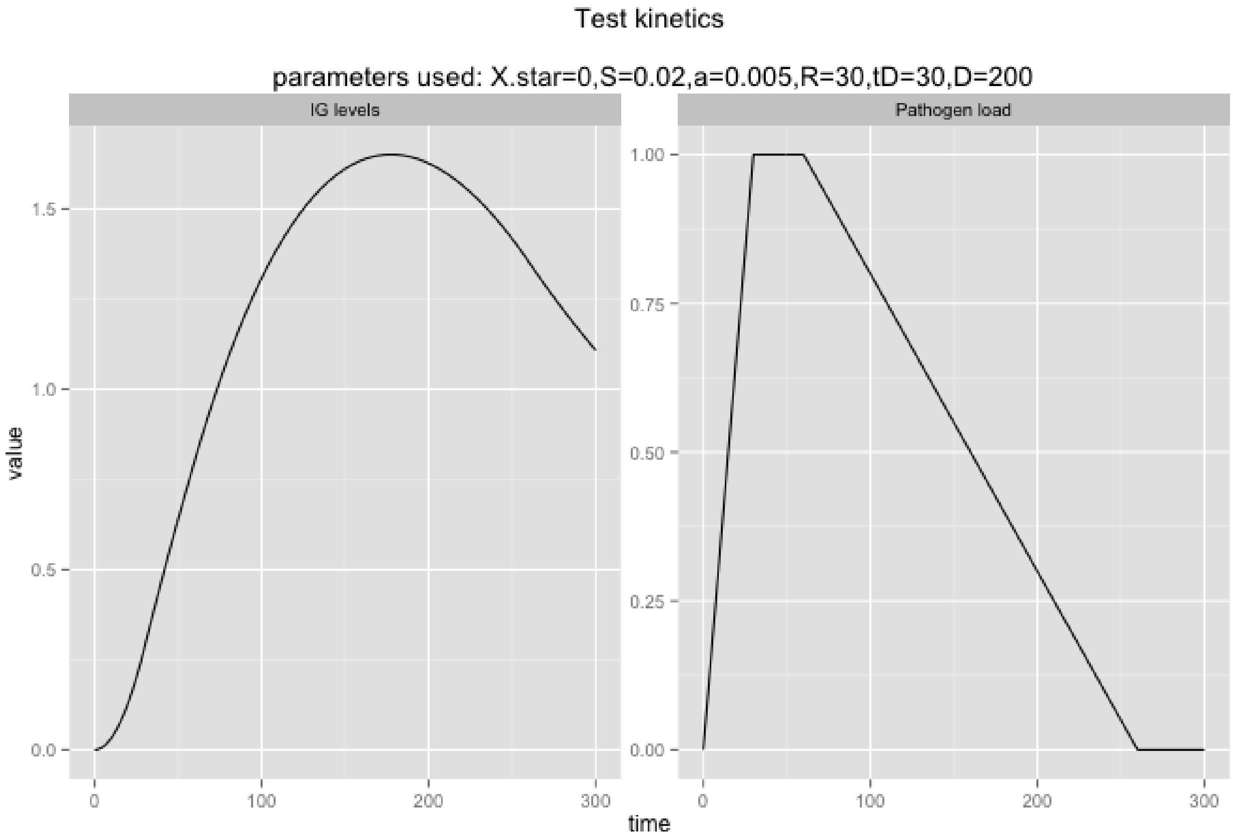

Figure 5.2: *Example trajectories for test kinetics using piecewise linear trends for the pathogen load kinetics*

Unfortunately, preliminary naive implementations of the piecewise linear model in RStan struggled with the discontinuous derivatives inherent in the piece wise linear formulation. In general, models in which the parameterisation of a curve changes at known or unknown "knots" are known as *change point models*, and there is an extensive literature for fitting these in an MCMC context (the introduction of Fearnhead (2006) provides an overview). There is no reason to believe that the piecewise linear curves described here could not be fitted via such approaches, but it requires additional work on the sampling implementation, which is beyond the scope of this thesis.

5.3 DISCUSSION

This chapter has described two extensions to existing methods for the estimation of test kinetics; first a description and a worked example of how to infer antibody- and antigen-based tests jointly; and second, a suggestion for how the kinetics of antigen based tests could be modelled more realistically using piece-wise linear curves. The study by Simonsen et al. (2009) that was used here as a starting point is part of a group of studies (Graaf et al., 2014; Simonsen et al., 2009; Teunis et al., 2012; Versteegh et al., 2005) that estimate test kinetics using a statistical approach for fitting mathematical equations (derived as the solutions to simple differential equations) to longitudinal measurements. All of these are restricted to only measuring antibodies, and including the antigen development in their model as an unobserved variable. Based on the results shown here, these studies would likely benefit from expanding the scope of inference to including antigen measurements (note that for many pathogens, this would require that cases are sampled at an early stage before the infection has cleared). In order for resulting estimates of antigen kinetics to be usable as clocks indicating time since infection, such expansions should use a flexible parametrisation for the antigen kinetics. A piecewise linear curve would be one option for doing so, while still allowing for the joint antibody and antigen kinetics be solved analytically. Such tractability is extremely useful in the implementation of inference using the current generation of MCMC engines. However, such practical requirements are likely to be relaxed for the next generation of such tools.

In that case another option would be to model the dynamic systems without the constraint of having to provide analytical solutions. There exists a substantial literature on the simulation of viral within-host kinetics (Canini and Perelson, 2014), that makes use of complex systems of differential equations. These equations typically use a large number of parameters; Heffernan and Keeling (2008) model the dynamics of Measles infection using a complex system of differential equations with 19 parameters to capture the interplay of CD8 T-cells, peripheral blood mononuclear cells (PBMCs) and measles virus, while Ciupe et al. (2007) use a system of differential equations with 14 parameters to model Hepatitis B infection. Indivdual parameters estimates are taken from the literature - fitting all model parameters simultaneously would require that a wide range of measurements be taken on a large number of subjects for the parameters to be identifiable. As a consequence, this approach is limited to very intensely studied systems of pathogen and hosts.

Systems of ordinary differential equations (ODE), such as the Lotka-Volterra equations used for simulation test kinetics in chapter 3, that are not analytically tractable but less complex than those used in the within-host kinetic models mentioned above, would be a natural intermediate step between these two extremes. As an added benefit, such relatively parsimonious models would be able to be fitted not only to well-studied disease systems, but also to rare, neglected, or emerging diseases. However, this approach would require implementing a custom MCMC sampler, since the ODE solvers in off-the-shelf software such as RStan are not yet stable enough to be used in fitting models to noisy data.

The approach by Simonsen et al. (2009) and others replicated and improved upon in this chapter rely on longitudinal follow-up measurements of infected individuals, and furthermore treat times of infection as known without uncertainty. For pathogens where repeated (longitudinal) testing of cases is difficult, the only available data may be single cross-sectional samples taken from infected individuals. In this situation, the only known information available for an individual would be the value of a test measurement at a particular time, and no information on how those values behaved earlier or later. Consequently, it would not be possible to estimate how trajectories differ between individuals, or indeed identify individuals that deviate in a systematic manner from the average. However, if the individuals from whom samples are taken cover the whole range of the infection process (i.e. different individuals in the sample were infected very recently, a very long time ago and at all points in between), population-average test kinetics may still be estimable. For the purpose of inference approaches such as hindcasting, which focus on population-level patterns, this may still be sufficient. An initial attempt was made to fit such a model, and initial results were promising, but had difficulty converging, and time did not permit further exploration or development.

The assumption of infection times being known without uncertainty could also be problematic, in particular for non-livestock animals which are not continuously monitored for disease, or for diseases with a substantial delay between the time of exposure and onset of symptoms. In such situations, the times of exposure $E = \{e_{ij}\}$ could be replaced with the times of observation $O = \{o_{ij}\}$. These could then be related to exposure times by assuming that the time elapsed from exposure until observation $\{\epsilon_{ij}\}$ follow some random distribution modelling the incubation period of the disease (and possibly other factors, such as the probability of an individual being observed), such as a gamma distribution (Nishiura, 2007). In this way, it would be possible to replace the times of exposure in the framework described in 5.2 via the relation $\{e_i j\} = \{o_{ij} - \epsilon_{ij}\}$. This would require modelling the time delay until observation together with the other parameters and would increase the uncertainty of estimated test kinetics, but would likely still be possible. Attempts were made to fit such models including incubation terms, and initial attempts were promising.

A major obstacle to the use of hindcasting and similar methods is the difficulty of finding usable data describing the kinetics of diagnostic tests as a function of time since infection. Most published experimental infection studies (Cray and Moon, 1995; Hoffman et al., 2006; Komar et al., 2003; López-Olvera et al., 2010; Major et al., 2004) only present their results in the form of graphs and summary statistics, which poses an obstacle to incorporating the results in statistical procedures. Similarly, while the longitudinal studies and the mathematical modelling studies mentioned above sometimes (but not always) present parameter estimates, producing kinetic curves from these estimates can be very difficult. The lack of easily accessible data may be because statistical methods exploiting such kinetics in epidemiological studies are not yet in wide use. It would be highly beneficial if kinetics data were more widely available in easily usable formats. Each study may be expensive, but making the results available would allow the study to have a long-lasting impact, which is clearly beneficial to the individual scientists, as well as to society as a whole. Most importantly, having the kinetics from diagnostic tests available would make approaches such as the hindcasting described in previous chapters, usable on a wide scale and in a wide range of settings.

CHAPTER 6: GENERAL DISCUSSION: CONTEXT, IMPLICATIONS AND FUTURE WORK

In the field of infectious disease surveillance, there is often a paucity of data, nationally (Parliamentary Office Of Science & Technology, 2014) and globally (World Health Organization, 2000), which constitute a substantial challenge for monitoring endemic trends, and detecting epidemics, and the emergence of novel pathogens. Lack of information further complicates the already challenging task of managing disease risks and deciding on appropriate actions. It is therefore imperative to make full use of the data that has been collected.

With the the rapid scientific development of computer-based methods for Bayesian inference (Lunn et al., 2009), it is increasingly feasible to integrate multiple sources of data into coherent models. Woolhouse and Matthews (Matthews and Woolhouse, 2005) give an extensive overview of studies that incorporate different data sources to recover the underlying dynamics of disease spread (Haydon et al., 2003; Presanis et al., 2011), and argue that the future of disease analysis lies in models taking into account a wider range of inputs, such as quantitative diagnostic test measurements, disease pathogenesis, or transmission mechanics, in addition to standard case count data. The work presented in the thesis is in line with such arguments, combining information on quantitative test measurements with knowledge of the within-host dynamics in the form of test kinetics, and developing methods that can be used to inform policy response in situations where only cross-sectional data are available. The thesis was carried out as part of the WildTech Framework 7 project.

WildTech was a project to develop advanced multiplex diagnostics for detecting wildlife diseases, using both nucleic acid and immune response based tests. The inspiration for hindcasting came from this setting of multiple tests and combined nucleic acid and immune response data. The simple case of two paired diagnostic test measurements was used as natural starting point for developing methods that exploit the synergy from multiple tests. However, as noted earlier the Bayesian and computational tools used here lend themselves to the integration of data from multiple sources. Therefore, as the use of multiplex testing develops further the framework for inference developed in this thesis could be applied to the results from multiple (i.e. >2) diagnostic tests.

6.1 SUMMARY OF THE RESULTS

The simplest case of multiple testing is that of using two tests with binary outcomes. Compared to a single test, two tests used in two populations with different incidences enable estimation of test sensitivity and specificity as well as the incidence levels. This is commonly done using a latent class model approach, first introduced in Hui and Walter (1980). This approach has become very popular; a recent review by Smeden et al. (2013) identified 111 published papers that used latent class models to evaluate diagnostic tests in humans and animals, the large majority published after 2000. In chapter two of this thesis, a Hui-Walter type latent class analysis was developed for a situation involving vaccinated animals. A simulation study was used to analyse the relationship between the properties of the newly developed Distinguishing between Infected and Vaccinated Animals (DIVA) tests for bovine tuberculosis, and the study size used to estimate these properties.

The extension of the classical Hui Walter approach used in chapter 2 made it possible to estimate the efficacy of the vaccine used in the study as opposed to assuming it was known, in addition to estimating unknown properties of the two DIVA tests. It was demonstrated that in such a situation, vaccine efficacy can be estimated with good precision, as well as estimating the sensitivity and specificity of two diagnostic tests with unknown sensitivity and specificity. This has important practical implications for the feasibility of simultaneously introducing novel tests and implementing control strategies using vaccines for which the efficacy is not fully known. As far as we are aware, this is the first study demonstrating the joint estimation of vaccine efficacy, incidence, and the sensitivity and specificity of diagnostic test. Of the 111 papers covered in the review mentioned earlier (Smeden et al., 2013), only one considered the use of latent class models in the context of vaccinated animals (Engel et al., 2008), and this study never included vaccine efficacy as a model parameter.

Of particular interest was the relationship between study size, assumptions of true DIVA test properties, and the resulting estimates of the specificity of the tests, when the specificity reached values of 99.9% and higher. An important objective of the study was to demonstrate that the DIVA tests have a sufficiently high specificity, with a threshold level of 99.85%, as this is critical for a large-scale bTB vaccination program to be cost effective (Conlan et al., 2015). The results of the simulation study indicated that demonstrating that the tests superseded the performance threshold would require either a very large study, (above 100 000 animals), or that true specificity was near perfect, with less than 1 in 100 000 truly negative animals testing positive. As an alternative, the results indicated that strengthening the main study with a pilot study that made use of reference animals that are verified positive or negative to bTB a priori would reduce the required sample size. In light of these results, a joint field trial of a bTB vaccine together with DIVA tests as recommended in AHAW (2013) may allow for a good evaluation of the vaccine efficacy. However, it is unlikely to provide sufficient evidence of DIVA test performance to satisfy the 99.85% specificity threshold, and therefore other types of evidence will need to be considered to demonstrate that the test satisfies this criteria. The full results of the simulation study have been included in a report submitted to DEFRA and the Welsh Government, to be used in discussions as to whether to conduct a large-scale vaccine/DIVA trial in the UK.

In the third chapter, the use of multiple tests with non-binary results was considered, using a model that assumes that the tests have quantitative results depending on the state of infection, and that they follow a particular known trajectory after exposure to a pathogen. By considering the diagnostic tests measurement in combination with the kinetics of the tests, a "clock" was incorporated that could be used to estimate time since infection. Earlier studies have combined test values and kinetics (Baguelin et al., 2011; Nielsen et al., 2011; Simonsen et al., 2008) using a single test, and estimated the mean incidence of infection over some period of time. This chapter demonstrated that by incorporating information from more than one diagnostic test, and by considering their joint kinetic pattern, it becomes possible to relax the assumption of a constant level of incidence and estimate not only the historic mean incidence, but linearly increasing or decreasing epidemic trends. In this thesis, the the terminology of "hindcasting" was introduced for this type of analysis. In particular, the results used syntethic data generated under a wide range of conditions to demonstrate that a cross-sectional sample using two diagnostic tests could reliably be used to estimate whether the incidence trend has been increasing or decreasing. Counterintuitively, it was also discovered that the estimation of an increasing trend was more difficult than estimating a decreasing trend, implying that different levels of effort are required for monitoring the effect of policy interventions depending on their expected success.

The approach used in chapter 3 models the distribution of times since infection in a cross-sectional sample, in effect modelling the cumulative incidence. This approach is closely related to "force of infection" studies (Hens et al., 2010) which considers age-stratified prevalence data to estimate how the so-called "force of infection" varies with age. As an example of the similarity, the chapter rediscovered an equation first described by Griffiths (1974) that models the distribution of times since infection in a cross-sectional sample, assuming a previous linear trend β, an instantaneous incidence at the time of sampling α, and including possible censoring of times since infection larger than some value C:

$$p(t = T) = \frac{(\alpha + \beta T)e^{-(\alpha+\beta T/2)T}}{1 - e^{-(\alpha+\beta C/2)C}}$$

The context is different however. Griffiths modelled the exposed fraction of a population in a situation where the incidence is stable over time, but differs by age, and where the force of infection follows a linear trend as a function of age. In contrast, our equation assumes a homogenous population, but a changing incidence over time, and is implemented in a framework where the times since infections (equivalent to age in Griffiths' implementation) are inferred rather than known. Moreover if ages of individuals or other covariates can be estimated for the individuals tested in our cross-sectional data then the methods introduced here could be extended to account for them.

In chapter four, the application of hindcasting was expanded to that of epidemic diseases where reinfections can be ignored, but where the trend of incidence develops in a more dynamic fashion, e.g. rising and then falling. By considering epidemics, the study can be compared with recent papers that focus on recovering transmission networks with a combination of epidemiological and phylogenetic data. The recent study by Jombart et al. (2014) combined genetic data with case counts to estimate unobserved epidemic parameters of the 2003 SARS outbreak. In a similar fashion, Biek et al. (2012) used the number of genetic mutations between cases to reconstruct the network of spread between cattle herds and badger groups. These studies also recover epidemic patterns from cross-sectional data, but rely on genetic mutation rates as "clocks". Since genetic mutation rates are measured in transmission events, such phylogenetic models can describe epidemics on a time scale of infected generations, but have difficulty with estimating the absolute time between events unless the mutation rates are very high. By using a test kinetic based approach in the hindcasting model we measure the absolute time since infection, which is more suited to describing the overall population-level development of the epidemic. Since test kinetics can be measured on a time scale of days or weeks, it should also be better suited to estimate rapidly developing early phase of an epidemic than phylogenetic models where the genetic mutation rates are measured on the order of weeks or months for viruses (Duffy et al., 2008) and much slower for bacteria.

The content of chapter 4 is a reproduction of the paper "Recovering epidemic trends from cross-sectional data using multiple diagnostic tests" (submitted to PloS Computational Biology on the 18th of June 2015). The paper is based on two case studies; a 2007 Bluetongue outbreak in Cattle in the UK, and a *B. Pertussis* outbreak in humans in Wisconsin, USA. In the chapter it is demonstrated that epidemic trends can be fitted by using a single log-normal parametrization of the trend for both epidemic trends in an exponentially increasing phase as well as epidemic trends past their peak. It also demonstrated that in addition to the trend, the approximate length of time since the start of the outbreak can be estimated, information that would be of benefit to outbreak investigations. The statistical framework described in the paper can be applied to estimate both complex trends and time since introduction from cross-sectional data *provided* that two or more sufficiently different diagnostic tests are used to measure the state of disease in the host. The paper also contains an indepth analysis of how different diagnostic tests combine to add information on individual-level time since infection, highlighting in particular the problem of non-identifiability. The analysis indicate that when combining multiple tests, the tests should be chosen so as to measure aspects of the host-pathogen interaction that develop over different time scales, as opposed to measuring the same process.

The implementation of the epidemic hindcasting model in Chapter 4 utilizes a novel parametrization of the lognormal distribution describing the trend, which is described in the included supplementary information. This was motivated by the fact that in the framework utilized in the chapter, the overall likelihood of times since infection given observed test data often have multiple modes. Even if one mode had the dominant mass, the multi modality can lead to the MCMC algorithm getting stuck around the local maximum. In the case of the hindcasting model, for many types of kinetic curves, the time since infection that would maximise the likelihood for an observed data point can be either before or after a peak in the test kinetics, with a range of time since infection in between (where the test kinetic peaks) that would result in low likelihood values. If the time of infection is parametrized with a single parameter, it can then be difficult for the sampler to go from a sample of the posterior with recent time-of-infection parameter value for an individual to sample with the pre-peak, or distant time of infection for the same individual. In order to make the overall posterior computationally tractable, the log-normal distribution was parametrized in terms of a binary variable signifying upper or lower tail and a positive variable measuring distance in number of standard deviations away from the mean, and the mean and standard deviation of the distribution:

$$Lognorm(\mu, \Delta) =$$

$$\begin{cases} e^{\mu + \Delta S \sigma}, \\ S \in \{-1, 1\} \\ \Delta \sim N_{standard}(\Delta) I[\Delta \geq 0] \end{cases}$$

Using this parameterization, the MCMC sampler can jump between either tail of the lognormal distribution by switching the value of S, leading to better mixing than the standard formulation, and avoiding either tail being isolated from the other in the sampling space. So-called mode-hopping MCMC algorithms that can jump between different modes of the posterior are well described in the literature (Behrens, 2008; Sminchisescu et al., 2003; Tjelmeland and Hegstad, 2001), but tend to focus on general solutions that are implemented in the jumping kernel. The solution describe above is simple and straightforward to implement in any MCMC-engine and can be easily generalised to any situation where modes of the posterior are likely to lie in either tail of a symmetric distribution.

Together, chapters three and four demonstrated the importance of detailed quantitative diagnostic test kinetics for extracting maximum information from infectious disease studies. Since these methods require us to have knowledge of the test kinetics, access to studies generating such information are crucial. Some of the studies on test kinetics found in the literature use highly sophisticated models of within host dynamics (Baccam et al., 2006; Hancioglu et al., 2007; Heffernan and Keeling, 2008; Mallet et al., 2013). These typically require a large number of parameters (Heffernan and Keeling use 19 parameters), which are fitted by assigning values or priors from the literature and expert opinion. This type of approach generates detailed predictions of the entire dynamic of the host pathogen system, but is only feasible for systems such as Measles-humans which have been very intensely studied. On the other end of the spectra, the experimental infection studies found tend to only report measured values without the use of any statistical inference framework (Cray and Moon, 1995; Hoffman et al., 2006; Komar et al., 2003; López-Olvera et al., 2010). There is thus a need for approaches that can be used to generate parametric descriptions of test kinetics in non-experimental settings or for emerging diseases. A group of studies focusing on the use of seroincidence models (Graaf et al., 2014; Simonsen et al., 2009; Teunis et al., 2012; Versteegh et al., 2005) strike an appealing balance between statistical sophistication and use of observed longitudinal data. Chapter five explores how the approach used in these studies could be taken further by replicating the study by (Simonsen et al., 2009), detailing the potential benefit of using two or more diagnostic tests, and showing how it would be possible to increase the realism of the model used by assuming a piecewise linear antigen curve. The chapter also discusses additional research that needs to be done in this field to strengthen diagnostic test usage, and as a result thereof strengthen disease surveillance.

6.2 FUTURE DIRECTIONS

The results in this thesis have highlighted the importance of considering the diagnostic tests used for disease surveillance and epidemiological studies, and in particular the use of multiple diagnostic tests. They have also shown the kind of unexpected and useful conclusions that can be drawn from data collected with more than one test. However, there are a number of different research questions that have yet to be pursued owing to lack of time and resources.

In the Hui-Walter work conducted in chapter 2, it was assumed that the results of the two DIVA tests were mutually independent. This is a strong assumption that can be questioned. An extension of the described work could thus be to consider the issue of dependence. Since introducing covariance structures in a two-test Hui Walter framework lead to a non-identifiable system (Georgiadis et al., 2003), this would imply either considering the situation with three tests with different properties, or to consider tests that produce a quantitative rather than binary results. Quantitative type tests could also be used to consider the issue of disease progression. The presented analysis intentionally ignored this issue, but it is well known that the progression of bTB in the host has a strong impact on the ability to identify the pathogen (Whelan et al., 2010) , whether by immune response or by pathology. Assuming either a latent quantitative indicator variable, or an explicitly quantitative test would make it possible to model this dependency.

The hindcasting framework introduced and developed in this thesis shows a lot of promise in endemic settings for tracking trends in populations where ongoing surveillance is lacking. Chapter 3 demonstrated the example of linear trends, but there is a need to evaluate both the robustness of this modelling approach when assumptions of linearly changing incidences are not met. Non-linear trends could be considered as well, such as exponential trends, or even fitting smoothly varying trends with the use of splines or lasso. However, there are a number of complications that need to be considered when generalising hindcasting in endemic settings. If the incidence is high enough, the issue of reinfections of individuals needs to be considered. The assumption used in the chapter was that the test kinetics of a reinfected individual followed the same pattern as a naive individual infected for the first time. This is a simplification, and a more nuanced approach would be needed for pathogens where infection confer a substantial level of immunity. For many diseases, reinfections are an important aspect of the epidemiology; in particular for diseases where behaviour is an important risk factor, such as Chlamydia (Edgardh et al., 2009). When generalising the type of trend considered, a statistical complication is that the approach used for hindcasting in this thesis requires that the distribution of ″times since infection″ at the time of sampling is modelled. The "times since infection" distribution is the integral of the continuous-time trend expression, and thus have a substantially more complicated parametrisation than is commonly used in regular curve fittings. Parameterising the models used for hindcasting in a different way may circumvent this issue, but it is unclear what other approach could be used. A related issue is that the current approach requires that there is a limit to the time span for which hindcasting is applied, or that the incidence goes down to zero at some point in the past. This is not a fully satisfactory approach, and finding a different way of considering this issue would be beneficial. The use of hindcasting in epidemic settings, as described in chapter 4, could prove important for outbreak response, in particular for outbreaks where detecting is delayed. The work described in the chapter used a lognormal parametrisation which seems to be capable of capturing the essence of even quite irregular epidemics. However, here as for endemic settings, more complex parametrisations could be considered. One interesting approach would be to model the disease dynamics explicitly, by the use of SIR type models, as this might allow for estimating pathogen characteristics such as R_0 in addition to estimating the epidemic trend. As SIR models are analytically intractable this would require the solving of a system of ODEs. Unfortunately at present none of the generalist Bayesian

MCMC packages have ODE solvers that are sufficiently capable, and so it would require coding the sampler in a general programming language.

The hindcasting approach could be extended even further by incorporating additional aspects of the disease system such as population demography (e.g. age- and sex-structure), contact networks, or spatial location. This might enable the estimation of disease dynamics not only in time but also in spatial spread, or of the pattern of spread through the demographic structure of the population. In Birrell et al. (2011), a number of data sources is combined to estimate the dynamics of the H1N1 influenza pandemic in London during 2009, including age-structured patterns of disease spread. This is done using observed binary data, and thus cannot estimate time since infection for observed cases. Similar types of evidence synthesis using quantitative diagnostic measures combined with a hindcasting type approach would have the potential for even more in-depth understanding of the disease dynamics.

The approach used in both chapter 4 and 5 for describing test kinetics is intentionally simplistic. By only using one parameter to describe the combined effect of measurement error and individual variation in the trajectory, inference in the MCMC framework is simplified, allowing the chapters to focus on highlighting the potential of hindcasting. It might well be that using a one-parameter description of error is enough to recover trends even in more realistic settings, as the hindcasting procedure has been surprisingly robust to even high levels of noise, at least in an epidemic setting where the overall shape of the curve is more important than producing estimates of e.g trend or location parameters. However, it is likely that when this approach is used on field data, it would benefit from modeling the various influences on test measurements separately: from individual-level variation in the underlying process being measured, via contamination and preservation issues, to noise introduced in the laboratory process of sample preparation and measurement. In order to model the individual-level variation, the lookup-table approach used in chapters 3 and 4 would have to be replaced with parametric hierarchical descriptions of the kinetics, where the parameters for each individual come from population-level probability distributions.

Modelling the test kinetics in a more complex manner requires that there are data available on the kinetics of the test, of sufficient quality to support such models. The proposal in chapter 5 to recover the test kinetics from observational data would be one way to make it more feasible to collect such data. The work described in the chapter is just a very first step, and there are a number of improvements that could be implemented. Instead of assuming that the time of infection reported is correct, one could incorporate the incubation time into the model as an unobserved variable. Another improvement concerns the parametrisation of the kinetic curve: in Simonsen et al. (2009) it is assumed that the pathogen level declines linearly from a high starting point, which is clearly a strong simplification. As an alternative chapter 5 suggested using a piecewise linear pathogen curve. Another option would be to model the interaction of immune response and pathogen levels directly as a pair of differential equations instead of attempting to find explicit expressions for the curves. Making use of this would allow for much more realistic descriptions of the test kinetics.

The hindcasting models developed in this thesis incorporate a "clock", that translates observed data to a time since infection via knowledge of diagnostic test levels as a function of time since infection. This clock could be constructed in other ways as well. As mentioned above, in phylogenetic analyses, the rate of mutation of a pathogen is known, and this rate of mutation combined with genetic distance translates into a time since the last common ”ancestor pathogen”. Hindcasting could be generalized by combining phylogenetic data, quantitative diagnostic test data, and test kinetics. Potential other ”clocks” exist that could also be used for the purpose of hindcasting. On the farm level, the diffusion pattern of pathogens into the environment means that the distance from e.g. a farm that a pathogen was detected, possibly combined with knowledge of past weather and rain volume, could be used to determine how long the pathogen had been diffusing from the farm, and thus the time since the farm was first infected. A common practice in animal disease surveillance is the use of pooled samples, such as sampling from the bulk milk tank at a dairy cattle farm rather than from individual animals, or combining blood samples from multiple individuals. In such pooled samples, the direct relationship between time and indicators becomes diffuse; however, the relative prevalence of different disease indicators would still reflect aspects of the disease dynamic of the sampled population, providing information that could be utilized in a hindcasting type framework. Looking at the individual level, clinical descriptions of stages of disease could be used as ordinal measurements of time since infection. Such measurements might not be precise enough in themselves, but when combined with one or more of the other clocks mentioned here, could help decrease uncertainty and consequently increase the accuracy of estimated population level trends.

6.3 policy implications

Current systems of disease surveillance are often limited in terms of timeliness and completeness. The quality and existence of disease surveillance is particularly limited in developing countries (Butler, 2006; US General Accounting Office, 2001), in wildlife (Mörner et al., 2002), and for pathogens that are not currently considered high risk (US General Accounting Office, 2001). The risk posed by infectious diseases is increasing (Jones et al., 2008), due to climate change (Fox et al., 2011; McMichael et al., 2006), increasing air traffic volume which increases the risk that a local epidemic becomes a global problem (Lipkin, 2013), and a number of other factors. As a consequence of this, the capacity and flexibility of disease surveillance needs to improve. Because of the costs involved in establishing surveillance systems, improvements in the use of existing systems and the data they produce are likely to be most cost effective.

DEFRA (2013), provides a series of suggested objectives for improvement in the animal disease surveillance sector. The policy objectives include: timeliness/ early identification; risk assessment; trend analysis and detection of change in endemic diseases; effective surveillance data gathering and analysis; and development of improved surveillance methodology.

This thesis has presented a number of results that can be used to reach objectives such as these by making better use of existing surveillance systems, extracting more information from cross-sectional data than was previously thought possible, and supporting the decision making of epidemiologists and policy makers. The extent of the time delay between an infection event, via its detection, to it being registered by the overall disease surveillance system, plays an especially important role during outbreaks of disease or when a novel pathogen has been introduced. While multiple testing in itself will not improve the time from infection to registration, hindcasting the times of infection before the point of detection could help mitigate the effects of delayed detection by allowing one to act as if the disease had been tracked all along. In this way, it can also help improve the risk assessment of emerging diseases once the disease has been identified, by providing information of its previous trajectory and rate of spread, thus informing predictions about future trajectories and rates of spread. This in turn would help to inform a more proportionate response in terms of disease control effort.

In situations where a novel pathogen has been discovered, two of the most pressing questions are how long the pathogen has been present, and if it is on the increase. A current example is the recent discovery of a leprosy-like disease in Scotland among squirrels (Meredith et al., 2014). Being able to sample across the squirrel population and establish whether it has been spreading over the last few years, or if it has already reached an endemic steady state, would be highly relevant for helping decision-makers decide on what actions to take.

In the context of endemic diseases, one of the most important objectives for assessing the burden of disease and evaluating the impact of policy measures, is to estimate and track trends of incidence. As suggested by the results in chapter 3, cross-sectional studies could be used to estimate trends and detect changes in incidence over multi-year periods. In settings where ongoing surveillance systems exists, the results could be used to validate data and trends as estimated by routine surveillance. In addition, it would enable the tracking of trends in populations for pathogens where routine surveillance is *not* conducted and only occasional cross-sectional studies are implemented.

Because of the high cost of surveillance systems, it is important to utilise available effort and collected data as efficiently as possible. In DEFRA (2013), it is stated that "Development and use of new or improved methodologies for collating, interpreting and analysing such diverse data is likely to be needed to aid in the implementation of improved surveillance methodologies". With the rapid development of novel diagnostic methods, it becomes possible to conduct statistical analyses that exploit the additional information inherent in these new diagnostic tests. As the results in this thesis demonstrate, by taking the quantitative nature of diagnostic tests test into account, and by combining multiple tests, there is a substantial gain in statistical power that can be used to estimate critical aspects of the epidemiology of disease.

Thus, increasing the collection and usage of quantitative information at all levels of the surveillance system should be a priority. Often, what is registered in databases as case/non-case is the result of applying a cutoff to a quantitative result of a diagnostic test. Collecting the underlying quantitative test data as well as the cut-off classification would be a good first step that could likely be taken without a great increase in expenditures. Switching from binary tests in situations and for diseases where such are used, to tests with quantitative information, and ideally with known test kinetics, would be a benefit, both at the point-of-care or point-of-diagnosis in terms of modern tests with better performance and faster turn-around, as well as to epidemiologists and statisticians working with the collected data to understand population level patterns. With quantitative data enabling richer analysis of the dynamics and state of disease in a population, this would in turn provide a stronger evidence base from which to conduct research, manage the effects of disease, and decide on policy measures to reduce the burden of disease.

6.4 CONCLUSION

The research put forward in this thesis highlights the importance of considering the choice of diagnostic tests not only from a cost perspective but from the perspective of gaining as much information as possible from collected data. Despite the additional cost, using two or more diagnostic tests is from this perspective a tremendous gain in terms of the kinds of analyses it enables, and the information these can provide.

This thesis has shown the rich possibilities for analyses that open up with the advent of quantitative diagnostic test data. At all levels, there needs to be an understanding of the additional information that can be derived from such analyses; those conducting tests and filling out forms; those designing databases for the collection of surveillance data; those analyzing the data; those implementing the surveillance systems; and those responsible for using these data to inform their decisions on management of infectious diseases.

In summary, multiple quantitative tests can measure the progression from a naive to an infected individual, and recover disease trends in a population. Embracing their use will allow us to shed light on the past to inform our actions in the future!

BIBLIOGRAPHY

AHAW, 2013. Scientific Opinion on field trials for bovine tuberculosis vaccination 1. EFSA Journal 11, 1–35. doi10.2903/j.efsa.2013.3475

Alban, L., Verheyen, K., Martinez, T., Velthuis, A., Saatkamp, H., Mourits, M., Koeijer, A. de, Elbers, A., 2010. Financial consequences of the Dutch bluetongue serotype 8 epidemics of 2006 and 2007. Preventive Veterinary Medicine 93, 294–304.

Ali, M., Emch, M., Donnay, J.P., Yunus, M., Sack, R.B., 2002. The spatial epidemiology of cholera in an endemic area of Bangladesh. Social Science and Medicine 55, 1015–1024. doi10.1016/S0277-9536(01)00230-1

Andersson, E., Kühlmann-Berenzon, S., Linde, A., Schiöler, L., Rubinova, S., Frisén, M., 2008. Predictions by early indicators of the time and height of the peaks of yearly influenza outbreaks in Sweden. Scandinavian Journal of Public Health 36, 475–82. doi10.1177/1403494808089566

Armitage, P., Berry, G., Matthews, J.N.S., 2008. Statistical methods in medical research. John Wiley & Sons.

Asmussen, S., Glynn, P.W., 2011. A new proof of convergence of MCMC via the ergodic theorem. Statistics and Probability Letters 81, 1482–1485. doi10.1016/j.spl.2011.05.004

Baccam, P., Beauchemin, C., Macken, C. a, Hayden, F.G., Perelson, A.S., 2006. Kinetics of influenza A virus infection in humans. Journal of Virology 80, 7590–9. doi10.1128/JVI.01623-05

Baguelin, M., Hoschler, K., Stanford, E., Waight, P., Hardelid, P., Andrews, N., Miller, E., 2011. Age-specific incidence of A/H1N1 2009 influenza infection in England from sequential antibody prevalence data using likelihood-based estimation. PloS One 6, e17074. doi10.1371/journal.pone.0017074

Banakar, V., Constantin de Magny, G., Jacobs, J., Murtugudde, R., Huq, A., Wood, R.J., Colwell, R.R., 2011. Temporal and spatial variability in the distribution of Vibrio vulnificus in the Chesapeake Bay: a hindcast study. EcoHealth 8, 456–67. doi10.1007/s10393-011-0736-4

Behrens, G.R., 2008. Mode jumping in MCMC. Scandinavian Journal of Statistics 28, 205–223.

Berkelman, R.L., Bryan, R.T., Osterholm, M.T., LeDuc, J.W., Hughes, J.M., 1994. Infectious disease surveillance: a crumbling foundation. Science 264, 368–370.

Bernard, K.A. et al., 2001. West Nile virus infection in birds and mosquitoes, New York State, 2000. Emerging Infectious Diseases 7, 679–85. doi10.3201/eid0704.010415

Bidet, P. et al., 2008. Real-time PCR measurement of persistence of Bordetella pertussis DNA in nasopharyngeal secretions during antibiotic treatment of young children with pertussis. Journal of Clinical Microbiology 46, 3636–8. doi10.1128/JCM.01308-08

Biek, R. et al., 2012. Whole genome sequencing reveals local transmission patterns of Mycobacterium bovis in sympatric cattle and badger populations. PLoS Pathogens 8, e1003008. doi10.1371/journal.ppat.1003008

Birrell, P.J. et al., 2011. Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. Proceedings of the National Academy of Sciences of the United States of America 108, 18238–43. doi10.1073/pnas.1103002108

Boven, M. van, Ferguson, N.M., Rie, A. van, 2004. Unveiling the burden of pertussis. Trends in Microbiology 12, 116–9. doi10.1016/j.tim.2004.01.002

Bronsvoort, B.M.D.C., Radford, A.D., Tanya, V.N., Nfon, C., Kitching, R.P., Morgan, K.L., 2004. Epidemiology of Foot-and-Mouth Disease Viruses in the Adamawa Province of Cameroon. Journal of Clinical Microbiology 42. doi10.1128/JCM.42.5.2186

Brooks, S.P., Gelman, A., 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics 7, 434–455.

Buehler, J.W., Berkelman, R.L., Hartley, D.M., Peters, C.J., 2003. Syndromic surveillance and bioterrorism-related epidemics. Emerging Infectious Diseases 9, 1197–1204.

Butler, D., 2006. Disease surveillance needs a revolution. Nature 440, 6–7. doi10.1038/440006a

Caboche, S., Audebert, C., Hot, D., 2014. High-Throughput Sequencing, a VersatileWeapon to Support Genome-Based Diagnosis in Infectious Diseases: Applications to Clinical Bacteriology. Pathogens 3, 258–279. doi10.3390/pathogens3020258

Canini, L., Perelson, A.S., 2014. Viral kinetic modeling: state of the art. Journal of Pharmacokinetics and Pharmacodynamics 431–443. doi10.1007/s10928-014-9363-3

Carrat, F., Vergu, E., Ferguson, N.M., Lemaitre, M., Cauchemez, S., Leach, S., Valleron, A.-J., 2008. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. American Journal of Epidemiology 167, 775–85. doi10.1093/aje/kwm375

Casadevall, A., Pirofski, L.-a., 2000. Host-Pathogen Interactions: Basic Concepts of Microbial Commensalism, Colonization, Infection, and Disease. Infection and Immunity 68, 6511–6518. doi10.1128/IAI.68.12.6511-6518.2000

Casadevall, A., Pirofski, L.-a.L.-a., 2001. Host-pathogen interactions: the attributes of virulence. The Journal of Infectious Diseases 184, 337–44. doi10.1086/322044

Casella, G., George, E.I., 1992. Explaining the Gibbs Sampler. The American Statistician 46, 167–174. doi10.1080/00031305.1992.10475878

CDC, 2001. Updated Guidelines for Evaluating Public Health Surveillance Systems. MMWR. Morbidity and Mortality Weekly Report 50.

Chan, E.H. et al., 2010. Global capacity for emerging infectious disease detection. Proceedings of the National Academy of Sciences of the United States of America 107, 21701–21706. doi10.1073/pnas.1006219107

Chan, M., 2009. World now at the start of 2009 influenza pandemic (WHO statement to the press). Http://www.who.int/mediacentre/news/statements/2009/.

Chatzinasiou, E., Dovas, C.I., Papanastassopoulou, M., Georgiadis, M., Psychas, V., Bouzalas, I., Koumbati, M., Koptopoulos, G., Papadopoulos, O., 2010. Assessment of bluetongue viraemia in sheep by real-time PCR and correlation with viral infectivity. Journal of Virological Methods 169, 305–15. doi10.1016/j.jviromet.2010.07.033

Ciupe, S.M., Ribeiro, R.M., Nelson, P.W., Perelson, A.S., 2007. Modeling the mechanisms of acute hepatitis B virus infection. Journal of Theoretical Biology 247, 23–35. doi10.1016/j.jtbi.2007.02.017

Clegg, T.A., Duignan, A., Whelan, C., Gormley, E., Good, M., Clarke, J., Toft, N., More, S.J., 2011. Using latent class analysis to estimate the test characteristics of the γ-interferon test, the single intradermal comparative tuberculin test and a multiplex immunoassay under

Irish conditions. Veterinary Microbiology 151, 68–76. doi10.1016/j.vetmic.2011.02.027

Cleveland, W.S., Devlin, S.J., 1988. Locally Weighted Regression : An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association 83, 596–610.

Communications, B., 2006. HIV-1 and HCV sequences from Libyan outbreak. Nature 444, 836–837. doi10.1038/nature444836a

Conlan, A.J.K., Brooks Pollock, E., McKinley, T.J., Mitchell, A.P., Jones, G.J., Vordermeier, M., Wood, J.L.N., 2015. Potential Benefits of Cattle Vaccination as a Supplementary Control for Bovine Tuberculosis. PLOS Computational Biology 11, e1004038. doi10.1371/journal.pcbi.1004038

Cray, W., Moon, H., 1995. Experimental infection of calves and adult cattle with Escherichia coli O157:H7. Appl. Envir. Microbiol. 61, 1586–1590.

Daszak, P., Cunningham, A.A., Hyatt, A.D., 2000. Emerging Infectious Diseases of Wildlife– Threats to Biodiversity and Human Health. Science 287, 443–449. doi10.1126/science.287.5452.443

Declich, S., Carter, A.O., 1994. Public health surveillance : historical origins, methods and evaluation. Bulletin of the World Health Organization 72, 285–304.

DEFRA, 2008. Report on the distribution of Bluetongue infection in Great Britain on 15 March 2008. Nobel House, 17 Smith Square, London, SW1P 3JR, United Kingdom.

DEFRA, 2013. New and re-emerging diseases , endemic diseases and enhanced surveillance methodology Evidence Plan PB13920.

DEFRA, 2014. The Strategy for achieving Officially Bovine Tuberculosis Free status for England PB 14088.

Del Rio Vilas, V.J., Pfeiffer, D.U., 2010. The evaluation of bias in scrapie surveillance: a review. The Veterinary Journal 185, 259–64. doi10.1016/j.tvjl.2009.06.014

Domingo, M., Vidal, E., Marco, a., 2014. Pathology of bovine tuberculosis. Research in Veterinary Science 97, S20–S29. doi10.1016/j.rvsc.2014.03.017

Dórea, F.C., Sanchez, J., Revie, C.W., 2011. Veterinary syndromic surveillance: Current initiatives and potential for development. Preventive Veterinary Medicine 101, 1–17. doi10.1016/j.prevetmed.2011.05.004

Duffy, S., Shackelton, L. a, Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews. Genetics 9, 267–276. doi10.1038/nrg2323

Durodié, B., 2011. H1N1 – the social costs of élite confusion. Journal of Risk Research 14, 511–518. doi10.1080/13669877.2011.576767

Echevarría-Zuno, S. et al., 2009. Infection and death from influenza A H1N1 virus in Mexico: a retrospective analysis. Lancet 374, 2072–9. doi10.1016/S0140-6736(09)61638-X

Edgardh, K., Kühlmann-Berenzon, S., Grünewald, M., Rotzen-Ostlund, M., Qvarnström, I., Everljung, J., 2009. Repeat infection with Chlamydia trachomatis: a prospective cohort study from an STI-clinic in Stockholm. BMC Public Health 9, 198. doi10.1186/1471-2458-9-198

EEC, 1977. Council Directive 78/52/EEC. Official Journal of The European Communities 1977, 34–41.

Engel, B., Buist, W., Orsel, K., Dekker, A., Clercq, K. de, Grazioli, S., Roermund, H. van, 2008. A Bayesian evaluation of six diagnostic tests for foot-and-mouth disease for vaccinated and non-vaccinated cattle. Preventive Veterinary Medicine 86, 124–138. doi10.1016/j.prevetmed.2008.03.009

FAO, 2013. Declaration of global freedom from rinderpest – Thirty-seventh Session of the FAO Conference, Rome 25 June-2 July 2011, in: FAO Animal Production and Health Proceedings No. 17. FAO. Rome, Italy.

Farrington, C.P., Andrews, N.J., Beale, a D., Catchpole, M. a, 1996. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. Journal of the Royal Statistical Society. Series A 159, 547–563. doi10.2307/2983331

Fauci, a S., 2001. Infectious diseases: considerations for the 21st century. Clinical Infectious Diseases 32, 675–685. doi10.1086/319235

Fearnhead, P., 2006. Exact and efficient Bayesian inference for multiple changepoint problems. Statistics and Computing 16, 203–213.

Fine, P.E.M., 1993. Herd Immunity : History, Theory, Practice. Epidemiologic Reviews 15.

Fineberg, H.V., 2014. Pandemic preparedness and response–lessons from the H1N1 influenza of 2009. The New England Journal of Medicine 370, 1335–42. doi10.1056/NEJMra1208802

Flynn, P., 2010. The handling of the H1N1 pandemic : more transparency needed. Council of Europe Parliamentary Assembly.

Fox, N.J., White, P.C.L., McClean, C.J., Marion, G., Evans, A., Hutchings, M.R., 2011. Predicting Impacts of Climate Change on Fasciola hepatica Risk. PLoS ONE 6, 9.

Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S., 2008. HealthMap : Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. Journal of the American Medical Informatics Association 15. doi10.1197/jamia.M2544.Introduction

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 515–533.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, Second Edition (Texts in Statistical Science).

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics 2, 1360–1383. doi10.1214/08-AOAS191

Gelman, A., Rubin, D.B., 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7, 457–472.

Geman, S., Geman, D., 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6, 721–741. doi10.1109/TPAMI.1984.4767596

Georgiadis, M.P., Johnson, W.O., Gardner, I.a., Singh, R., 2003. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. Journal of the Royal Statistical Society: Series C (Applied Statistics) 52, 63–76. doi10.1111/1467-9876.00389

Gibbons, C.L. et al., 2014. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health 14, 147. doi10.1186/1471-2458-14-147

Gilbert, M., Mitchell, a, Bourn, D., Mawdsley, J., Clifton-Hadley, R., Wint, W., 2005. Cattle movements and bovine tuberculosis in Great Britain. Nature 435, 491–6. doi10.1038/nature03548

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–4. doi10.1038/nature07634

Giorgi, E.E., Funkhouser, B., Athreya, G., Perelson, A.S., Korber, B.T., Bhattacharya, T., 2010. Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. BMC Bioinformatics 11, 532. doi10.1186/1471-2105-11-532

Girard, M.P., Tam, J.S., Assossou, O.M., Kieny, M.P., 2010. The 2009 A (H1N1) influenza virus pandemic: A review. Vaccine 28, 4895–902. doi10.1016/j.vaccine.2010.05.031

Godfray, H.C.J. et al., 2013. A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain. Proceedings of the Royal Society B 280. doi10.1098/rspb.2013.1634

Götz, H.M., 2005. Screening for Chlamydia trachomatis: Whom and How ? (PhD thesis). University Medical Center Rotterdam.

Graaf, W. de, Kretzschmar, M., Teunis, P., Diekmann, O., 2014. A two-phase within-host model for immune response and its application to serological profiles of pertussis. Epidemics 9, 1–7. doi10.1016/j.epidem.2014.08.002

Greeff, S.C. de, Melker, H.E. de, Gageldonk, P.G.M. van, Schellekens, J.F.P., Klis, F.R.M. van der, Mollema, L., Mooi, F.R., Berbers, G. a M., 2010. Seroprevalence of pertussis in The Netherlands: evidence for increased circulation of Bordetella pertussis. PloS One 5, e14183. doi10.1371/journal.pone.0014183

Greiner, M., Gardner, I.A., 2000a. Epidemiologic issues in the validation of veterinary diagnostic tests. Preventive Veterinary Medicine 45, 3–22. doi10.1016/S0167-5877(00)00114-8

Greiner, M., Gardner, I.A., 2000b. Application of diagnostic tests in veterinary epidemiologic studies. Preventive Veterinary Medicine 45, 43–59.

Grenfell, B.T., Bjørnstad, O.N., Kappey, J., 2001. Travelling waves and spatial hierarchies in measles epidemics. Nature 414, 716–723. doi10.1038/414716a

Griffiths, D. a, 1974. A Catalytic Model of Infection for Measles. Journal of the Royal Statistical Society. Series C (Applied Statistics) 23, 330–339. doi10.2307/2347126

Hallander, H.O., Andersson, M., Gustafsson, L., Ljungman, M., Netterlid, E., 2009. Seroprevalence of pertussis antitoxin (anti-PT) in Sweden before and 10 years after the introduction of a universal childhood pertussis vaccination program. APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica 117, 912–22. doi10.1111/j.1600-0463.2009.02554.x

Hallander, H.O., Gustafsson, L., Ljungman, M., Storsaeter, J., 2005. Pertussis antitoxin decay after vaccination with DTPa. Response to a first booster dose 3 1/2-6 1/2 years after the third vaccine dose. Vaccine 23, 5359–64. doi10.1016/j.vaccine.2005.06.009

Hammes, F., Egli, T., 2010. Cytometric methods for measuring bacteria in water: Advantages, pitfalls and applications. Analytical and Bioanalytical Chemistry 397, 1083–1095. doi10.1007/s00216-010-3646-3

Hancioglu, B., Swigon, D., Clermont, G., 2007. A dynamical model of human immune response to influenza A virus infection. Journal of Theoretical Biology 246, 70–86. doi10.1016/j.jtbi.2006.12.015

Hanson, K.M., 2001. Markov Chain Monte Carlo posterior sampling with the Hamiltonian method, in: Proc. SPIE 4322, Medical Imaging 2001: Image Processing. pp. 456–467. doi10.1117/12.431119

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Haydon, D.T., Chase-Topping, M., Shaw, D.J., Matthews, L., Friar, J.K., Wilesmith, J., Woolhouse, M.E.J., 2003. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proceedings. Biological Sciences / The Royal Society 270, 121–7. doi10.1098/rspb.2002.2191

Heffernan, J.M., Keeling, M.J., 2008. An in-host model of acute infection: Measles as a case study. Theoretical Population Biology 73, 134–147. doi10.1016/j.tpb.2007.10.003

Heffernan, J.M., Smith, R.J., Wahl, L.M., 2005. Perspectives on the basic reproductive ratio. Journal of the Royal Society, Interface / the Royal Society 2, 281–293. doi10.1098/rsif.2005.0042

Hempel, S., 2013. John Snow. Lancet 381, 1269–1270. doi10.1016/S0140-6736(13)60830-2

Hens, N., Aerts, M., Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P., Beutels, P., 2010. Seventy-five years of estimating the force of infection from current status data. Epidemiology and Infection 138, 802–812. doi10.1017/S0950268809990781

Herrmann, B., Törner, A., Low, N., Klint, M., Nilsson, A., Velicko, I., Söderblom, T., Blaxhult, A., 2008. Emergence and spread of Chlamydia trachomatis variant, Sweden. Emerging Infectious Diseases 14, 1462–1465. doi10.3201/eid1409.080153

Hethcote, H.W., 2007. The Mathematics of Infectious Diseases. SIAM Review 42, 599–653.

Hoffman, M. a, Menge, C., Casey, T. a, Laegreid, W., Bosworth, B.T., Dean-Nystrom, E. a, 2006. Bovine immune response to shiga-toxigenic Escherichia coli O157:H7. Clinical and Vaccine Immunology : CVI 13, 1322–7. doi10.1128/CVI.00205-06

Hoffman, M.D., Gelman, A., 2014. The No-U-Turn Sampler : Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15, 1351–1381.

Hogan, R.J., Mathews, S.a., Mukhopadhyay, S., Summersgill, J.T., Timms, P., 2004. Chlamydial Persistence: Beyond the Biphasic Paradigm. Infection and Immunity 72, 1843–1855. doi10.1128/IAI.72.4.1843-1855.2004

Höhle, M., An der Heiden, M., 2014. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. Biometrics. doi10.1111/biom.12194

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–71.

Hulth, A., Rydevik, G., Linde, A., 2009. Web Queries as a Source for Syndromic Surveillance. PLoS ONE 4, 10.

Isakbaeva, E.T. et al., 2005. Norovirus Transmission on Cruise Ship. Emerging Infectious Diseases 11, 2003–2006.

Johnson, W.O., Gastwirth, J.L., Pearson, L.M., 2001. Screening without a "gold standard": the Hui-Walter paradigm revisited. American Journal of Epidemiology 153, 921–4.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., Ferguson, N., 2014. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Computational Biology 10, e1003457. doi10.1371/journal.pcbi.1003457

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990–3. doi10.1038/nature06536

Juve, G., Deelman, E., Vahi, K., Mehta, G., Berriman, B., Berman, B.P., Maechling, P., 2009. Scientific Workflow Applications on Amazon EC2–includesz HPC and cloud, in: 5th IEEE International Conference on E-Science. pp. 59–66.

Kao, R.R., 2002. The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. Trends in Microbiology 10, 279–286. doi10.1016/S0966-842X(02)02371-5

Kathryn, M., Carlin, B.P., Cowles, M.K., 1996. Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review. Journal of the American Statistical Association 91, 883–904.

Kaufmann, S.H.E., Schaible, U.E., 2005. 100th anniversary of Robert Koch's Nobel Prize for the discovery of the tubercle bacillus. Trends in Microbiology 13, 469–75. doi10.1016/j.tim.2005.08.003

Kirschner, D.E., Linderman, J.J., 2009. Mathematical and computational approaches can complement experimental studies of host-pathogen interactions. Cellular Microbiology 11, 531–539. doi10.1111/j.1462-5822.2009.01281.x

Kleczkowski, A., Gilligan, C.A., 2007. Parameter estimation and prediction for the course of a single epidemic outbreak of a plant disease. Journal of the Royal Society, Interface / the Royal Society 4, 865–77. doi10.1098/rsif.2007.1036

Koch, A.L., 1966. The Logarithm in Biology 1. Mechanisms generating the Log-Normal Distribution exactly. Journal of Theoretical Biology 12, 276–290.

Komar, N., Langevin, S., Hinten, S., Nemeth, N., Edwards, E., Hettler, D., Davis, B., Bowen, R., Bunning, M., 2003. Experimental infection of North American birds with the New York 1999 strain of West Nile virus. Emerging Infectious Diseases 9, 311–22.

Kosek, M., Bern, C., Guerrant, R.L., 2003. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. Bulletin of the World Health Organization 81, 197–204. doiS0042-96862003000300010 [pii]

Kuiken, T., Gortázar, C., 2011. Establishing a European network for wildlife. Revue Scientifique et Technique (International Office of Epizootics) 30, 755–761.

Kulldorff, M., 1997. A spatial scan statistic. Communications on Statistical Theory and Methodology 26, 1481–1496.

Kulldorff, M., Rand, K., Gherman, G., Williams, G., DeFrancesco, D., 1998. SaTScan v 2.1: Software for the spatial and space-time scan statistics. Bethesda, MD: National Cancer Institute.

Langmuir, A.D., 1976. William Farr: Founder of Modern Concepts of Surveillance. International Journal of Epidemiology 5, 13–18. doi10.1093/ije/5.1.13

Last, J., 1963. The Iceberg: "Completing the Clinical Picture" in General Practice. The Lancet 282, 28–31.

Lessler, J., Reich, N.G., Brookmeyer, R., Perl, T.M., Nelson, K.E., Cummings, D.A., 2009. Incubation periods of acute respiratory viral infections: a systematic review. The Lancet Infectious Diseases 9, 291–300. doi10.1016/S1473-3099(09)70069-6

Leung, G.M., Nicoll, A., 2010. Reflections on pandemic (H1N1) 2009 and the international response. PLoS Medicine 7, 6. doi10.1371/journal.pmed.1000346

Lewandowski, D., Kurowicka, D., Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis 100, 1989–2001. doi10.1016/j.jmva.2009.04.008

Limpert, E., Stahel, W.a., Abbt, M., 2001. Log-normal Distributions across the Sciences: Keys and Clues. BioScience 51, 341. doi10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Lipkin, W.I., 2013. The changing face of pathogen discovery and surveillance. Nature Reviews. Microbiology 11, 133–41. doi10.1038/nrmicro2949

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J.L., 2006. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet 367, 1747–57. doi10.1016/S0140-6736(06)68770-9

López-Olvera, J.R. et al., 2010. Experimental infection of European red deer (Cervus elaphus) with bluetongue virus serotypes 1 and 8. Veterinary Microbiology 145, 148–152.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project : Evolution , critique and future directions. Statistics in Medicine 3049–3067. doi10.1002/sim

Madoff, L.C., 2004. ProMED-mail : An Early Warning System for Emerging Diseases. Clinical Infectious Diseases 02115, 227–232.

Major, M.E., Dahari, H., Mihalik, K., Puig, M., Rice, C.M., Neumann, A.U., Feinstone, S.M., 2004. Hepatitis C virus kinetics and host responses associated with disease and outcome of infection in chimpanzees. Hepatology 39, 1709–1720. doi10.1002/hep.20239

Mallet, D.G., Bagher-Oskouei, M., Farr, a C., Simpson, D.P., Sutton, K.-J., 2013. A mathematical model of chlamydial infection incorporating movement of chlamydial particles. Bulletin of Mathematical Biology 75, 2257–70. doi10.1007/s11538-013-9891-9

Matthews, L., Woolhouse, M., 2005. New approaches to quantifying the spread of infection. Nature Reviews. Microbiology 3, 529–36. doi10.1038/nrmicro1178

McMichael, A.J., Woodruff, R.E., Hales, S., 2006. Climate change and human health: present and future risks. Lancet 367, 859–69. doi10.1016/S0140-6736(06)68079-3

Mercer, G.N., Glass, K., Becker, N.G., 2011. Effective reproduction numbers are commonly overestimated early in a disease outbreak. Statistics in Medicine 30, 984–94. doi10.1002/sim.4174

Meredith, A., Del Pozo, J., Smith, S., Milne, E., Stevenson, K., McLuckie, J., 2014. Leprosy in red squirrels in Scotland. The Veterinary Record 175, 285–6. doi10.1136/vr.g5680

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21.

Morabia, A., Zhang, F.F., 2004. History of medical screening : from concepts to action. Postgrad Medical Journal 80, 463–470. doi10.1136/pgmj.2004.018226

Mörner, T., Obendorf, D.L., Artois, M., Woodford, M.H., 2002. Surveillance and monitoring of wildlife diseases. Scientific and Technical Review of the Office International Des Epizooties 21, 67–76.

Muench, H., 1934. Derivation of Rates from Summation Data by the Catalytic Curve. Journal of the American Statistical Association 29, 25–38.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., H, E., 1986. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction, in: Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory, pp. 263–273.

Nagelkerke, N.J.D., Arora, P., Jha, P., Williams, B., McKinnon, L., Vlas, S.J. de, 2014. The Rise and Fall of HIV in High-Prevalence Countries: A Challenge for Mathematical Modeling. PLoS Computational Biology 10, e1003459. doi10.1371/journal.pcbi.1003459

Nielsen, S.S., Toft, N., 2009. A review of prevalences of paratuberculosis in farmed animals in Europe. Preventive Veterinary Medicine 88, 1–14. doi10.1016/j.prevetmed.2008.07.003

Nielsen, T.D., Nielsen, L.R., Toft, N., 2011. Bayesian estimation of true between-herd and within-herd prevalence of Salmonella in Danish veal calves. Preventive Veterinary Medicine 100, 155–62. doi10.1016/j.prevetmed.2011.02.014

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. doi10.1186/1742-7622-4-2

OIE, 2009. Bovine Tuberculosis, in: Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. pp. 1–16.

OIE, 2013. Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, Chapter 1.1.5: Principles and methods of validation of diagnostic assays for infectious diseases., 2013th ed. Office International des Epizooties, Paris, France.

OIE, 2014. Paratuberculosis (Johne ' s disease), in: Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. Office international des épizooties, Paris, France.

Parashar, U.D., Hummelman, E.G., Bresee, J.S., Miller, M.a., Glass, R.I., 2003. Global illness and deaths caused by rotavirus disease in children. Emerging Infectious Diseases 9, 565–572. doi10.3201/eid0905.020562

Parliamentary Office Of Science & Technology, 2014. Surveillance of Infectious Disease. POSTNOTE 462, 1–4.

Patel, M.M., Hall, A.J., Vinjé, J., Parashar, U.D., 2009. Noroviruses: A comprehensive review. Journal of Clinical Virology 44, 1–8. doi10.1016/j.jcv.2008.10.009

Pawlotsky, J.M., 2002. Use and interpretation of virological tests for hepatitis C. Hepatology 36, 65–73. doi10.1053/jhep.2002.36815

Perez, a., AlKhamis, M., Carlsson, U., Brito, B., Carrasco-Medanic, R., Whedbee, Z., Willeberg, P., 2011. Global animal disease surveillance. Spatial and Spatio-Temporal Epidemiology 2, 135–145. doi10.1016/j.sste.2011.07.006

Plummer, M., 2003. JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling, in: Proceedings of the 3rd

International Workshop on Distributed Statistical Computing (DSC 2003).

Plummer, M., 2014. rjags: Bayesian graphical models using MCMC.

Presanis, A.M., Pebody, R.G., Paterson, B.J., Tom, B.D.M., Birrell, P.J., Charlett, A., Lipsitch, M., De Angelis, D., 2011. Changes in severity of 2009 pandemic A/H1N1 influenza in England: a Bayesian evidence synthesis. BMJ (Clinical Research Ed.) 343, d5408.

Pugliese, A., Gandolfi, A., 2008. A simple model of pathogen-immune dynamics including specific and non-specific immunity. Mathematical Biosciences 214, 73–80. doi10.1016/j.mbs.2008.04.004

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; R Foundation for Statistical Computing, Vienna, Austria.

Robert, C., Casella, G., 2011. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. Statistical Science 26, 102–115. doi10.1214/10-STS351

Robert, C.P., 2007. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd ed. Springer, New York, NY.

Rodrigue, D.C., Tauxe, R.V., Rowe, B., 1990. International increase in Salmonella enteritidis : A new pandemic ? Epidemiology and Infection 105, 21–27.

Savill, N.J., St Rose, S.G., Keeling, M.J., Woolhouse, M.E.J., 2006. Silent spread of H5N1 in vaccinated poultry. Nature 442, 757. doi10.1038/442757a

Schiller, I. et al., 2011. Bovine tuberculosis in Europe from the perspective of an officially tuberculosis free country: trade, surveillance and diagnostics. Veterinary Microbiology 151, 153–9. doi10.1016/j.vetmic.2011.02.039

Scudamore, J.M., Harris, D.M., 2002. Control of foot and mouth disease : lessons from the experience of the outbreak in Great Britain in 2001. Revue Scientifique et Technique (International Office of Epizootics) 21, 699–710.

Simonsen, J., Mølbak, K., Falkenhorst, G., Krogfelt, K.A., Linneberg, A., Teunis, P.F.M., 2009. Estimation of incidences of infectious diseases based on antibody measurements. Statistics in Medicine 28, 1882–1895. doi10.1002/sim.3592

Simonsen, J., Strid, M. a, Mølbak, K., Krogfelt, K. a, Linneberg, A., Teunis, P., 2008. Sero-epidemiology as a tool to study the incidence of Salmonella infections in humans. Epidemiology and Infection 136, 895–902. doi10.1017/S0950268807009314

Simonsen, J., Teunis, P., Pelt, W. van, Duynhoven, Y. van, Krogfelt, K. a, Sadkowska-Todys, M., Mølbak, K., 2011. Usefulness of seroconversion rates for comparing infection pressures between countries. Epidemiology and Infection 139, 636–43. doi10.1017/S0950268810000750

Smeden, M. van, Naaktgeboren, C.a., Reitsma, J.B., Moons, K.G.M., Groot, J.a.H. de, 2013. Latent Class Models in Diagnostic Studies When There is No Reference Standard–A Systematic Review. American Journal of Epidemiology 179, 423–431. doi10.1093/aje/kwt286

Sminchisescu, C., Welling, M., Hinton, G., 2003. A Mode-Hopping MCMC sampler. University of Toronto Technical Report CSRG-478.

Smith, G.J.D. et al., 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459, 1122–5. doi10.1038/nature08182

Sotir, M.J., Cappozzo, D.L., Warshauer, D.M., Schmidt, C.E., Monson, T.A., Berg, J.L., Zastrow, J.A., Gabor, G.W., Davis, J.P., 2008. A countywide outbreak of pertussis: initial transmission in a high school weight room with subsequent substantial impact on adolescents and adults. Archives of Pediatrics & Adolescent Medicine 162, 79–85. doi10.1001/archpediatrics.2007.7

Sperlova, a., Zendulkova, D., 2011. Bluetongue: A review. Veterinarni Medicina 56, 430–452.

Stan Development Team, 2014. RStan: the R interface to Stan, Version 2.4.

Ståhl, K., Alenius, S., 2012. BVDV control and eradication in Europe — an update. Japanese Journal of Veterinary Research.

Steenbergen, M.R., Jones, B.S., 2002. Modeling Multilevel Data Structures. American Journal of Political Science 46, 218–237.

Stott, A.W., 2003. Costs and Benefits of Preventing Animal Diseases: A review focusing on endemic diseases (No. October). SAC.

Stritch, C., Naulty, F., Zintl, A., Callanan, J.J., McCullough, M., Deane, D., Marnell, F., McMahon, B.J., 2015. Squirrelpox virus reservoir expansion on the east coast of Ireland. European Journal of Wildlife Research. doi10.1007/s10344-015-0909-5

Szmaragd, C., Wilson, A.J., Carpenter, S., Wood, J.L.N., Mellor, P.S., Gubbins, S., 2009. A modeling framework to describe the transmission of bluetongue virus within and between farms in Great Britain. PLoS ONE 4. doi10.1371/journal.pone.0007741

Teunis, P.F.M., Eijkeren, J.C.H. van, Ang, C.W., Duynhoven, Y.T.H.P. van, Simonsen, J.B., Strid, M.A., Pelt, W. van, 2012. Biomarker dynamics: estimating infection rates from serological data. Statistics in Medicine 31, 2240–8. doi10.1002/sim.5322

Teunis, P.F.M., Heijden, O.G. van der, Melker, H.E. de, Schellekens, J.F.P., Versteegh, F.G. a, Kretzschmar, M.E.E., 2002. Kinetics of the IgG antibody response to pertussis toxin after infection with B. pertussis. Epidemiology and Infection 129, 479–89.

The Institute of Medicine, 2007. Global Infectious Disease Surveillance and Detection : Assessing the Challenges — Finding Solutions. The National Academies Press, Washington DC.

The Royal Society, 2002. Infectious diseases in livestock: Scientific questions relationg to the transmission, prevention and control of infectious disease in livestock in Great Britain. The Royal Society Publishing.

Thiboutot, M.M., Kannan, S., Kawalekar, O.U., Shedlock, D.J., Khan, A.S., Sarangan, G., Srikanth, P., Weiner, D.B., Muthumani, K., 2010. Chikungunya: a potentially emerging epidemic? PLoS Neglected Tropical Diseases 4, e623. doi10.1371/journal.pntd.0000623

Thomas, A., Spiegelhalter, D.J., Gilks, W., 1994. A language and program for complex bayesian modelling. Journal of the Royal Statistical Society. Series D 43, 169–177.

Thorner, R.M., Remein, Q.R., 1961. Principles and procedures in the evaluation of screening for disease. U.S. Department of Health, Education, and Welfare. Public Health Service. Public Health Monograph No. 67 24 pp.

Tjelmeland, H., Hegstad, B.K., 2001. Mode jumping proposals in MCMC. Scandinavian Journal of Statistics 28, 205–223.

Toft, N., Akerstedt, J., Tharaldsen, J., Hopp, P., 2007. Evaluation of three serological tests for diagnosis of Maedi-Visna virus infection using latent class analysis. Veterinary Microbiology 120, 77–86. doi10.1016/j.vetmic.2006.10.025

Toft, N., Innocent, G.T., Reid, S.W.J., 2004. Assessment of convergence in Hui-Walter models for diagnostic test evaluation. American Journal of Epidemiology 1–22.

Triple S Consortium, 2013. Guidelines for designing and implementing a syndromic surveillance system. Http://syndromicsurveillance.eu/Triple-S_guidelines.pdf.

Triveritas Ltd for a consortium, 2015. Field trial design to test and validate the performance of the CattleBCG vaccine and associated DIVA diagnostic test in England and Wales - Grant SE3287.

Tuite, A.R. et al., 2010. Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. CMAJ : Canadian Medical Association Journal = Journal de L'Association Medicale Canadienne 182, 131–136. doi10.1503/cmaj.091807

United States Department of Agriculture, 2010. National Scrapie Surveillance Plan. Fort Collins, CO.

Unkel, S., Farrington, P., Garthwaite, P., Robertson, C., Andrews, N., 2012. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Journal of the Royal Statistical Society Series A Statistics in Society 175, 49–82. doi10.1111/j.1467-985X.2011.00714.x

US Department of Health and Human Services, 1999. Biosafety in Microbiological and Biomedical Laboratories. Public Health Service 5th Editio, 1–250. doiciteulike-article-id:3658941

US General Accounting Office, 2001. Challenges in Improving Infectious Disease Surveillance Systems. United States General Accounting office, GAO-01-722 Global Health.

Uttamchandani, M., Neo, J.L., Ong, B.N.Z., Moochhala, S., 2009. Applications of microarrays in pathogen detection and biodefence. Trends in Biotechnology 27, 53–61. doi10.1016/j.tibtech.2008.09.004

Versteegh, F.G. a, Mertens, P.L.J.M., Melker, H.E. de, Roord, J.J., Schellekens, J.F.P., Teunis, P.F.M., 2005. Age-specific long-term course of IgG antibodies to pertussis toxin after symptomatic infection with Bordetella pertussis. Epidemiology and Infection 133, 737–48.

Wangersky, P.J., 1978. Lotka-Volterra Population Models. Annual Review of Ecology and Systematics 9, 189–218.

Warns-Petit, E., Morignat, E., Artois, M., Calavas, D., 2010. Unsupervised clustering of wildlife necropsy data for syndromic surveillance. BMC Veterinary Research 6, 56. doi10.1186/1746-6148-6-56

Waters, W.R., Palmer, M.V., Buddle, B.M., Vordermeier, H.M., 2012. Bovine tuberculosis vaccine research: historical perspectives and recent advances. Vaccine 30, 2611–22. doi10.1016/j.vaccine.2012.02.018

Watzinger, F., Ebner, K., Lion, T., 2006. Detection and monitoring of virus infections by real-time PCR. Molecular Aspects of Medicine 27, 254–98. doi10.1016/j.mam.2005.12.001

Wethey, D.S., Woodin, S.A., 2008. Ecological hindcasting of biogeographic responses to climate change in the European intertidal zone. Hydrobiologia 606, 139–151. doi10.1007/s10750-008-9338-8

Whelan, C., Whelan, A.O., Shuralev, E., Kwok, H.F., Hewinson, G., Clarke, J., Vordermeier, H.M., 2010. Performance of the enferplex TB assay with cattle in Great Britain and assessment of its suitability as a test to distinguish infected and vaccinated animals. Clinical and Vaccine Immunology 17, 813–817. doi10.1128/CVI.00489-09

WHO, 2012a. Global incidence and prevalence of selected curable sexually transmitted infections - 2009.

WHO, 2012b. Rapid Risk Assessment of Acute Public Health Events.

WHO Ebola Response Team, 2014. Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections. The England New Journal of Medicine 371, 1–15. doi10.1056/NEJMoa1411100

WHO, Global Commission for the Certification of Smallpox, 1980. The global eradication of smallpox. World Health Organization, Geneva.

Wilson, A.J., Mellor, P.S., 2009. Bluetongue in Europe: past, present and future. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 364, 2669–81. doi10.1098/rstb.2009.0091

Winter, K., Harriman, K., Zipprich, J., Schechter, R., Talarico, J., Watt, J., Chavez, G., 2012. California pertussis epidemic, 2010. The Journal of Pediatrics 161, 1091–6. doi10.1016/j.jpeds.2012.05.041

World Health Organization, 2000. WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases.

Wright, P.F., Nilsson, E., Van Rooij, E.M., Lelenta, M., Jeggo, M.H., 1993. Standardisation and validation of enzyme-linked immunosorbent assay techniques for the detection of antibody in infectious disease diagnosis. Revue Scientifique et Technique (International Office of Epizootics) 12, 435–450.

Yang, S., Rothman, R.E., 2004. PCR-based diagnostics for infectious diseases : uses , limitations , and future applications in acute-care settings. The Lancet Infectious Diseases 4, 337–348.

Zeimes, C.B., Olsson, G.E., Ahlm, C., Vanwambeke, S.O., 2012. Modelling zoonotic diseases in humans: comparison of methods for hantavirus in Sweden. International Journal of Health Geographics 11, 39. doi10.1186/1476-072X-11-39

Zhang, Y., Lopez-Gatell, H., Alpuche-Aranda, C.M., Stoto, M.A., 2013. Did advances in global surveillance and notification systems make a difference in the 2009 H1N1 pandemic?–a retrospective analysis. PloS One 8, e59893. doi10.1371/journal.pone.0059893

Zinsstag, J., Schelling, E., Waltner-Toews, D., Tanner, M., 2011. From "one medicine" to "one health" and systemic approaches to health and well-being. Preventive Veterinary Medicine 101, 148–56. doi10.1016/j.prevetmed.2010.07.003