HIERARCHICAL MODELLING FOR INFECTIOUS DISEASES

ZHENG XIAOHUI

(BSc.(Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Zheng Xiaohui

13 Mar 2014

Acknowledgements

First and foremost, I wish to express my genuine gratitude and heartfelt thanks to my supervisor, Dr Alex R. Cook. He has been very kind, patient and encouraging in his guidance and I have learnt a lot from him. I thank him for introducing me to hierarchical modelling. Through all the uncountable motivating discussions, we made the research projects feasible. Without all his priceless advices and timely feedback, this thesis would not have been possible.

I would like to thank all the faculty members, support staffs and my postgraduate classmates of Department of Statistics and Applied Probability of National University of Singapore (NUS) for providing me a great learning experience and all the precious advices. I thank NUS for awarding me with NUS Research Scholarship as a financial support during the first four years of my PhD studies. I am also surrounded by amazing colleagues in Saw Swee Hock School of Public Health who are accommodating during the last year of my PhD studies.

I thank Dr Vernon Lee J. for giving me the chance to get involved in the dengue project. I am grateful to him for allowing the access to the data used in this dengue project and also giving comments to the H1N1 project. I also thank Dr Mark Chen I-Cheng for his feedback on my work in relation to the H1N1 pandemic.

I wish to thank my family members who have always supported me in everything I do. They are always there when I needed them and I appreciate them for their unwavering love and care for me. Without their understanding and support through all the difficult times, I would not have embarked on this journey or even be able to made it through.

Contents

Summary

Bayesian methods have become increasingly used in infectious disease modelling, both statistical and mathematical models. This evolution of infectious disease modelling has led to demands for more sophisticated models that make use of expensive, yet messy, data as efficiently as possible. Fitting such models is challenging, however.

My thesis will address this issue by considering flexible hierarchical models, which pool information from related datasets to provide more accurate estimates of key parameters, and use appropriate algorithms to fit data for three infectious disease applications.

First, we used Bayesian methods to analyse clinical data from patients admitted to Tan Tock Seng Hospital, Singapore, for either dengue or Chikungunya, two mosquito-borne infections that have similar presentation. In the first part of this analysis, a Bayesian logistic regression model was developed to predict the aetiology using the significant variables found in our previous publication (V. J. Lee et al., 2012), with different prior distributions for regression coefficients. In the second part of the analysis, hierarchical models are fitted to clinical or laboratory temporal data from these patients to infer differences in these two similar diseases over time, to guide clinical management and diagnosis. Just Another Gibbs Sampler (JAGS) was used to estimate the key parameters in characterizing the observations trend that were modelled hierarchically. The routine was repeated for four significant variables, Haematocrit, Platelet Counts, Leukocytes and patient's temperature.

Next, we developed a hierarchical model for the 2009 H1N1 pandemic in a network or basket of countries. Data in relation to the influenza pandemic were collated via a literature search and Bayesian evidence synthesis was used to combine information from these data to infer accurate severity metrics. A hierarchical adaptation of the common Susceptible-Infected-Removed (SIR) compartmental model was fitted to the

datasets. Markov Chain Monte Carlo (MCMC) was used to establish an initial, rough estimate of the parameters' posterior distribution and sequential importance sampling was used to perform parameter estimations more efficiently.

Last, we examined the age-specific prevalence of Enterovirus 71 (EV71) in Asian countries using a hierarchical model. An MCMC algorithm is used to build the posterior samples of the parameters and hyperparameters, which are used within a Bayesian optimal design routine to plan future studies of EV71 seroprevalence in other Asian populations. We probe the possibility of different optimization criteria and design search methodologies. We finally selected the criterion that maximises the reciprocal of the absolute determinant of the variance-covariance matrix from a Weibull survival regression model fitted classically, allowing the use of prior information to design a study that would be analysed within a classical framework. Using a good experimental design such as that developed in this thesis for expensive serological studies can reduce costs without appreciably reducing information content.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Infectious diseases are of great concern for they impact public health and the economy. Mathematical and statistical models can be developed to understand how diseases spread and predict the severity of outbreaks in real-time for effective policy making.

Models for real-time analysis allow policy makers and hospitals to prepare by forecasting the magnitude of outbreak before they happen. Preventive measures can be assessed on computer experiments to decide the most appropriate response during the course of the epidemic. The 2009 influenza pandemic illustrated the importance of disease models, from assessing effectiveness of interventions *in silico* (Cook, Gibson, Gottwald, & Gilligan, 2008) to forecasting burden and severity (Ong et al., 2010).

1.2 General Infectious Diseases Modelling

Accurate and useful forecasts require infectious diseases models that have been appropriately selected and rigorously fitted to disease outbreak data. They must encapsulate the rate of infection and recovery from disease within host dynamics. Here, we will discuss two typical infectious diseases models, the Susceptible-Infected-Removed *(SIR)* and Susceptible-Infected-Susceptible *(SIS)* models.

The nature of the disease determines the appropriate model. If recovery confers immunity, a Susceptible-Infected-Removed (SIR) model may be appropriate. In this model, there are three classes of people in the population: Susceptible (S), Infected (I) and Removed (R). Prior to infection, individuals are classified under S. When

infection occurs, they shift from S to I. Transition from S to I is controlled by the rate of infection parameter. Upon recovery, or death, individuals move to the R class, an event governed by the rate of removal. These two parameters determine whether the epidemic might spread or become extinct as they determine the Basic Reproduction Number (R_0), a key quantity in infectious disease epidemiology.

R_0 is the ratio of the total instantaneous rate of infection to the total instantaneous rate of removal in an immunonaive population (Lee, 1997). If $R_0 > 1$, at the start of an epidemic, each case typically causes more than 1 secondary infections over his lifetime and thus the infectious disease might persist to cause a large outbreak. On the other hand, if $R_0 < 1$, more removals happen than infections, so the epidemic will die out. R_0 also determines the strength of the response as interventions must bring R_0 below 1 if they are to contain an outbreak.

Susceptible-Infected-Susceptible (SIS) model is another common infectious disease model appropriate if one can be reinfected after recovery. There are two groups of people in the population: Susceptible (S) and Infected (I). In contrast to the SIR model, upon recovery, individuals return to the S class. The calculation of R_0 in this model is similar to that in the SIR model.

There are many other infectious disease models other than the two that were discussed above. Thus, the behavior of the emerging disease outbreak has to be known first so that the most appropriate model can be chosen and fitted, as described in chapter 2, to give the most suitable analysis.

1.3 Bayesian Statistics

Parameters for rate of infection and removal are two of the key quantities characterizing an epidemic. At the start of an epidemic, these parameters are unknown but they must be estimated to understand and forecast the epidemic. Because even simply stated epidemic models create complex likelihood functions, the difficulty in parameter estimation is the main problem tackled in this thesis.

Prior information of the parameters can be drawn from observed data or past experiences to provide information for the actual characteristic of the parameters. In predicting the type of disease based on simple clinical and laboratory predictors and observational time course analysis for Dengue Hemorrhagic Fever and Chikungunya

(Lee et al., 2012), we will demonstrate how different priors of logistic regression coefficients can affect the probability of making a correct prediction.

Often, the actual number of infected cases cannot be recorded. In the H1N1 example, we will utilise the idea of Bayesian evidence synthesis with strong priors to gather information about the I and R classes of the SIR model from related data.

In Bayesian parameter estimation, Markov Chain Monte Carlo (MCMC) can be used. The advantage of MCMC is its ability to fit a complex model without existing solutions. There are several tools to perform MCMC: Just Another Gibbs Sampler (JAGS) is a program that will work for simple models (Plummer, 2013). Due to the rigidness of this program, the intricacy of some problems can only be resolved by carrying out the analysis in another language, such as R (R Core Team, 2013).

1.4 Hierarchical Modelling

In analyzing an epidemic, observations from similar epidemics can be useful because epidemics do not always happen in isolation, and so if a disease is affecting a different population, aspects such as the rate of infection and removal might be similar across different populations. Although the epidemic trajectories might not always be in synchrony, some things will generalise, like the removal rate.

H1N1 is an example of a potentially infectious disease that needed to be analysed while the pandemic was still at an early stage. Opportunities for prediction based on other countries' surveillance systems arise. In the early stage of the H1N1 pandemic at July 2009, Singapore's Minster of Health, Mr Khaw had correctly predicted the peak of the number of H1N1 infections in Singapore based on the information collected from the New York City which had already experienced their peak (Chua, 2009). This simple approach motivates the formal use of scientific evidence synthesis with hierarchical modelling for analysis of pandemic progression. We will show that pooling information from different countries allows better analysis of the infectious disease using hierarchical modelling via an application to the H1N1 pandemic example where the influenza outbreak was observed in 15 different countries or territories during the worldwide spread in 2009.

Other than forming hierarchical models for different populations, we can apply the same idea by using data from patients to predict the course of disease better by

pooling information from other patients. Information can be used to guide diagnosis for unknown pathogen. This concept will be implemented in the Dengue example. Daily information from each patient is contributed to the hierarchical model that will characterise the time course of several symptoms for patients with Dengue Hemorrhagic Fever or Chikungunya (Lee et al., 2012). Because patients with each disease are apprehended to provide similar information on the syndromes, a hierarchical model was used to amalgamate this knowledge. Supplementary to the ability to borrow strength, hierarchical model can also measure the variability of the parameters across different patients.

To decide on the control measures when faced with an outbreak, we need to know the burden of the infectious disease. This is often done by using serological studies to access past exposure which is very expensive. If a serological study were badly designed, it would lead to a wastage of money. The flexibility of hierarchical models allow parameter modelling for infectious diseases in different countries at different time point if there is no available information from our country. This will be demonstrated in the Enterovirus 71 (EV71) serology optimal design experiment. A hierarchical model will provide information on the means and variances of the parameters measuring the crucial rates in the epidemic for each population. This gives insights for the current prevalence situation so that we can derive the best design experiment that can save time and cost for the best experimental effect.

1.5 Structure of Thesis

In the next chapter, we will illustrate the methodologies that will be used in this thesis. Following that we will demonstrate the Bayesian logistic regression and the time course hierarchical modelling on Dengue Hemorrhagic Fever and Chikungunya in chapter 3. In chapter 4, we will devise the hierarchical model for the H1N1 pandemic in 2009 based on the Bayesian evidence synthesis techniques to combine multiple sources of information. In chapter 5, we will analyse past EV71 serological studies using hierarchical model to provide strong prior information for setting up an optimal design to examine the prevalence of EV71 efficiently.

Chapter 2

Tools and Methodology

2.1 Modelling of Infectious Diseases

Infectious disease epidemics are one of the leading causes of mortality. On top of the direct impacts on society—in the form of mortality, admissions to intensive care units, hospitalisations—are indirect impacts, such as work and school absenteeism (Meltzer, Cox, Fukuda, et al., 1999), and impacts on tourism. For example, the 2003 SARS outbreak was estimated to have caused a drop of 0.47% to Singapore's GDP (Lee & McKibhin, 2004). Policy makers such as ministries of health have to make decisions whether to implement counter-measures such as quarantine, vaccination, or school closure, and when to step up and step down such interventions (Cauchemez et al., 2006), which themselves have costs that impact the economy. Mathematical modelling can play an important role in guiding these decisions (Hethcote, 2000).

Predicting the spread of an emerging disease outbreak and the effect of control measures *in silico* is faster, cheaper and safer than waiting for an actual outbreak to occur and performing a randomised controlled trial, which may be infeasible on ethical and practical grounds. It can also play a role in providing real time information to the public, satiating their demand for information on what is happening and how the outbreak may evolve (Ong et al., 2010). Forecasting the progression of spread is also essential for deciding the most appropriate measure against the infectious diseases in the shortest possible time.

Hammond and Tyrrell (1971) demonstrated the use of deterministic mathematical model to study upper respiratory tract infection outbreaks in Tristan da Cunha, an island located in the Southern Atlantic Ocean where, due to its remoteness,

outbreaks of influenza and similar viruses would occur only after the arrival of a ship from South Africa. Their approach used a series of ordinary differential equations, fit to observed data on the number of islanders symptomatic using least squares. As the solution of ordinary differential equations (ODE) is only one, fixed path for fixed initial, or boundary, conditions and parameter values, their approach disregards other possible trajectories that the epidemic might have taken, limiting its usefulness for forecasting. However, for large outbreaks—unlike those on Tristan da Cunha, for instance, influenza pandemics in large, globally connected cities—the trajectory traced by ODE solutions will fall close to that of more complicated, stochastic models, but with the deterministic model's output requiring less computing power to derive. As a result, deterministic mathematical models are popular and efficient, especially when the involved population is large. This thesis will apply deterministic models to model influenza A (H1N1-2009) later in Chapter 4.

In contrast to deterministic models, stochastic models are formulated to allow chance events to impact epidemic trajectory. They are often set up as temporally inhomogeneous Poisson processes, but may also use estimates of sojourn time to parametrise non-exponential within-host event times. One method of simulation uses Gillespie's algorithm (Keeling & Ross, 2008; Gillespie, 1977) in which the rates of successive events are recalculated after each event is simulated, reflecting the fact that rates change with time in an inhomogeneous Poisson process. Such a procedure has to be repeated until the end of the time period of interest, or until no further events are possible, and is hard to parallelise, thus making it a computer intensive process to simulate. Stochastic models are more suitable for small populations or small outbreaks where chance events can happen. In principle, stochastic models for large populations are possible, but they will look and behave like deterministic model which is more computationally efficient for large populations. According to Barbour (1974), the asymptotic properties of a stochastic model will approximate a deterministic model by the Central Limit Theorem for a population growth model as illustrated in the following figure.

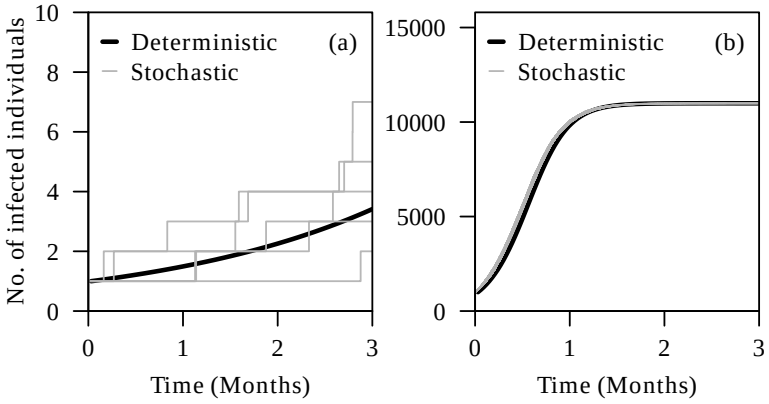

Figure 2.1: **Comparison of Deterministic and Stochastic SIR model for small population (in panel (a)) and large population (in panel (b)).** The solution of deterministic ODE for the number of infected individual is represented by the solid black lines in both panels. The grey lines are the possible trajectories simulated using Gillespie's algorithm on 5 different trials. For both situations, the rate of infection per Susceptible-Infected pair and rate of removal per infected individual is 0.25 and 0.2 respectively. The total population sizes used in panel (a) and (b) are 1 000 and 10 000 respectively, where panel (a) started with 1 infected individual and panel (b) started with 1 000 infected individual. Computation details can be found in section 4.3.6.

Because parametrisation, which may involve simulation, will lead to much complication, much work has focused on developing statistical methods to fit these to observational data, as described in the following paragraphs.

Approximate Bayesian computation (ABC) has become widely used in population genetics where likelihood cannot be calculated easily (Wilkinson, 2013; McKinley, Cook, & Deardon, 2009; Marjoram, Molitor, Plagnol, & Tavaré, 2003). It has also been used for epidemic models (Blum & Tran, 2010) where, similarly, the exact event times or natures cannot be directly observed. This method works by measuring the discrepancy between the simulated data, generated using proposed parameter values, and the observed data. There are many variants but within a Markov chain Monte Carlo approach, the proposed parameters will be accepted with higher probability if the discrepancy is smaller (Wilkinson, 2013). On top of the metric for discrepancy, a tolerance, $\epsilon > 0$, should be set which governs the acceptance or rejection the simulated parameters (Wilkinson, 2013). In many scenarios, this approach is highly efficient, for it replaces the need to calculate likelihoods in a Metropolis-Hastings algorithm, which would otherwise be required to calculate the posterior density. A weakness of ABC, as described by Robert et al. (2011), is that theoretical discussions on its convergence properties are missing due to the algorithm's typical use of non-sufficient summary

statistics, which lead to an unidentified quantum of loss of information. Robert et al. (2011) have demonstrated that ABC methodology in some examples is unreliable for Bayesian model selection, for instance.

Another approach for fitting stochastic models, as used Ross et al. (2006), is the Cross-Entropy (CE) method to find the parameter that will maximise the likelihood function of a model. CE involves simulating a large number of potential parameter values, typically in a swarm around the current best guess, and using these to update the best guess. The algorithm repeatedly explores the parameter space locally until a criterion is met for stopping. This method will be presented in the serology example in Chapter 5.

Another approach that is commonly used in other settings with partially observed data is the expectation maximization (EM) algorithm. In EM algorithm, there are two main steps. First, the E step will compute the expected value of the log likelihood based on the conditional distribution of the augmented data simulated from the current set of parameters given the observed data. Second, the M step finds the parameter that will maximise the expected value from the E step. These two steps have to be repeated until the parameter converges. However, O'Neill et al. (2000) have argued that the EM algorithm is not suitable for epidemic outbreak models as the required conditional expectation is very difficult to compute with the heavily censored data from epidemics. Instead, they argue that data augmentation is required for the computation of the likelihood, integrating over unobserved event times and states. This approach has been widely used to fit temporal and spatio-temporal models (Cook, Otten, Marion, Gibson, & Gilligan, 2007; Gibson & Renshaw, 1998). It is, however, extremely computationally expensive if the population is large or the time frame is long since the augmented parameter space, including actual parameters and the augmented variables, will in such cases be massive. To illustrate the difficulties, consider the following example, based on incomplete data on pneumococcal infection in school children (Cauchemez et al., 2006). They sampled 2 807 children, 3 to 6 years old, from 50 schools in France to collect 5 swabs over 5 months to investigate transmission of 15 pneumococcoal serotypes. As, however, not all children provided all 5 swabs, Cauchemez et al. (2006) used data augmentation to infer the missing number of bacterial serotypes and actual event times. In that analysis, a Bayesian

hierarchical model was added to reduce the heteroskedasticity of estimates of the time course of infection for each child. This combination (of data augmentation and hierarchical modelling) is valuable for the typically highly censored data that are unavoidable in observational field studies of infectious diseases. In that context, the approach was feasible as a mere 2 807 children formed the dataset, so the resulting parameter dimensionality was not excessive. When the population or study size is large, however, the amount of augmentation required may be prohibitively large.

Statistical modelling of outbreak progression can provide valuable information in the course prediction for planning. At the early stage of an outbreak, there is insufficient information from that outbreak to appraise the risk appropriately, as evinced by the initial uncertainty of the WHO and several governments in their response to the H1N1 pandemic of 2009 (Chang, Southard, & Sullivan, 2010). However, if information can be drawn from other sources in real time (to be presented in the H1N1 example), or historical outbreaks (to be demonstrated in the serology example) then better decision making can be made. Bayesian inference provides a natural mechanism to do this.

2.2 Bayesian Inference

In this thesis, we use Bayesian statistical concepts to estimate model parameters from, typically, messy and partially observed data. In this section, we will justify our use of Bayesian statistics and why it is preferred for our applications, as well as provide a brief overview of how Bayesian statistics works and the distinction with Classical statistics.

Under the Classical paradigm, parameters are fixed, i.e. non-random, numbers; in contrast, they are random variables in Bayesian statistics (O'Neill, 2002; Gelman, Carlin, Stern, & Rubin, 2003), where their distribution characterises the uncertainty in their values after observing data. Although in many cases, Classical approaches perform well and may be preferable to Bayesian ones, in others they are too inflexible for complex modelling problems. Unlike Classical approaches that are often based on the asymptotic normality assumption, and hence rely on large enough samples to justify their use, Bayesian methods typically are not (Congdon, 2001), and may be used for small (or, non-infinite) samples. When its assumptions are violated by the

data, the result from a Classical analysis may be unacceptably approximate. In such cases, Bolstad (2004) has claimed that Bayesian methods often outperform Classical methods, even when judged by Classical criteria.

We illustrate one advantage of Bayesian methods over Classical ones using the example of the ALVAC/AIDSVAX HIV vaccine trial in Thailand (Rerks-Ngarm et al., 2009). In their study, Rerks-Ngarm et al. (2009) vaccinated $n = 8\,197$ individuals of whom $x = 51$ were infected (similar numbers were given a placebo). If p is the probability of HIV infection over the study time frame for a vaccinated individual, we can assume a binomial model as $x \sim \text{Bin}(n, p)$. Under the standard, Classical approach, p is estimated by $\hat{p} = \frac{x}{n} = 0.00622$, the maximum likelihood estimate and the empirical fraction infected, and the 95% confidence interval of p is calculated by $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, i.e. $(0.00452, 0.00792)$. This confidence interval calculation assumes that the sample size is sufficiently large that the MLE is normally distributed. (In contrast, a Bayesian approach taking a uniform prior for p and using either Markov Chain Monte Carlo sampling or direct calculation quantifies the full distributional profile of the parameter without requiring the sample size be approximately infinite (Congdon, 2001).)

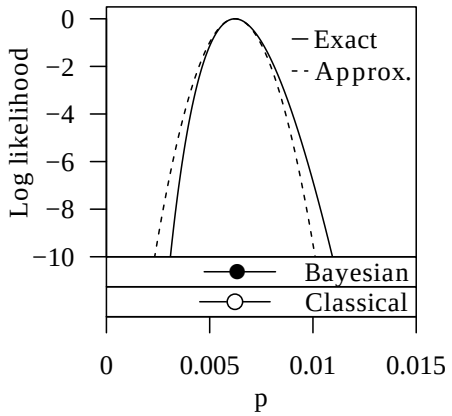

Figure 2.2: **Comparison of Bayesian and Frequentist estimates for the Thailand HIV trial example.**
The exact log likelihood of the data based on the binomial distribution is represented by the solid line. The dotted line is the approximated log likelihood based on a normal distribution where the mean is $\hat{p}$ and standard deviation is the estimated standard error $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The mean of posterior samples is represented by the solid dot and the 95% credible interval is represented by the line in the lower panel. The MLE $\hat{p}$ and the Classical confidence interval is represented in the lowest panel by the hollow dot and line.

Even for the large sample size of 8 197 in the ALVAC/AIDSVAX trial, a pronounced asymmetry in the likelihood can be discerned in figure 2.2, which is not adequately characterised in the Classical confidence interval. In contrast, the Bayesian credible interval can account for this asymmetry and arguably give a more accurate depiction of the uncertainty in the parameter.

Typically both approaches employ the same fundamental statistical concept, the likelihood function. Suppose the observed data are D and the parameter is θ, in which case the likelihood (function) of θ is the probability of observing the data given the parameter,

$$L(\theta) = f(D|\theta). \tag{2.1}$$

In many non-infectious disease applications, the likelihood can be factorised into a product of terms, one for each datum, but as infectious diseases are communicable, the disease states of different individuals are positively correlated, and so in general this factorisation cannot be assumed. If a parameter value of θ, and the model it belongs to, fit the data D well, $L(\theta)$ will be relatively large, and this is often exploited in Classical statistics by calculating the value of the parameter, $\hat{\theta}$, also known as maximum likelihood estimate (MLE), that will give the largest $L(\theta)$, i.e.

$$\hat{\theta} = \arg\max_{\theta} L(\theta). \tag{2.2}$$

Although loosely interpreted as the value which promises to be the 'most likely' given the observed data, for that model, such probabilistic statements about parameters are not permissible in Classical statistics and the term 'maximum likelihood' is unfortunate.

In contrast, Bayesian estimation uses Bayes' theorem to combine both sources of information, prior and data (Bolstad, 2004), by converting from the probability distribution of the (known) data given knowledge of the (unknown) parameters, to the probability distribution of the (unknown) parameters given knowledge of the known data.

For events A and B, Bayes' rule states that the joint probability of A and B can be derived from the conditional probability

$$\Pr(A, B) = \Pr(A|B)\Pr(B) \tag{2.3}$$

$$= \Pr(B|A)\Pr(A). \tag{2.4}$$

Through simple manipulation, Bayes' theorem states that

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}. \tag{2.5}$$

Replacing variables, we obtain the posterior distribution, $\Pr(\theta|D)$ (or, henceforth, $f(\theta|D)$ as typically θ is continuous and therefore has a probability density). The prior is the distribution of the parameter that we assume before accounting for the observed data. In the Bayesian framework (Gelman et al., 2003), the density of the posterior can be represented by

$$f(\theta|D) = \frac{f(D|\theta) \cdot f(\theta)}{f(D)} \tag{2.6}$$

where

$$f(D) = \int_{\theta} f(D|\theta) \cdot f(\theta) d\theta \tag{2.7}$$

is a constant that can be found by integration but can sometimes be ignored (in popular methods such as Markov Chain Monte Carlo and importance sampling) to get a direct proportionality between the posterior density and the product of likelihood and prior density:

$$f(\theta|D) \propto f(D|\theta) \cdot f(\theta) \tag{2.8}$$

$$\text{Posterior} \propto \text{Likelihood} \cdot \text{Prior}. \tag{2.9}$$

2.2.1 Prior choice

A perceived limitation of Bayesian statistics is the requirement, and occasional difficulty, to select a prior distribution for the parameters. As the posterior distribution, which leads directly to reported estimates of model parameters, is proportional to the likelihood and prior distribution, different prior distributions will lead to different posterior distributions in an apparently subjective way. Although prior distributions are chosen based on our preliminary beliefs (Lee, 1997), which may differ from analyst to analyst, Bolstad (2004) and others have argued that in practice, although different prior distributions may be used, the posterior distributions are typically similar as the data will swamp the prior if they contain sufficient information content. With adequate data, the effect of the exact choice of prior specification is minimal compared to the effect of the data themselves (Bolstad, 2004). This will be shown in the

subsequent example of logistic regression to distinguish Dengue from Chikungunya infection.

Priors should be chosen based on the objective of the analysis. If there is sufficient information in the data about all the parameters, a flat (or approximately flat) non-informative prior is commonly used so that the posterior distribution is proportional to the likelihood density. If external data are available and the information is relevant to the current analysis, an informative prior can be derived from those data to accumulate more evidence about the parameter values. An example of this approach to developing an informative prior will be shown for the recovery rate of H1N1 in Chapter 4, based on an analysis of the time course of infection in a volunteer challenge study by Carrat et al. (2008). Conversely, if a subjective prior is not objectionable, for example if the analysis is being used as a guide to decision making, it could be set from personal or experts' belief. Hierarchical models, to be elaborated in the next section, also act as an indirect form of informative prior, which allows information to be borrowed between different parts of the datasets by assuming a common distribution specified by hyper-parameters.

When there are insufficient details of parameter, a non-informative or flat prior, which does not favour any value, is used to avoid undue influence on the posterior distribution (Bolstad, 2004) and to allow the data to speak for themselves. By using a flat prior distribution, no parameter values are given additional 'weight' beyond the information in the data (Congdon, 2001). An extreme example of a flat prior is a uniform distribution from negative to positive infinity, which gives the same prior density for any real value. In such cases, the posterior is only proportional to the likelihood.

Other non-informative priors can also be used, depending on the parameter support (Gelman et al., 2003). If a parameter should be positive, an exponential distribution or log normal distribution could be used, for instance. For the exponential, a small rate parameter gives a distribution with a large mean, that may be effectively flat over the range of values with high likelihood.

Choosing a flat prior distribution is not always the best solution in situations where we wish or need to pool information from multiple data sources, in which case an informative prior distribution built from an earlier dataset can allow a better

posterior distribution for the model parameters after analysing a later dataset (Lee, 1997).

If informative priors are chosen, it may be valuable to perform a sensitivity analysis, performing several analyses using different prior distributions, comparing the resulting posteriors, and checking the robustness of the conclusions to these assumptions, in a similar way that sensitivity analyses of data-model assumptions are sometimes conducted (O'Neill & Roberts, 1999). If all posteriors are similar, any prior distributions can be adopted with confidence, but if the results are contrasting, extra care must be taken to ensure the prior used in the reported analysis is appropriate.

2.2.2 Computational issues

Markov Chain Monte Carlo (MCMC) is a technique for computing posterior distributions that are not otherwise analytically tractable. The exact posterior density can rarely be calculated, because although the posterior is proportional to the product of prior and likelihood, it is hard to calculate the unknown proportionality constant $f(D)$ (cf. Section 2.2) for the integral of the posterior to be equal to 1. (This problem is comparatively harder than maximising the likelihood in Classical statistics due to the difference in complexity between maximisation and quadrature.) The problem is especially severe in a high dimensional parameter space. However, for most purposes, the problem is obviated using MCMC for two reasons, described below.

MCMC is an extension of the Monte Carlo (MC) technique, which involves drawing samples from a distribution. If (large) samples are drawn directly from the posterior distribution, the statistics required to describe the posterior can be calculated with ease simply by calculating characteristics of the sample (thus avoiding subsequent integration). For instance, instead of integrating to get the posterior mean, $E(\theta|D) = \int_{\mathbb{R}} \theta \cdot f(\theta|D) d\theta$, it can instead be calculated using the average of the sample $\{\theta_i : i = 1, \ldots, n\}$ from the posterior, i.e. $E(\theta|D) = \frac{1}{n} \sum_{i=1}^{n} \theta_i$. By the Strong Law of Large Numbers, if $\{\theta_i : i = 1, \ldots, n\}$ are pairwise independent and identically distributed (*i.i.d.*) such that either $E[(\theta_1)_+]$ or $E[(\theta_1)_-]$ is finite, then, $\frac{1}{n} \sum_{i=1}^{n} \theta_i \to E(\theta_1)$ almost surely as $n \to \infty$ (Gilks, Spiegelhalter, & Richardson, 1996). If the sample is large enough, the estimate from posterior sample is effectively the actual quantity from the posterior.

However, except for special cases, it is difficult to sample from the posterior distribution as its properties are not generally known. MCMC overcomes this by drawing samples from a Markov Chain with the posterior as the stationary distribution. Under certain conditions (ergodicity, irreducibility, and aperiodicity), Markov chains will converge to a stationary distribution (Gilks et al., 1996). Convergence can be assessed heuristically after comparing different simulations (Gelman et al., 2003), which if simulated independently should converge to the same distribution. The Metropolis-Hastings algorithm is one way to construct a Markov chain such that its stationary distribution is the posterior of interest, and requires only being able to evaluate the (log) posterior to a constant of proportionality. The sampled posterior distribution is a collection of values from a Markov Chain, and as a result the sampled values are typically autocorrelated. The usual solution to mitigate the dependency of successive simulations is to thin the posterior sample, where only the kth simulated value (say, $k = 10$) is kept, so that the correlation between subsequent simulations can be reduced (Gelman et al., 2003). Gelman et al. (1995) added that thinning is beneficial as large numbers of parameters poses a problem for computer storage. However, with today's technology, this problem is less of an issue and throwing away simulations wastes information, and thus thinning is avoided unless the number of parameters to store is large and the mixing poor.

The general concept of MCMC is to set up a Markov Chain whose stationary distribution is equal to the unknown posterior distribution. Because even if different initial values θ_0 are used, the resulting chains ought still to converge to the stationary distribution, multiple chains with different initial values can be simulated and a comparison of their apparent stationary distribution made to assess this property.

By simulating the Markov Chain, the set of $\{\theta_t : t > n\}$ is the posterior sample for analysis of the model. The first n values of the Markov Chain are typically removed from the posterior sample as they might precede convergence to the stationary distribution. This process is called 'burn in'. Usually, burn-in discards the first 1% to 2% of the simulation, provided the simulations have shown convergence (Gilks et al., 1996).

One way to obtain a Markov chain with the desired stationary distribution, is to use the popular Metropolis-Hastings Algorithm:

1. Choose an initial value, θ_0.

2. Calculate the posterior density under the chosen parameter, θ_0, and represent it by h_{old}.

3. Simulate θ_{new} using a proposal distribution, where the probability of proposing θ_{new} from θ_0 is represented by $q(\theta_{\text{old}} \rightarrow \theta_{\text{new}})$.

4. Calculate the posterior density under the new parameter, θ_{new}, and represent it by h_{new}.

5. A decision to accept or reject θ_{new} is made by calculating the acceptance probability, P_{acc}. θ_{new}, will be accepted as the next value in the chain, θ_1, with probability P_{acc}. Otherwise, θ_{new} will be rejected with probability $1 - P_{\text{acc}}$ and the next value in the chain remains as θ_0,

$$P_{\text{acc}} = \min\left(1, \frac{h_{\text{new}}}{h_{\text{old}}} \cdot \frac{q(\theta_{\text{new}} \rightarrow \theta_{\text{old}})}{q(\theta_{\text{old}} \rightarrow \theta_{\text{new}})}\right). \tag{2.10}$$

6. A random number U is generated from a uniform distribution with range [0,1]. If P_{acc} is more than U, accept θ_{new}. Otherwise, reject θ_{new} and retain θ_0.

7. Step 2 to 6 is repeated as θ_i is updated as θ_{i+1}. The length of the chain determines how many repetitions are required. As mentioned, the first fraction (say 1%) of the chain will not be stored due to the requirement of burn-in. To perform thinning, every jth value of the chain will be stored, where j will be preselected.

The choice of the proposal distribution will affect the jumping rules and efficiency of the simulations. The ideal Metropolis-Hastings algorithm will simulate parameters directly from the posterior distribution, in which case the acceptance probability is 1 (Gelman et al., 2003) and the algorithm can be seen to be equivalent to vanilla Monte Carlo. An alternative is to simulate one parameter, or a block of parameters, from their conditional distribution given the data and all other parameters, in which case, again, the acceptance probability is 1. This is known as Gibbs Sampling.

In principle, almost any distribution can be used for proposal distribution (Congdon, 2001). A normal distribution is usually preferred as it is symmetrical which allows equal chance of positive and negative jumps, if its mean is set to the current value

of the parameter. This also implies that $q(\theta_{\text{new}} \to \theta_{\text{old}}) = q(\theta_{\text{old}} \to \theta_{\text{new}})$ which simplifies the calculation of the acceptance probability, P_{acc}, as the ratio of the proposal distributions cancels. Thus, all the proposal distributions used in this project will be normal distributions, unless otherwise stated. The calculation of P_{acc} for θ in the Metropolis-Hastings Algorithm is, in this case:

$$P_{\text{acc}} = \frac{h_{\text{new}}}{h_{\text{old}}} \cdot \frac{q(\theta_{\text{new}} \to \theta_{\text{old}})}{q(\theta_{\text{old}} \to \theta_{\text{new}})} \tag{2.11}$$

$$= \frac{h_{\text{new}}}{h_{\text{old}}}. \tag{2.12}$$

The proportionality constant, where the posterior density is proportional to the product of likelihood and prior density, will be cancelled in this ratio, making h be just the product of likelihood and prior density and the acceptance probability:

$$P_{\text{acc}} = \frac{h_{\text{new}}}{h_{\text{old}}} \tag{2.13}$$

$$= \frac{l(\theta_{\text{new}}|D) \cdot f(\theta_{\text{new}})}{l(\theta_{\text{old}}|D) \cdot f(\theta_{\text{old}})}. \tag{2.14}$$

Because some probabilities have small values, the calculation of the acceptance probability is typically done working with a logged likelihood function, $\log l(\theta|D)$ and log prior density $\log f(\theta)$,

$$P_{\text{acc}} = \exp\left(\log \frac{l(\theta_{\text{new}}|D) \cdot f(\theta_{\text{new}})}{l(\theta_{\text{old}}|D) \cdot f(\theta_{\text{old}})}\right) \tag{2.15}$$

$$= \exp[\log l(\theta_{\text{new}}|D) - \log l(\theta_{\text{old}}|D) + \log f(\theta_{\text{new}}) - \log f(\theta_{\text{old}})]. \tag{2.16}$$

The variance (or covariance matrix) of the normal distribution used as a proposal needs to be chosen carefully. To be efficient when the posterior is unknown, the proposal distribution should be sufficiently wide to allow large jumps so as to achieve convergence quickly. If the jump steps were too small, i.e. if the proposal distribution is too narrow, the trace plot will take a long time to reach convergence. However, the proposal variance cannot be too large, as in this case most of the simulations will be rejected. Thus, a proposal with appropriately intermediate variance should be selected for good sized jump steps.

In models with a high dimensional parameter space, the Markov Chain may exhibit 'stickiness'. This can be due to a weirdly shaped posterior distribution, caused by correlations between two or more parameters, a poor choice of proposal distribution, or strong correlations between a multitude of parameters. A 'stuck' chain may

spend an inordinate time at a certain region of the parameter and take a long time to reach another, well supported region. This is a great challenge faced in this project, due to the complexity of the models used, and several attempted solutions will be discussed.

2.3 Bayesian Hierarchical Modelling

Estimation in sparse, scattered datasets can be made more robust by borrowing strength between the points where evidence, or information, accumulate. When parameters are analysed individually, similarities between different components of the model are neglected. Hierarchical modelling allows related identities in different sampling units to be brought together, reducing variability, or uncertainty, in parameters from different submodels while at the same time measuring the irregularity between these parameters using a hyper-distribution.

We need a set of parameters for each dataset to be coming from a distribution that is governed by the second-level parameters called hyper-parameters (Schervish, 1995). The n sets of independently observed data and the parameters θ_i, $i = 1, \ldots, n$, are iid and governed by the hyper-parameters (for instance, μ and σ) which will also have their own distribution.

The prior distributions of the parameters θ_i (or a transformation to make their support the real line) are typically represented as normal distributions with mean μ and standard deviation σ,

$$\theta_i \sim N(\mu, \sigma^2) \tag{2.17}$$

where the hyper-prior distributions for $\eta = (\mu, \sigma)$ might be chosen to be uniform distributions which are non-informative if there is no prior knowledge or beliefs for them, for instance

$$\mu \sim U(-1000, 1000) \tag{2.18}$$

$$\sigma \sim U(0, 1000). \tag{2.19}$$

Hence, from the posterior samples of the parameters and hyper-parameters, the characteristics of the posterior distributions of the parameters of interest can be obtained.

MCMC can be used to explore the parameters space of (θ_i, μ, σ) based on the stationary property discussed earlier, with some extension to the posterior calculations.

Using the Bayes' theorem,

$$f(\theta, \eta|D) \propto f(D|\theta, \eta) \cdot f(\theta, \eta) \tag{2.20}$$

$$\propto f(D|\theta, \eta) \cdot f(\theta|\eta) \cdot f(\eta). \tag{2.21}$$

The algorithm is:

1. Choose the initial values for all the parameters, $(\theta_i)_0$ for $i = 1, \ldots, n$, and hyper-parameters, $\eta_0 = (\mu_0, \sigma_0)$.

2. Calculate the posterior density under the chosen parameters, $(\theta_i)_0$, and hyper-parameters, $\eta_0 = (\mu_0, \sigma_0)$, and represent it by h_{old}.

3. Fixing the hyper-parameters at their current values, simulate $(\theta_i)_{\text{new}}$ using a proposal distribution, where the probability of proposing $(\theta_i)_{\text{new}}$ from $(\theta_i)_0$ is represented by $q((\theta_i)_{\text{old}} \to (\theta_i)_{\text{new}})$.

4. Calculate the posterior density under the new parameter, $(\theta_i)_{\text{new}}$ and the original hyper-parameters, $\eta_0 = (\mu_0, \sigma_0)$, and represent it by h_{new}.

5. Decision to accept or reject $(\theta_i)_{\text{new}}$ is made by calculating the acceptance probability, P_{acc}. The newly simulated parameter, $(\theta_i)_{\text{new}}$, will be accepted as the next value in the chain, $(\theta_i)_1$, with probability P_{acc}. Or, $(\theta_i)_{\text{new}}$ will be rejected with probability $1 - P_{\text{acc}}$ and the next value in the chain remains as $(\theta_i)_0$,

$$P_{\text{acc}} = \min\left(1, \frac{h_{\text{new}}}{h_{\text{old}}} \cdot \frac{q((\theta_i)_{\text{new}} \to (\theta_i)_{\text{old}})}{q((\theta_i)_{\text{old}} \to (\theta_i)_{\text{new}})}\right). \tag{2.22}$$

6. A random number U is generated from a uniform distribution with range [0,1]. If P_{acc} is more than U, accept $(\theta_i)_{\text{new}}$. Otherwise, reject $(\theta_i)_{\text{new}}$ and retain $(\theta_i)_0$.

7. Having made the decision for $(\theta_i)_1$, use the calculated posterior density under the parameters, $(\theta_i)_1$, and hyper-parameters, $\eta_0 = (\mu_0, \sigma_0)$ as h_{old}.

8. Fixing the parameters at their current values, $(\theta_i)_1$, simulate η_{new} using a proposal distribution, where the probability of proposing η_{new} from η_0 is represented by $q(\eta_{\text{old}} \to \eta_{\text{new}})$.

9. Calculate the posterior density under the new hyper-parameters, η_{new}, and the current parameters, $(\theta_i)_1$, and represent it by h_{new}.

10. Decision to accept or reject η_{new} is made by calculating the acceptance probability, P_{acc}. The newly simulated parameter, η_{new}, will be accepted as the next value in the chain, η_1, with probability P_{acc}. Otherwise, η_{new} will be rejected with probability $1 - P_{\text{acc}}$ and the next value in the chain remains as η_0,

$$P_{\text{acc}} = \min\left(1, \frac{h_{\text{new}}}{h_{\text{old}}} \cdot \frac{q(\eta_{\text{new}} \to \eta_{\text{old}})}{q(\eta_{\text{old}} \to \eta_{\text{new}})}\right). \tag{2.23}$$

11. A random number U is generated from a uniform distribution with range [0,1]. If P_{acc} is more than U, accept η_{new}. Otherwise, reject η_{new} and retain η_0.

12. Step 2 to 11 is repeated as $(\theta_i)_j$ is updated as $(\theta_i)_{j+1}$ and η_j is updated as η_{j+1}. The length of the chain determines how many repetitions are required. The first 1% of the chain will not be stored due to the requirement of burn-in. To perform thinning, every kth value of the chain will be stored, where k is preselected.

There are many calculations and some of them can be reused to save time. In step 5, as the hyper-parameters are fixed, the hyper-prior densities, $f(\eta)$, will be the same, and so P_{acc} can be simplified to

$$P_{\text{acc}} = \frac{l((\theta_i)_{\text{new}}|D) \cdot f((\theta_i)_{\text{new}}|\eta)}{l((\theta_i)_{\text{old}}|D) \cdot f((\theta_i)_{\text{old}}|\eta)}. \tag{2.24}$$

In step 10, the parameters are held fixed, and as a result the likelihood densities, $l((\theta_i)_1|D)$, will be the same for both proposed and current values. As long as the hyper-parameters' values fall within the interval in the uniform hyper-prior distribution, $f(\eta_{\text{new}}) = f(\eta_{\text{old}})$. Thus, P_{acc} can be simplified to

$$P_{\text{acc}} = \frac{f((\theta_i)_1|\eta_{\text{new}})}{f((\theta_i)_1|\eta_{\text{old}})}. \tag{2.25}$$

In step 10, $f((\theta_i)_1|\eta_{\text{old}})$ can be reused from step 5, taking the same value as $f((\theta_i)_{\text{new}}|\eta)$ if $(\theta_i)_{\text{new}}$ was accepted or $f((\theta_i)_{\text{old}}|\eta)$ otherwise. In the next round of step 5, $f((\theta_i)_{\text{old}}|\eta)$ can be reused from step 10 in the previous round, taking the same value as $f((\theta_i)_1|\eta_{\text{new}})$ if η_{new} was accepted as η_1 or $f((\theta_i)_1|\eta_{\text{old}})$ otherwise.

With the information on the parameters and hyper-parameters provided by the observed data, predictions can be done for other datasets. Based on the common dependence of all the parameters on the hyper-parameters characteristic of hierarchical models, random samples are collected from the hyper-parameters for representing other similar models (Schervish, 1995). Each set of $\eta = (\mu, \sigma)$ sampled from their respective posterior samples is used to generate θ_k, the parameter for the intended kth dataset, using the normal distributions stated for the model.

This method can be repeated a large number of times to estimate the distribution of the forecast θ_k. Having these as predictive posterior distributions, we can have a good idea of the range of parameter values that might arise in a similar dataset. This method of estimating the future when there is only limited information about the model is known as forecasting (Kleczkowski & Gilligan, 2007). We will demonstrate this in the dengue disease and serology examples.

Apart from predicting parameters for future data, hierarchical models can also improve the estimation of parameters for partially collected data. This is more useful in reality as incomplete real time data from similar populations may be available and the parameters for these incomplete current data are of public health interest. Pooling strength from each dataset to supplement the insufficient knowledge in other parts of the data will be exemplified in the H1N1 example in chapter 4.

2.3.1 Importance Sampling

The expected value of any function $g(\theta)$ of the parameters may be represented by the following integral,

$$E(g(\theta)) \quad = \quad \int g(\theta) f(\theta) d\theta \tag{2.26}$$

where $f(\theta)$ is the posterior distribution.

This integration can be approximated by taking the average of a large sample of size N of the parameters from the posterior,

$$E(g(\theta)) \quad \approx \quad \frac{1}{N} \sum_{i=1}^{N} g(\theta_i). \tag{2.27}$$

If it is difficult to sample from the posterior itself, we might instead simulate θ from another proposal distribution, $q(\theta)$. The expectation of $g(\theta)$ in this case can be

represented as

$$E(g(\theta)) \quad = \quad \int g(\theta)\frac{f(\theta)}{q(\theta)}q(\theta)d\theta. \tag{2.28}$$

The approximation of this expectation becomes

$$E(g(\theta)) \quad \approx \quad \frac{1}{N}\sum_{i=1}^{N}\left(g(\theta_i)\frac{f(\theta_i)}{q(\theta_i)}\right) \tag{2.29}$$

where $\frac{f(\theta_i)}{q(\theta_i)}$ is termed the weight of the ith draw, and denoted w_i. As the large sample size may still not be exactly a true representation of the actual model distribution, these weights will compensate for the discrepancy. The larger the probability density $f(\theta_i)$, the larger the weight w_i will be for a higher contribution of the $g(\theta_i)$ in the expectation approximation.

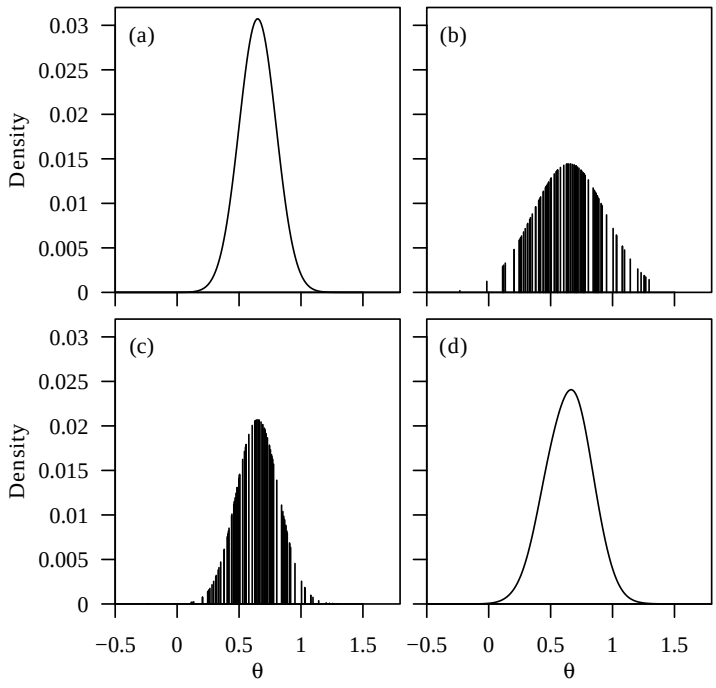

Figure 2.3: **Demonstration of how the weight w_l can be larger for better points with higher posterior density than those points with lower posterior density in Importance Sampling.**
In this example, the actual posterior distribution is represented in (a), normal with mean 0.65 and standard deviation 0.15. In panel (b), 100 particles are simulated from a normal proposal distribution with the same mean but the standard deviation is doubled to 0.3 and the proposal density is represented y axis. In panel (c), we calculate the weights for the simulated particles as the ratio of posterior to proposal density. Panel (d) shows the kernel density estimate of the simulated particles based on the weights in (c).

The main concept of importance sampling is to sample from an arbitrary distribution, rather than the (unknown) posterior directly, and correct the sample by allocating weights to all the sampled particles. The weights will distinguish the better particles from the inferior ones, but also correct for over or under sampling in some regions of the parameter space.

Resampling the initial sampled particles based on their weights resembles the thinning process in MCMC by discarding particles with very low weights, as well as reducing the problem of memory and storage. This resampling procedure also converts the weighted samples into equally weighted samples, which allows methods that require equal weights to be used within this context.

As demonstrated, weights can be associated with the ratio of the posterior density to the proposal density, where the posterior density is proportional to the likelihood and prior density.

In the hierarchical model context, many parameters will be involved and as a result, it may be difficult to achieve convergence in the MCMC routine. We therefore will demonstrate the use of (sequential) importance sampling in the H1N1 example in chapter 4 to propose samples based on the output from an MCMC routine which has not yet converged. Our approach was to approximate the distribution thus obtained by the product of independent normals with the mean and variance of each parameter estimated from the (unconverged) MCMC routine and used in the importance sampler.

After sampling particles from this (multivariate) normal proposal distribution and calculating the weights, the particles, together with their respective weights, will be the information used to generate the next round's multivariate normal proposal distribution, with weighted mean of the particles and weighted covariance matrix of the particles. To facilitate gradual refinement of the proposal distribution, we introduce a temperature variable, like the temperature in simulated annealing. Specifically, we introduce a temperature variable T that flattens the likelihood function on initial pilot runs. Under this scheme, the weight of the lth particle at every round, with parameter values, $\boldsymbol{\theta_l}$, and intensity, T, is

$$w_l = \frac{(f(D|\boldsymbol{\theta_l}))^T f(\boldsymbol{\theta_l})}{q(\boldsymbol{\theta_l})}. \tag{2.30}$$

Taking logarithms, we can better see how the intensity T will be relating the

likelihood to the weights

$$\log w_l \quad = \quad T \log f(D|\boldsymbol{\theta_l}) + \log f(\boldsymbol{\theta_l}) - \log q(\boldsymbol{\theta_l}). \tag{2.31}$$

In this way, the proposed values are more diffuse than they ought to be at early pilot runs, allowing the proposal distribution to be gradually improved, but as the number of rounds increases, this routine converges towards the desired posterior. Thus, the values of T will increase stepwise from a small number (we used 0.1) to 1 in each subsequent sampling rounds. Only the sample from the round with $T = 1$ are used for inference.

As shown in Figure 2.3, parameter values that do not model the data well will be represented by low weights. The few points that suit the data better will have much larger weights. The temperature starts small to prevent over-concentration of the few particles with good weights. As Sequential Importance Sampling progressed from the first round to the tenth round, the particles will gradually become a good realization of the posterior.

If we take exponential transformation to convert $\log w_l$ back to w_l for particle l, many of the values will become 0 as the $\log w_l$ is close to $-\infty$. Hence, we overcome numerical overflow issues by transforming the $\log w_l$ to

$$(\log w_l)^* \quad = \quad \log w_l - \max(\log \boldsymbol{w}) \tag{2.32}$$

before exponentiating to get

$$w_l^* \quad = \quad \exp (\log w_l)^* \tag{2.33}$$

and rescaling all w_l^* by

$$w_l \quad = \quad \frac{w_l^*}{\sum w_l^*} \tag{2.34}$$

so that all the weights, w_l, sum to 1.

This approach of flattening the likelihood and gradually returning it to its original shape, as we develop a better idea of the posterior to sample from, is useful for complex models like those discussed in this thesis where MCMC converge is problematic. In chapter 4, we will use this technique to explore the high dimensional parameter space in the H1N1 setting.

Chapter 3

Dengue and Chikungunya Infections in Tan Tock Seng Hospital

Unlike many other infectious diseases which spread directly from infectious host to another susceptible individual, both Dengue and Chikungunya are infectious diseases that transmit through the bites of infected female *Aedes* mosquitoes (WHO, 2013). Dengue infections can evolve into a more severe and life threatening condition, named Dengue Hemorrhagic Fever. While Chikungunya is rarely life threatening, it has long-term sequellae, and although caused by distinct viruses, the symptomology of the two diseases is similar and they may be mistaken for one another in places where both are endemic.

According to the World Health Organization (WHO) (2013), nearly half of the world population is now at risk of dengue infection, and there is no treatment available for the infected individuals, although clinical trials are ongoing (Debing, Jochmans, & Neyts, 2013). As the WHO claim that giving appropriate medical care can reduce the fatality rate of dengue fever to less than 1%, it is important to be able to develop ways to identify the infection type as quickly as possible (WHO, 2013), particularly in low resource settings. We will deal with this identification problem by developing a Bayesian logistic regression model.

While the aetiological agent can often be classified using highly accurate diagnostic tests (Lee et al., 2012), in some settings such resources might not be available

or affordable. The Times of India (2010) report that, in India, the gold standard test (reverse transcriptome polymerase chain reaction [RT-PCR]) can cost up to 5000 Rupee (approximately 80 USD), pricing it out of the means of most Indians. Therefore if an accurate diagnosis could be derived through symptomatic observations or simple clinical tests, it could be very beneficial in such settings.

Even after diagnosis, patients also need to be monitored over the time course of illness. Knowing the typical temporal trend of the infection can allow the attending physician to ascertain changes and anticipate behaviours of the patients' symptoms during the time course of illness. It might also prevent misdiagnoses. To this end, a hierarchical, temporal model of various clinical and laboratory characteristics will be developed.

3.1 Bayesian logistic regression

This project involved 117 individuals diagnosed with Chikungunya and 917 other Dengue Fever (DF) patients, including 55 individuals who had Dengue Hemorrhagic Fever (DHF) (Lee et al., 2012). The symptoms are very similar, including the sudden onset of an influenza-like illness with fever, muscle pain, headache and rashes, but Chikungunya can cause joint pains that can continue for months (WHO, 2008). We use the data collected from these observations on patients suffering from dengue or Chikungunya when they presented at the hospital to develop a model to predict what disease/infection the patient has upon admission to hospital, based on routine data on their symptoms and simple laboratory tests that are available on the day of admission itself.

These retrospective observations for the Chikungunya were made on hospitalised patients from Tan Tock Seng Hospital, Singapore, during the dengue outbreak in August 2008, while those observations for Dengue Fever were made on hospitalised patients from the same location during the 2004 dengue outbreak (Lee et al., 2012). The individuals were identified by reverse transcription-polymerase chain reaction (RT-PCR) and their demographic, epidemiological, serial clinical and laboratory, radiological, treatment and outcome data were collected but not recorded together with the patient's name for privacy issues (Lee et al., 2012).

Our exploration of the demographic and clinical factors associated with Dengue

Fever (DF) or Chikungunya infection using classical multivariate logistic regression has previously been published (Lee et al., 2012). In this thesis, we extend the analysis to use a Bayesian approach.

3.1.1 Data Processing

The variables at our disposal in the multivariate logistic regression model are: age, gender, hypertension, time since onset (in days), duration of fever (in days), presence of fever, headache, myalgia/athralgia, rash, any bleeding, sore throat, cough, nausea, vomiting, diarrhea, abdominal pain, anorexia, maximum temperature (°C), tachycardia (pulse >100/minute), leukocyte count, hemoglobin, serum hematocrit, platelet count, lymphocyte proportion, serum sodium, potassium, urea, creatinine, bilirubin, alanine (ALT) and aspartate aminotransferase (AST), alkaline phosphatase (ALP), protein and albumin, as measured on the day of hospital presentation (Lee et al., 2012). Because it involved retrospective chart review, some of these variables were missing for some patients.

Usually, when faced with missing entries, those individuals' data would be removed from the analysis. Out of the 917 dengue patients, about 2.5% (23 patients) have missing entries for the Hematocrit observations. But useful insights can be derived from the entries of other variables, and such cases may arise in regular clinical management of patients for whom a diagnosis is still required, so instead of discarding such individuals, we replace missing values with the imputed value (Lee et al., 2012): the mean for continuous variables and 0 for dichotomous variables coded 0 for absent and 1 for present. This will ensure the other variables of those individuals continue to give information but the imputed values will not distort the information from the available data. The side effect of this replacement is an unwanted reduction of the standard error which is indirectly proportional to the sample size (Cohen & Cohen, 1984).

Another important problem realised in this project is termed *separation* exhibited by the independent variables (Heinze & Ploner, 2003; Shen & Gao, 2008). Separation occurs when the binary outcome (Chikungunya or DF) can be perfectly separated by a single covariate or combination of several covariates. Quasi-Complete Separation is a less extreme case, which occurs when some values of the binary outcome

(Chikungunya or DF) overlap at a single covariate or several covariates. An example is depicted in the following figure.

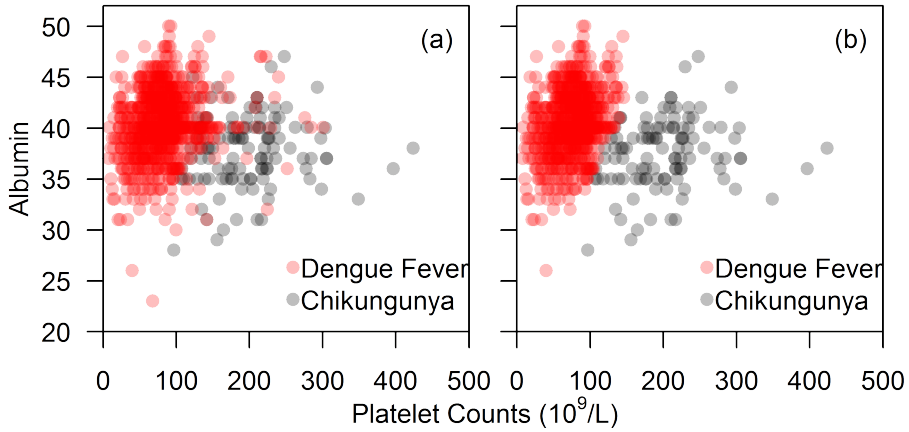

Figure 3.1: **Differentiating a Quasi-Complete Separation (in panel (a)) from a Complete Separation case (in panel (b)).**
We artificially removed overlapping points which amounts to about 8% (78 out of 979 patients) of the total to achieve the complete separation in panel (b). Both plots show two obvious regions on the plot of Albumin (g/L) versus Platelet counts (10^9/L) that can differentiate the two different diseases.

Figure 3.1 shows that both Chikungunya and DF patients exhibit a similar albumin level, but they could be easily distinguishable as suffering from DF if a patient is having low platelet counts ($< 100 \times 10^9$/L) and Chikungunya otherwise. Specifically, if the patient's Albumin level in g/L is more than $0.15 \times$Platelet Counts($\times 10^9$/L)$+20$, they are very likely to be infected with Dengue Fever. Although the plots show that these two variables are highly predictive, there is no finite MLE in a logistic regression model that uses them as predictors. In simple logistic regression, the estimated coefficients are the values that will maximise the likelihood. The algorithm used in the R statistical environment uses a Newton-Raphson approach to search for the coefficient. But when there is a Separation problem, no finite maximum likelihood estimates exist (Heinze & Schemper, 2002). The Newton-Raphson method will stop at the wrong parameter value when it has exhausted the maximum number of iterations or when the difference in log-likelihoods is smaller than a threshold and report a nonsensical estimate: an odds ratio that is too large, a standard error that is even more too large, and a p-value that is non-significant despite the obvious wealth of information (see figure 3.1). To visualise the problem, we present in table 3.1 estimates for two near-identical datasets (illustrated in figure 3.1) with a small number

(8%, or 78 out of 979) of patients removed from the second. The difference in the estimates caused by Separation is stark.

	Platelets Counts			
Data	coeff.	OR	Std. Err	p-value
Panel (a)	0.04	1.04	0.003	0
Panel (b)	45	4.2×10^{19}	670	0.95
	Albumin			
Data	coeff.	OR	Std. Err	p-value
Panel (a)	-0.29	0.75	0.045	0
Panel (b)	-301	1.6×10^{-131}	4 500	0.95

Table 3.1: **Comparison of the MLE of logistic regressions of Platelet and Albumin for the dataset in figure 3.1 panel (a), before the removal of patients with overlapping case, and in panel (b), where there is Complete Separation problem.**
The estimates were attained by fitting the disease outcome to the linear predictive variables of platelet counts and albumin using a generalised linear model. The `glm` function can be found in the `stats` package in R (R Core Team, 2013).

As demonstrated by Heinze and Schemper (2002), the separation problem often depends on the sample size. Intuitively, the smaller the collected sample, the higher the chance of having the responses separated by the independent variables. However, it is often infeasible to collect more data to resolve what is really a statistical, not a data, problem. The risk of observing separation also increases with the number of independent variables (Heinze & Schemper, 2002).

One solution to the separation problem is combining classes of categorical variables, like classifying ethnicity into four groups—Chinese, Malay, Indian and Others (Heinze & Schemper, 2002). Alternatively, continuous variables can be structured into categorical variables which may rectify observed separation.

Another, sadly common, solution for the separation problem is to exclude the variable responsible. This is unfortunate as the variable itself is typically highly predictive of the outcome and so discarding it reduces the predictive power. In the current context, this would lead to more misdiagnoses.

A more satisfying alternative is to used a form of penalised regression (Heinze and Schemper (2002) recommend Firth's penalised likelihood method (Firth, 1993), to estimate adjusted odds ratios with reduced bias relative to maximum likelihood estimation. If using Firth's approach, one can derive p-values and confidence intervals using the profile-penalised likelihood function that could be found from the algorithm

of Venzon and Moolgavkar (1988) and the `logistf` package (Heinze, Ploner, Dunkler, & Southworth, 2013) in the R statistical environment (R Core Team, 2013).

In particular, Firth (1993) proposed an approach that yields parameter estimates by reducing the score function, $U(\theta)$, which is the gradient of the log likelihood, to $U^*(\theta)$ and solving for $U^*(\theta) = 0$. The modified score function,

$$U^*(\theta) \quad = \quad U(\theta) - i(\theta)b(\theta), \tag{3.1}$$

is the reduction of the original score function, $U(\theta)$, by the product of the gradient of the score function, $i(\theta)$, and a bias term, $b(\theta)$, that depends on the model. This modification of the score function will lead to a modified estimate, θ^*, that satisfies $U^*(\theta^*) = 0$. In our paper, we applied Firth's logistic regression to solve the Separation problem without throwing away potentially precious information that was collected during the study (Lee et al., 2012).

After solving the problem of non-existent MLE, we need to find a way to identify the significant variables in the model selection step. Two common methods for determining the significant variables are forward and backward stepwise model selection. The former starts off with the simplest model with no variables and adds one variable, which provides the most information to the model, at a time until no other variables can improve the model (Pasha, 2002). On the other hand, backward selection starts off with all possible variables and removes the variable with least benefit to the model until the best model is achieved (Pasha, 2002).

In our paper, the backward approach is used by including all the variables in the logistic regression model at the first step. The variable which corresponds to the maximum p-value of all parameters was removed one at a time from the model until all the p-values were below the level $\alpha = 0.05$. The remaining variables are deemed statistically significant; they are the duration of illness, duration of fever, whether there is fever at presentation of illness, any bleeding and platelet counts (Lee et al., 2012). The variable of platelet counts was initially spotted to exhibit Quasi-Complete Separation, which led to our adopting a penalised log likelihood approach to avoid losing valuable information for classifying patients' risk.

To demonstrate the effects of having the unnecessary, non-significant variables in the model, we will model the multivariate logistic regression with significant variables found in the paper (Lee et al., 2012), as well as some other non-significant variables

with different prior distributions for the regression coefficients within a Bayesian framework.

3.1.2 Methods

The number of significant variables for the multivariate logistic regression of Chikungunya versus DHF and Chikungunya against DF were 5 and 16, respectively (Lee et al., 2012), and for this thesis, because of the greater danger attributable to DHF infections, we focus on predicting whether a patient has Chikungunya versus DHF. A multivariate logistic regression model is fitted to the data, using variables determined to be significant in the paper by Lee et al. (2012), as well as the last 5 non-significant variables that were removed from the model in the backward stepwise model selection, but this time using a Bayesian approach.

The response variable, Y, is binary and equals to 1 if the patient is diagnosed with Chikungunya (via RT-PCR) and 0 if DHF (via RT-PCR). Potential predictors are labelled as $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10})$ which represents the duration of illness, duration of fever, whether there is fever at presentation of illness, any bleeding, platelet counts, atypical lymphocytes counts, alkaline phosphatase (ALP) measurements, whether there is rashes, whether the patient is a Singaporean and whether the patient feels nausea respectively.

While this seems an odd choice, it makes considerable epidemiological sense for several reasons. The first diagnosed case of Chikungunya only occurred in Singapore in 2008 (Ng et al., 2009). Due to the lack of past exposure, the immunity of Singaporeans against Chikungunya would be lower than those foreigners who come from countries where Chikungunya is endemic. In addition, foreign patients are more likely outdoor workers who are more likely to be bitten by *Aedes albopictus*, the primary vector for Chikungunya. Thus, citizenship will indirectly have an effect on the probability of identifying whether the patient is infected with Chikungunya or Dengue.

The model for the jth individual is as follows:

$$\text{logit}(\Pr(Y_j = 1)) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_{10} x_{10j}. \tag{3.2}$$

The odds of Chikungunya (versus DHF) are

$$\frac{\Pr(Y_j = 1)}{1 - \Pr(Y_j = 1)} = \exp(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_{10} x_{10j}). \tag{3.3}$$

Thus, the probability of individual j having Chikungunya conditional on the model is

$$\Pr(Y_j = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_{10} x_{10j}))}. \quad (3.4)$$

There are five binary and five discrete independent variables. Because the range of these variables differ, it did not seem appropriate to use the same prior for the parameters $(\beta_0, \beta_1, \beta_2, \ldots, \beta_{10})$. Instead, we standardised all the covariates.

If x_{ij} is the ith covariate for the jth individual, then $z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$ is the standardised value, where $\bar{x}_i$ and s_i are the mean and standard deviation of the ith covariate. This ensures that the estimated coefficients are the overall strength of the relationship between the predictors and the response variable. The model for the jth individual is changed to the following:

$$\text{logit}(\Pr(Y_j = 1)) = b_0 + b_1 z_{1j} + b_2 z_{2j} + \ldots + b_{10} z_{10j} \quad (3.5)$$

where b_i is the new regression coefficient after the standardization.

Since regression coefficients can be positive or negative depending on the relationship between the predictor variable and the response variable, a Laplace prior distribution (double exponential distribution) centred at 0 (i.e. $b_i \sim \text{Laplace}(0, \lambda)$ for $i = 0, 1, 2, \ldots, 10$) was chosen to allow the regression coefficients to take any real numbers. The probability density function can be represented by

$$f(b_i|\lambda) = \frac{1}{2\lambda} \exp(-\frac{|b_i|}{\lambda}) \quad (3.6)$$

where the mean and variance are 0 and $2\lambda^2$ respectively.

Using a Laplace prior distribution acts as a penalizing procedure, like that of the LASSO estimator. The LASSO estimator can be expressed as

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \log L(D|\mathbf{b}) \text{ subjected to } \|\mathbf{b}\|_1 \leq c \quad (3.7)$$

$$= \arg\max_{\mathbf{b}} \log L(D|\mathbf{b}) - \lambda \|\mathbf{b}\|_1, \lambda \geq 0. \quad (3.8)$$

where $L(D|\mathbf{b})$ is the likelihood density, $\|\mathbf{b}\|_1 = \sum |b_i|$, and λ is the optimal penalty.

We can also represent the logarithm of posterior density by

$$\log \Pr(\mathbf{b}|D) = c + \log L(D|\mathbf{b}) + \log f(\mathbf{b}) \quad (3.9)$$

$$= c' + \log L(D|\mathbf{b}) - \frac{\sum |b_i|}{\lambda}. \quad (3.10)$$

where $f(\mathbf{b})$ is the Laplace prior density. With the similarity in the two methods, imposing a Laplace prior distribution has the effect of penalizing the estimating process. In a Bayesian framework, we will be characterising $\log \Pr(\mathbf{b}|D)$ instead of finding the value of $\hat{\mathbf{b}}$.

Other than using a Laplace prior distribution to mimic a penalizing process, a normal prior distribution with zero mean could also be used. This is similar to a ridge estimator which is represented by

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \log L(D|\mathbf{b}) \text{ subjected to } \|\mathbf{b}\|_2^2 \leq c \tag{3.11}$$

$$= \arg\max_{\mathbf{b}} \log L(D|\mathbf{b}) - \lambda \|\mathbf{b}\|_2^2, \lambda \geq 0. \tag{3.12}$$

where $\|\mathbf{b}\|_2^2 = \sqrt{\sum b_i^2}$. Correspondingly, we can present the logarithm of posterior density with a normal prior distribution with zero mean and variance σ^2 as

$$\log \Pr(\mathbf{b}|D) = c' + \log L(D|\mathbf{b}) - \frac{\sum b_i^2}{2\sigma^2}. \tag{3.13}$$

Laplace distributions have fatter tails than normal distributions with the same mean and variance. We need a prior distribution which does not overly favor regression coefficients values which is close to zero and at the same time allow deviation to both ends of the real numbers. If the regression coefficients are allowed to take the appropriate numbers, the probability of getting a correct prediction from the logistic regression will increase.

To show the different swamping effects in informative and non-informative prior distributions, the scale parameter, λ, of the Laplace prior distribution is allowed to vary in the parameter model.

To illustrate the results of using different prior distributions on the Separation issue, repeated Bayesian logistic regression Models with different values of the variance parameter, λ, are explored.

The algorithm uses MCMC to do so, as follows:

1. Standardise the dataset for every covariate i and individual j by

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}. \tag{3.14}$$

2. In the model specification, the disease type for jth individual will follow a Bernoulli distribution with success probability

$$\Pr(Y_j = 1) = \frac{1}{1 + \exp(-(b_0 + b_1 z_{1j} + b_2 z_{2j} + \ldots + b_{10} z_{10j}))}. \quad (3.15)$$

3. The posterior distribution of the parameters is sampled using JAGS (Just Another Gibbs Sampler) using the precision in the specification of the normal distribution, instead of the usual standard deviation or variance (Plummer, 2013). Fixing the prior distribution scale parameter, λ, say to 0.001, we use JAGS to get a posterior sample of b_i, using 90% of the data, over four chains with 1000 burn-in and 2500 iterations each. The training set of 90% of the data is randomly chosen based on the index of the patient using the `sample` function in R (R Core Team, 2013).

4. For each individual from the remaining 10% data, we estimate the probability of getting Chikungunya by

$$\hat{\Pr}(Y_j = 1) = \frac{1}{1 + \exp(-(\hat{b}_0 + \hat{b}_1 z_{1j} + \hat{b}_2 z_{2j} + \cdots + \hat{b}_{10} z_{10j}))} \quad (3.16)$$

where a set of regression coefficients, $\hat{\mathbf{b}} = \left(\hat{b}_0, \hat{b}_1, \hat{b}_2, \cdots, \hat{b}_{10}\right)$, are randomly sampled from the posterior sample for each individual.

5. The posterior predictive probability of correctly forecasting the aetiology is

$$p = \left[\hat{\Pr}(Y_j = 1)\right]^{y_j} \left[1 - \hat{\Pr}(Y_j = 1)\right]^{1-y_j}. \quad (3.17)$$

The score for this λ, the scale parameter of the Laplace prior distribution which we have fixed in Step 3, is obtained by taking the mean of these probability values from the 10% data.

6. The mean of each b_i, for $i = 0, 1, 2, \ldots, 10$, from the posterior sample are recorded.

7. Repeat step 3–5 for the same λ ten times, using a different set of 90% of the data in each round by re-sampling again based on the patients' index. The mean of all the ten scores in step 5 for that particular λ is taken to be the mean.

8. Then, repeat step 3–7 with different λ taking values from 0.001 to 0.1 with thinner spacing for smaller λ.

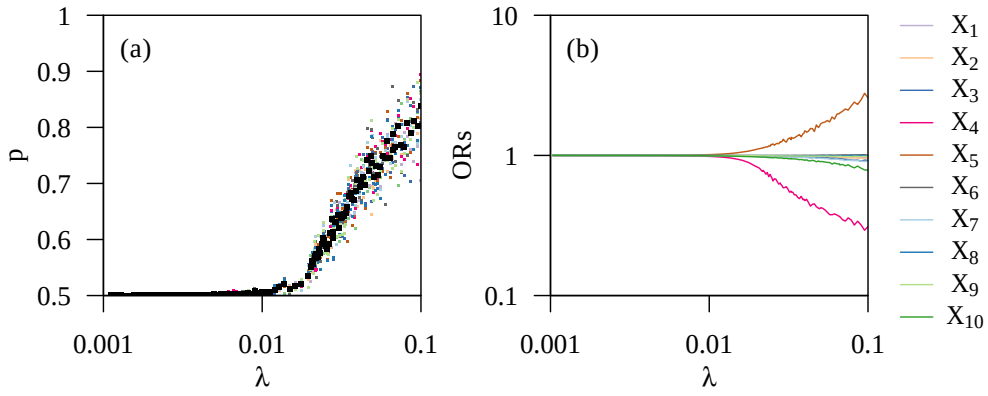

Figure 3.2: **Probability of correct diagnosis and odds ratio for different λ.** Panel (a) shows the mean of the probability of making a correct diagnosis, p, for different λ, represented by tiny, coloured points, for ten different runs. The bigger, black points are the mean of the ten values for each λ value. Panel (b) shows the mean odds ratios, $\exp(b_i)$ of all the 10 variables, X_i for $i = 1, 2, \ldots, 10$, for different λ.

As λ increases, the probability of making a correct diagnosis increases, implying a more predictive logistic regression model. For large λ, the prior distribution for b_i is effectively flat. Because a flat prior does not give particular weights to any values, the estimates of the parameters b_i are governed by the data. From the second panel, it further confirms that small λ will only make incorrect focus of b_i near 0, which will result in more failed diagnoses. Choosing the right prior distribution can help focus the posterior to suitable values while an ill-suited prior may reduce the accuracy of the model fit or forecast. Thus, when there is no prior belief for the parameter, a flat prior should be adopted.

Since the regression coefficients are still adjusting themselves to take larger positive values and smaller negative values (which corresponds to larger odds ratio and close to zero odds ratio) when the prior distribution permits in the second panel of figure 3.2, the algorithm stated above is done again with λ taking values from 0.001 to 100.

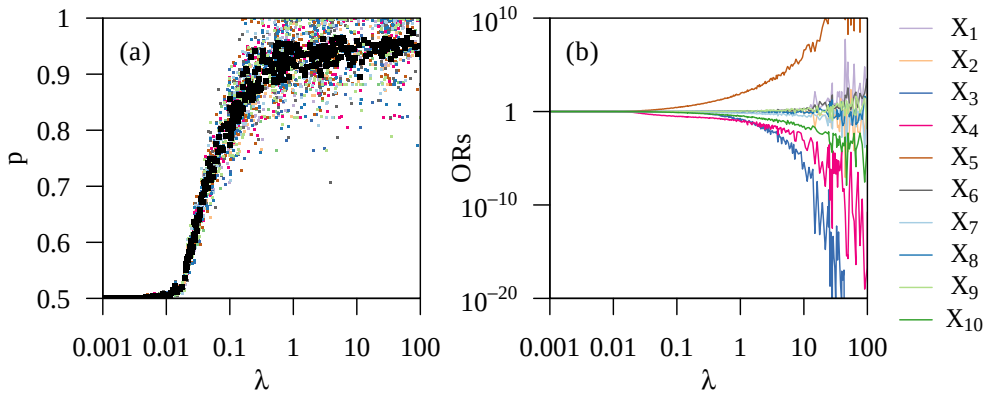

Figure 3.3: **Probability of correct diagnosis and odds ratio for different λ.** This figure has the same features as in figure 3.2 where λ is extended to 100.

In the first panel of figure 3.3, the probability of making a correct diagnosis, p, asymptotes between 0.9 and 1, indicating that the regression model cannot perform any better for even larger values of λ. The odds ratio for the binary variable of whether the patient is suffering from fever when presented with the illness, X_3, and whether the patient has any bleeding, X_4, are still decreasing as presented in the second panel. On the other hand, the odds ratio for the platelet counts, X_5, is still increasing. But the model's diagnostic ability is still good for these data implying that the effects of these deviating values cancel out and they will still continue to grow if larger λ is used as they will give unnecessary weights to these extreme values.

We observe changes in 'swamping'—the effect that occurs when the posterior distribution is mostly driven by the data, and not the prior—as λ varies. In this case, the predictive accuracy of the model is not influenced when the prior standard deviation λ increases from 10 to 100, but when λ is smaller than 10, different degrees of swamping are observed. If the prior distribution is too narrowly focused on 0, the data are unable to swamp the prior, and the effects are worse predictions.

3.2 Hierarchical Modelling of Disease Time Course

Dengue Hemorrhagic Fever (DHF) and Chikungunya (Chik) have similar symptoms and the diagnosis is expensive in locations where they cocirculate. It would be beneficial if the symptoms or laboratory observations can be modelled with time from the onset of illness. By modelling the disease time course for the symptoms or laboratory observations, the daily progression can help to guide the clinical management

for accurate diagnosis of different type of patients during their course of illness.

The hierarchical model accounts for the correlations within each patient's observation. Here, the chosen variables are Haematocrit, Platelet Counts, Leukocytes and the patient's temperature. These measurements come from the same dataset. The choice of these 4 variables was due to their clinical relevance: Haematocrit, Platelet Counts and Leukocytes are major components of blood. Daily observations of these variables are easily and readily obtainable through blood tests, while temperature taking is routine in clinical care for Dengue and Chikungunya patients.

3.2.1 Model and Method

The model was fitted separately for each type of observation and disease. For notational brevity, the measurement and virus are suppressed from the notation that follows.

We assume that these observations are conditionally independent and follow normal distributions:

$$y_{ij} \sim N(\mu_{ij}, \sigma^2) \tag{3.18}$$

where y_{ij} is the observation for patient i on day j, where day 0 corresponds to the day of symptom onset, and y_{ij} may not be completely observed over the whole observation period.

The observations have variance σ^2 which neither changes over time nor differs for each individual.

The mean observations represented by μ_{ij} will have both time effects and random effects. Correlations over time will be induced by the choice of prior. Details of μ_{ij} will be described below.

The unknown value of b_0 is given a flat, normal prior distribution with large standard deviation to accommodate different observation types,

$$b_0 \sim N(0, 100^2). \tag{3.19}$$

We believe, biologically, that b_j for day $j > 0$ will be dependent on the previous day $j-1$. To force this, we set the prior for b_j to depend on the previous time point b_{j-1} as the mean of its normal distribution,

$$b_j \sim N(b_{j-1}, \sigma_b^2). \tag{3.20}$$

We expect each individual to have a similar values to the mean of all patients' observations on each day of their time course. Differences between individuals are characterised via a random effect parameter, β_i. If the individual i had a larger observation than the others on day j, the mean observation μ_{ij} will be an amplification of the value of b_j if $e^{\beta_i} > 1$ or a reduction if $0 < e^{\beta_i} < 1$,

$$\mu_{ij} = e^{\beta_i} b_j. \tag{3.21}$$

Since there is no existing knowledge of whether each individual i should have observations greater or smaller than others, a normal prior centred at zero is used for β_i for equal chances of getting $e^{\beta_i} > 1$ and $0 < e^{\beta_i} < 1$,

$$\beta_i \sim N(0, \sigma_\beta^2). \tag{3.22}$$

This project has been done using JAGS (Just Another Gibbs Sampler) where the precision, σ^{-2}, is used in the specification of normal distribution, instead of the usual standard deviation σ (Plummer, 2013). The prior distribution for all the precision values in the normal distribution should be positive and focus near to 0. The choice used here is gamma distribution with both the shape and scale parameters taking small value,

$$\sigma^{-2}, \sigma_\beta^{-2}, \sigma_b^{-2} \sim \Gamma(0.01, 0.01). \tag{3.23}$$

The final model is thus

$$\begin{aligned}
y_{ij} &\sim N(\mu_{ij}, \sigma^2) && (3.24)\\
\mu_{ij} &= e^{\beta_i} b_j && (3.25)\\
b_0 &\sim N(0, 100^2) && (3.26)\\
b_j &\sim N(b_{j-1}, \sigma_b^2) && (3.27)\\
\beta_i &\sim N(0, \sigma_\beta^2) && (3.28)\\
\sigma^{-2}, \sigma_\beta^{-2}, \sigma_b^{-2} &\sim \Gamma(0.01, 0.01). && (3.29)
\end{aligned}$$

The algorithm is thus:

1. Prepare the data and initialisation for running the MCMC in JAGS within the R platform (Plummer, 2013).

2. After the JAGS process, the posterior sample for the parameters mentioned in the model can be used to get the 95% prediction interval for a certain observation.

3. Gather the parameter values from the posterior sample, including b_j, β_i, σ^2.

4. Compute the μ_{ij} for each set of posterior samples using the relation $\mu_{ij} = e^{\beta_i} b_j$.

5. $\bar{\mu}_j$, the daily mean of all the μ_{ij}s was computed and plotted in Figure 3.4.

6. The 95% credible interval for μ_{ij} will be presented in Figure 3.4.

7. Simulate the predicted observation values $\hat{y}_{ij}$ from normal distribution with mean μ_{ij} and variance σ^2.

8. With all the simulated values of $\hat{y}_{ij}$ as prediction, the 95% prediction interval is obtained by calculating the 2.5% and 97.5% quantiles for the simulated values at each day and plotted in Figure 3.4.

9. Step 1 to 8 should be repeated for the four chosen symptoms and observations for each of the two diseases.

3.2.2 Results and Inference

The algorithm was executed and the results is shown in Figure 3.4 for both diseases and the four chosen measurements.

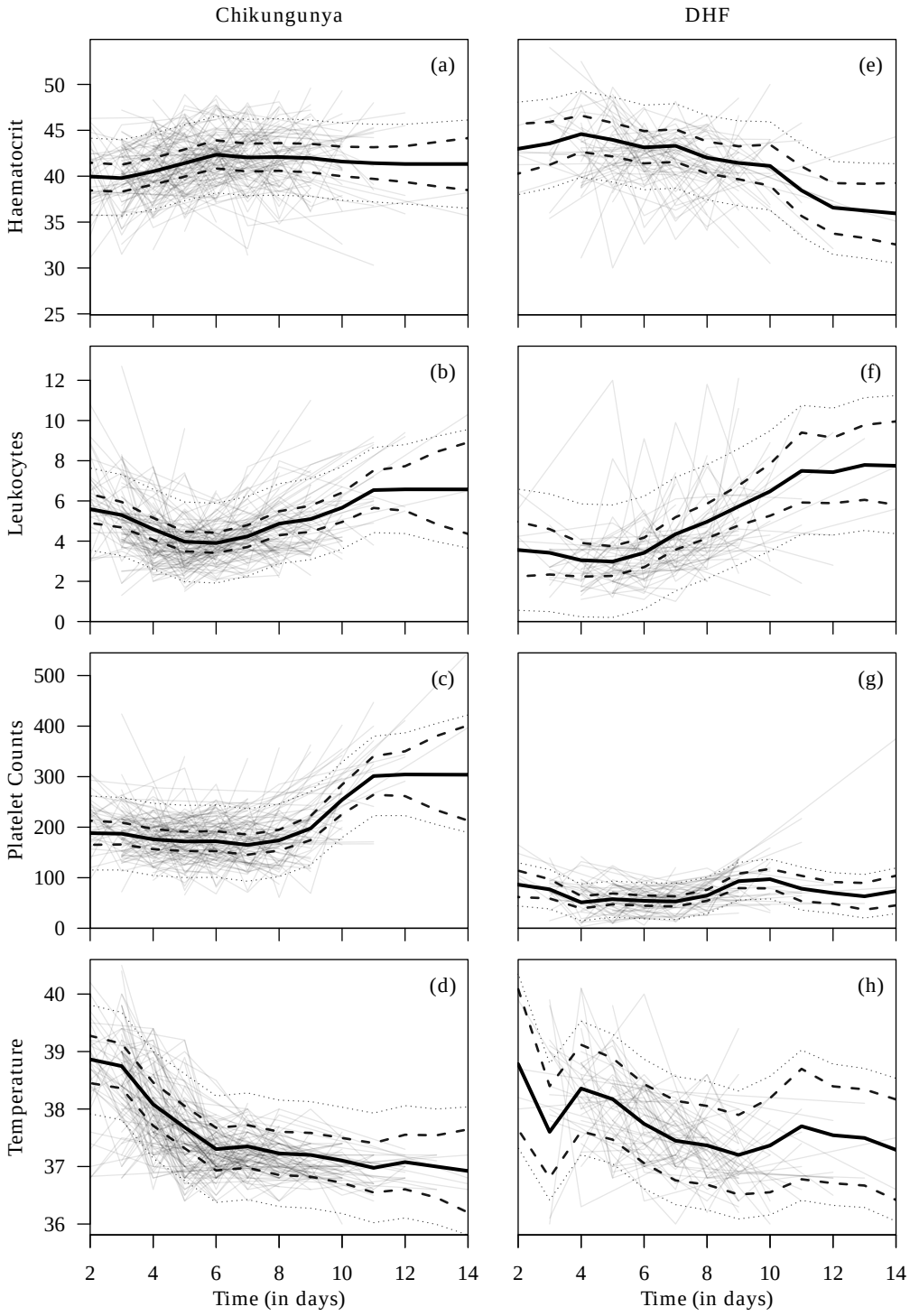

Figure 3.4: **Prediction for time course of the four selected variables, namely Haematocrit (in volume percentage) (Panel (a) & (e)), Leukocytes (in volume percentage) (Panel (b) & (f)), Platelet Counts (in $\times 10^9/L$) (Panel (c) & (g)), and the patient's temperature (in $^\circ C$) (Panel (d) & (h)) for Chikungunya and DHF respectively.**
The actual observations of the patients over a period of two weeks were plotted as light grey lines. The black solid lines show how the mean observations $\bar{\mu}_j$ of patients changes along day j; the black dashed lines show the credible interval for the mean μ_{ij}. The black dotted line shows the credible interval for the predicted observations $\hat{y}_{ij}$.

According to Nuraini (2012), haematocrit concentration will increase for dengue infected diseases and decrease to normal level of 40–50% for males and 37–47% for females upon recovery. In figure 3.4, panel (a) and (e) showed increment in haematocrit concentration at the start of the time course and a decline near to the end of observation for both diseases. The decrement in haematocrit concentration for DHF patients is more obvious than that for Chikungunya patients.

Leukocytes values are predicted to dip faster in DHF patients than Chikungunya patients (Lee et al., 2012). The normal percentage of leukocytes is 1% and figure 3.4 shows that patients of both diseases had leukocytes more than 1%, a response to the immune systems defending the body from the disease (Alberts et al., 2002). We can see that the mean observation $\bar{\mu}_2$ at day 2 dropped quickly until $\bar{\mu}_6$ for Chikungunya in panel (b) whereas that for DHF in panel (f) did not decrease that sharply. However, after the drop, leukocytes values for Chikungunya patients did not bounce back as quickly as how the DHF patients would have recovered.

Platelet counts are inversely related to Haematocrit concentration (Nuraini & Tasman, 2012).The predicted platelet counts decrease initially and rise towards the end of the observation window for both diseases which agrees with the claim by Nuraini (2012) for platelet counts to be in opposite direction of haematocrit concentration. This time course analysis, presented in Figure 3.4 panel (c) and (g), supports platelet count as the main variable for differentiating Chikungunya and DHF, as the average platelet count barely dropped below $200 \times 10^9/L$ in Chikungunya patients, but fell below $100 \times 10^9/L$ in DHF patients.

In both diseases, temperature is anticipated to reduce and asymptote at normal human temperature of slightly less than 37°C after day 2. As the temperature data collected from the Chikungunya patients are more coherent, the credible interval for mean observation $\bar{\mu}_j$ is narrower as portrayed in panel (d). The temperatures of DHF patients are more unsteady, leading to a wider credible interval for the mean temperature $\bar{\mu}_j$.

3.3 Conclusion

In the analysis of the clinical and laboratory predictors of Dengue and Chikungunya Disease, the main issue was the Quasi-Complete Separation problem of the data. In

the publication, we explored the use of Firth's penalised likelihood method with the logistic regression to overcome separation (Heinze et al., 2013).

The alternative described herein is a Bayesian analogue, in which we have shown the importance of using appropriate prior distributions for better performance in predicting the correct type of disease.

Data collection is incomplete because data are collected based on clinical need rather than for statistical purposes. As a result there are not always daily observations for each patient. The advantage of our hierarchical model is being able to set the course observation trend with missing observations from certain days by borrowing information from the other patients. This method of putting a hierarchical model in time course will, we hope, guide clinical management by providing daily trends for key variables for each type of illness. In the best case, the observation trend could help physicians in accurate diagnosis of the different type of patients during the course of illness and detect aberrant patterns that may indicate the patient's condition has changed unexpectedly.

The observations described in the hierarchical time course model were the major indicators of Dengue or Chikungunya diseases. Future work could use other observations to identify trends of the different observations.

Chapter 4

Hierarchical Model of 2009 Pandemic H1N1 Transmission

4.1 Introduction

In this increasingly globalised world, the volume of people traveling across borders allows pathogens to spread rapidly from their place of emergence to all corners of the world. When a new virus emerges and spreads to multiple countries, the World Health Organization (WHO) will declare pandemic and differentiate the seriousness with different stages (WHO, 2010).

Countries are mandated to have pandemic preparedness plans which detail their own policies and measures to deal with the pandemic (Poggensee et al., 2010; Ujike et al., 2011; Fuhrman et al., 2011). If the pandemic predictions can be improved, these plans can be tailored to the predicted severity, and better decisions may prevent unnecessary interventions while minimizing economic losses and morbidity and mortality.

The first new pandemic of the 21st century was announced in early 2009, an influenza A virus. The impact of this pandemic virus was so great that, in less than a year, it resulted in more than 15000 confirmed deaths worldwide (Halder, Kelso, & Milne, 2010).

Mexico was the first country to confirm cases of a novel variant of H1N1 (Trifonov, Khiabanian, & Rabadan, 2009), thought to be a recombinant version of viruses circulating in swine and birds (Neumann, Noda, & Kawaoka, 2009), in April of that year.

The Mexican Ministry of Health discontinued all schooling to counteract the spread of disease by reducing the contacts amongst the younger generations (Chowell et al., 2011), who constituted the majority of both cases and confirmed deaths in the early stages of the pandemic (Domínguez-Cherit et al., 2009). Further control measures to minimise physical interaction included the closure of movie theaters, restaurants, and other public assemblage locations (Chowell et al., 2011).

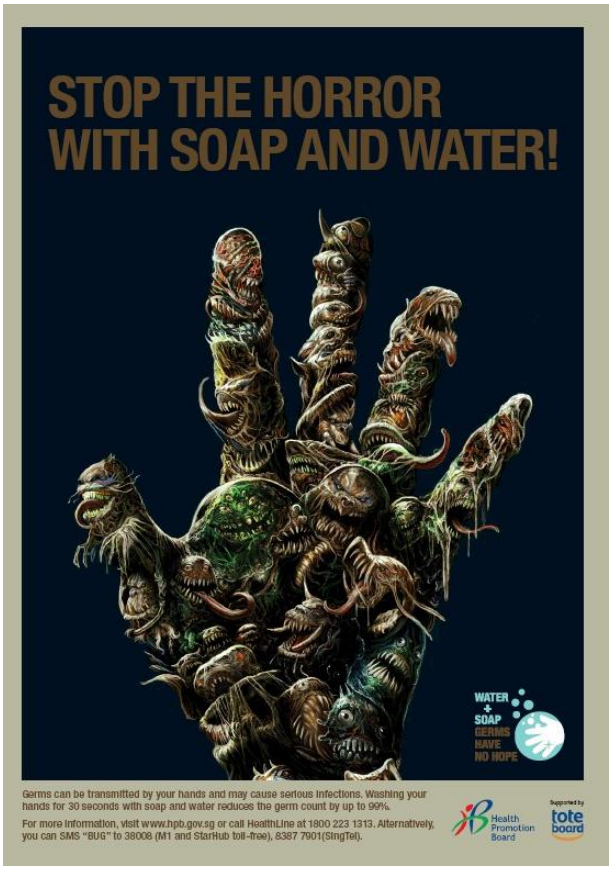

Figure 4.1: **Singapore's Health Promotion Board promotion poster.** This poster aims to inform the public about germs being transmitted by hand may cause serious infection.

Other countries took different approaches to contain or mitigate the pandemic. For instance, in Germany, improved sanitation methods were widely proposed, including public health education on the correct way of hand cleaning (see for example Singapore's Health Promotion Board promotion poster reproduced in figure 4.1), the use of face masks and the isolation of possible infected individuals to abate the circulation of the disease (Poggensee et al., 2010), with claims that carrying out these non-pharmaceutical interventions within 36 hours of symptom onset can reduce the rate of diseases spread (Poggensee et al., 2010). In Japan, according to Ujike et al.

(2011), the inventory of Oseltamivir during the pandemic was doubled. In France, during the initial phase of the pandemic from 1 May 2009 to 30 June 2009, all verified and possible cases of H1N1 were hospitalised, regardless of the degree of seriousness of illness (Fuhrman et al., 2011). This extreme prevention during the containment period resulted in the unsustainable rising need for hospitalization, requiring this policy be later amended so that only patients in a critical condition be monitored in hospitals (Fuhrman et al., 2011).

In contrast to past influenza pandemics, but similar to the situation following the SARS outbreak of 2003 (Naylor, Chantler, & Griffiths, 2004), attributes of the H1N1 influenza virus from different countries were shared quickly worldwide for the speedy development of a vaccine and identification of the viral strain (Ikonen et al., 2010). However, the downside to this was the initial panic caused by the unduly pessimistic estimates of the severity of the virus from Mexico (Goodwin, Haque, Neto, & Myers, 2009). Being able to merge, appropriately, data from multiple countries would allow better decision making while overcoming the weaknesses or gaps in individual countries' surveillance data, reflecting differences in the extensiveness of localized data.

The virulence of the pathogen, measured by the rate of infection and removal or the reproduction number, should be similar in different parts of the world, and although the pandemic virus might mutate during the outbreak, as it did in the 1918 pandemic (Taubenberger & Morens, 2006), we might assume that the viral characteristic remain unchanged over the first wave of the outbreak, when the data paucity is most severe. However, fatality, confirmation and hospitalization rates will depend on countries' healthcare capacity and surveillance systems and are expected to differ.

A powerful model should address these similarities and differences. In this chapter, we will present a framework to exploit valuable information on the spread of a pandemic in different countries combining hierarchical with transmission dynamic modelling, in particular, a Susceptible-Infected-Removed (SIR) model of a homogeneously mixing population is used to model observational data collected from different countries. Demonstrating the use of this approach via data published on the H1N1 pandemic, we propose that a hypothetical network of surveillance systems could be

set up to pool data from participating countries that would provide real-time data for analysis and prediction for a pandemic outbreak.

Several platforms for reporting information on infectious diseases exist but they do not have committed organizations from worldwide to provide actual real-time observational data. The Global Public Health Intelligence Network (GPHIN) has been supplying information to WHO, international governments and non-governmental organizations since 2004, using information extracted from reports of eight different languages (Mawudeku & Blench, 2006). From 1994, Program for Monitoring Emerging Diseases (ProMed-mail) has been sourcing infectious diseases information from the grey literature—media or official reports—and disseminating the materials to subscribers by email (Victor & Madoff, 2004a). This has been useful, and it picked up the emergence of SARS before the Chinese government shared data with the rest of the world (Victor & Madoff, 2004b), but data are partial, messy, unconfirmed and have many false alarms. With added languages for sourced documents, HealthMap is able to collate information automatically in collaboration with ProMed-mail in a quicker manner since 2006 (Brownstein, Freifeld, Reis, & Mandl, 2008). However, accuracy problems may arise due to the mechanized routine for data compilation. The most promising platform is the International Severe Acute Respiratory Infection Consortium (ISARIC), an international alliance with about 50 to 60 research networks worldwide for real-time infectious diseases data sharing since 2011 (Yong, 2012). Yet, the synchronization of shared data may be a problem if data were not collected based on fixed standard criteria, and data from academic institutes may lack the completeness of national surveillance.

In the situation where many countries experience an outbreak, the experience of each will differ, with differences in importation and establishment dates (Lau et al., 2012), interventions, seasonality, and potentially severity indices. This requires being characterised by a separate parameter vector for each country, resulting in a high dimensional parameter space for exploration. To perform a model fit, Bayesian solution can provide accurate information about the parameters using methods such as the Markov Chain Monte Carlo (MCMC) algorithm. On top of this, hierarchical modelling is commonly used in non-epidemic settings but heretofore has rarely been exploited within the infectious disease setting (Kleczkowski & Gilligan, 2007). We

build a hierarchical epidemic model with hyper-parameters to account for the variability between outbreaks of different countries and estimate the parameters using MCMC and importance sampling methods. This approach is demonstrated to be very successful in pooling information across multiple countries and in characterizing the variability between outbreaks, showing that non-epidemic methodology can, with suitable adaptation and development, contribute by giving better estimations for infectious disease epidemiology.

4.2 Literature review and data sources

On 23 April 2009, the first case of H1N1 was reported to WHO by Mexico (Chang et al., 2010). H1N1 was the first virus in the 21st century that has spread to most countries in the world, causing an influenza pandemic (Chang et al., 2010). Upon contact with an infected individual, a susceptible individual may get infected, potentially leading to confirmed death in the most serious cases (Zuno et al., 2009). The main symptoms of H1N1 are fever, cough, headache, muscle aches, and rhinorrhoea (Zuno et al., 2009), i.e. symptoms that are indistinguishable from a regular 'cold'.

Typically each country collects and analyses their data in isolation. Limited data, during the start of a pandemic, often may give a misleading impression and when used in a forecasting routine might not give sensible predictions due to data incompleteness. But a hierarchical model can be formed to pool information from different sources, which may collect different types of data, such as on hospitalizations or community transmission, simultaneously. To evaluate the utility of data sharing networks for future worldwide outbreaks, H1N1 pandemic data were collected from a literature review of research publications and government surveillance websites.

Individuals affected during the pandemic can be classified in several ways according to the severity of their infection and their healthcare utilisation (see Figure 4.2).

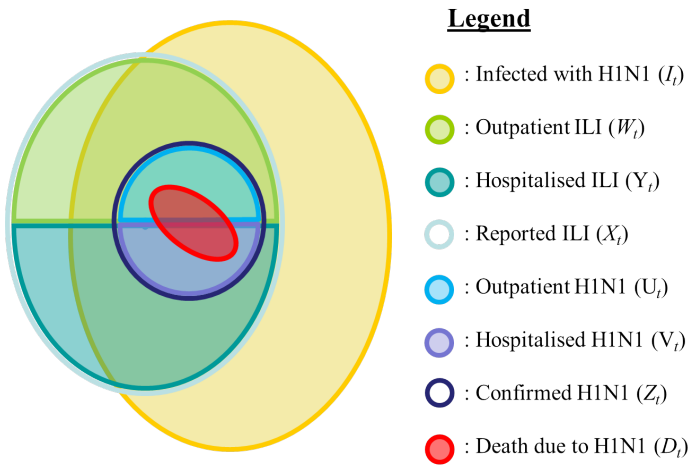

Figure 4.2: **Venn diagram for differentiating individuals at time t during the pandemic.**
The yellow oval represents all the individuals who are infected with H1N1, I_t. The aqua oval represents those who consulted a doctor and were reported to show influenza-like illnesses (ILI) symptoms, X_t, Within these doctor consultations, the outpatient ILI, W_t, is represented by the lime-green semi-oval; the hospitalised ILI, Y_t, are represented by the teal semi-oval. All patients who are confirmed to be infected with H1N1, Z_t, are represented by the indigo oval. The outpatient H1N1, U_t, is represented by the turquoise semi-oval; the hospitalised H1N1, V_t, are represented by the purple semi-oval. The red oval represents those who died due to H1N1 infection, D_t.

Not all individual infected with influenza will visit a doctor, and not all who do will be diagnosed as having an influenza infection (as opposed to another respiratory virus), and as a result it is impossible to record all the individuals infected with H1N1 at time t, I_t, or who have recovered at time t, R_t. The number of ILI, X_t, in the community is usually measured by a network of influenza sentinel clinics (Truscott et al., 2012), who can give useful insights to I_t. X_t does not in general represent the full number of ILI cases in the country or territory because the surveillance system would typically not include all the physicians in the country, but as long as the coverage does not vary over the course of an outbreak, we can justifiably assume proportionality in the model to overcome this shortcoming.

Influenza sentinel clinics may be hospital-based or clinic-based. ILI patients reported by doctors at outpatient clinics or general practitioners (GP) are represented by W_t, depicted by the lime-green semi-oval in figure 4.2. Hospitalised ILI patients are denoted as Y_t depicted by the teal semi-oval in figure 4.2. If observed for the whole population, W_t and Y_t should be subsets of X_t, but most studies were done in a limited number of hospitals, and so the numbers reported were only representative

of those hospitals. As a result, due to the imperfect data collection and the differences in the authorities' practices, the observed data may not satisfy this relationship.

We denote all confirmed H1N1 cases as Z_t, hospitalised H1N1 patients as V_t, those who are not hospitalised as U_t, and those who died due to infection as D_t. The number of confirmed confirmed deaths due to H1N1, D_t, may be an under-recording of the actual number as H1N1 might not be identified as the cause of confirmed death for some patients. Among the sources that we have explored, data on U_t and Y_t could not be found, but the remaining data provide rich information for the missing data of I_t and R_t. Notation for the states and data are provided in table 4.1.

States	Description	Availability
I_t	Individuals infected with H1N1 at time t	No
R_t	Individuals recovered or died due to H1N1 at time t	No
D_t	All confirmed confirmed deaths due to H1N1 at time t	Yes
Z_t	Patients diagnosed with H1N1 at time t	Yes
Y_t	Hospitalised ILI patients at time t	No
X_t	All reported ILI patients at time t (Hospitalised and outpatient)	Yes
W_t	Outpatient ILI patients at time t (Government clinic and GP)	Yes
V_t	Hospitalised H1N1 patients at time t	Yes
U_t	Outpatient H1N1 patients at time t	No

Table 4.1: **Summary of the different states and data used in this project.** These are the states that we have considered and the availability column indicates whether the data could be found in the literature reviews.

Differences in countries' reporting systems led to non-systematic data for the categories listed in table 4.1. We harvested data from publications from fourteen countries with twenty-five datasets, of which Z_t, the number of confirmed H1N1 cases, was the most common data type, whereas U_t, the number of outpatient H1N1, and Y_t, the number of hospitalised ILI, were not observed.

ILI counts are considered more reliable than confirmed H1N1 cases because typically they are collected under a consistent sampling protocol (Mark I-Cheng Chen, personal correspondence). For countries with a proper surveillance system, ILI data may be recorded consistently, even before an outbreak. Fox (2009) reports that there was a sudden increase in demand for influenza tests in the early stage of the pandemic, leading to more capacity to confirm suspected infections at this stage. Thus,

the number of confirmed H1N1 cases is subject to possible biases and might change with the testing paradigm according to the changing risk perceptions. An example of this was reviewed by the Australian Government Department of Health and Ageing (2011), where extensive laboratory tests were carried out at the start of the pandemic, but to be more efficient, the protocol was modified to direct the tests towards the more severe cases and vulnerable individuals to reduce the surge in demand for tests. As a result, where ILI data were available, we included those alongside H1N1 confirmations, to validate the latter.

The WHO declared the end of H1N1 pandemic on 10 August 2010 (WHO, 2010), though for most countries, the first wave—the time of greatest uncertainty—was completed by late 2009. Because our basket of countries have different seasons, and seasonality has been observed to influence the risk of influenza transmission (Balcan et al., 2009), we truncated all datasets at 1 October 2009 which marks the common change of seasons for countries in the northern and southern hemisphere. We initially attempted to factor in the seasonal effect, but this was deemed unduly complicated, as will be elaborated in the next subsection. As a result, our model focuses on the period 23 April 2009 (Day 113 of the year) to 1 October 2009 (Day 274), i.e. an interval of twenty-three weeks.

Countries were chosen such that they could form a good representation of the world. They are countries and cities from four continents, namely North America, South America, Eurasia and Australia; there is a marked paucity of data from Africa and no good sources could be found. Among these countries, there was a mix of middle and higher income countries. We will elaborate on how the data have been collected in the order of the countries' geographical latitude.

Finland: The first H1N1 confirmed cases occurred on 10 May 2009 (Ikonen et al., 2010). The National Infectious Disease Registry collated the weekly numbers of laboratory confirmed infections of the 2009 pandemic influenza A (H1N1) viruses, as reported by Ikonen et al (2010). The number of H1N1 confirmed cases was digitised using Engauge Digitiser from the article's bar chart.

England: Three data types are available, namely, the number of H1N1 confirmed deaths, the number of hospitalised H1N1 cases and the number of outpatient ILI cases. During the pandemic, McLean et al. (2010) from the Health Protection

Agency's (HPA, now Public Health England) Centre for Infection claimed that the General Registry Office of England reported daily numbers of confirmed deaths, based on laboratory confirmation or classification on the confirmed death certificate. Figure 28 of the Epidemiological Report informs about the four countries in United Kingdom (UK) but only England's number of H1N1 confirmed death was digitised (McLean et al., 2010) for consistency with the other sources for the UK.

In another HPA weekly report (McLean & Paterson, 2010), the number of hospitalised H1N1 cases and number of outpatient ILI cases for England were digitised from the same diagram in Figure 8. Under the National Laboratory Reporting Scheme, 230 National Health Service, HPA and independent sector microbiology laboratories provided data on the number of hospitalised H1N1 cases (McLean et al., 2010). The outpatient ILI cases were provided by approximately 50 physicians from the Royal College of General Practitioners (RCGP), who reported the weekly number of ILI cases to the RCGP Research and Surveillance Centre to provide the data for the HPA report regularly (McLean et al., 2010; McLean & Paterson, 2010).

France: The French GP sentinel surveillance system has been in place since 1984 to collect the number of ILI consultations and it is still ongoing (Sentinelles, 2012). The system relies on a 1300 volunteer GPs who submit weekly number of ILI consultations via secure internet connection (Sentinelles, 2012). Subsequently, average numbers were estimated at the national level. We digitised the graph presented by Fuhrman et al (2011) for the estimated number of ILI consultations.

New York: We found three datasets, including the number of H1N1 confirmed deaths, the number of hospitalised H1N1 cases and the number of reported ILI cases. The surveillance for H1N1 was in response to the expected severity of the pandemic as the Department of Health and Mental Hygiene (DOHMH) appealed for all the hospitalised H1N1 cases and confirmed deaths due to H1N1 to be reported (Lee et al., 2010; Balter, Gupta, Lim, Fu, & Perlman, 2010). The surveillance system for confirmed death due to influenza has been in place before the H1N1 pandemic but only collecting data on influenza confirmed deaths in children (Lee et al., 2010). According to Lee et al. (2010), the DOHMH removed the age criterion to make the documentation suitable for the pandemic situation, and the surveillance system was ready and credible for use. The bar chart on the number of H1N1 confirmed

death in New York was digitised from the Lee et al. (2010) paper. DOHMH also ensured the accuracy of the number of hospitalised H1N1 by following up with all the 57 hospitals in New York daily by telephone to be notified of the total number of hospitalised H1N1 patients within or outside the intensive care units (ICU) (Balter et al., 2010). The number of ILI cases for New York was available from the United States Centers for Disease Control and Prevention (CDC) website (CDC, 2009), which was collected by the U.S. Outpatient Influenza-like Illness Surveillance Network made up of more than 2,700 outpatient healthcare providers in all 50 states (CDC, 2009).

Japan: Ujike (2011) reports that the numbers of ILI and H1N1 cases were collected by influenza sentinel and non-sentinel clinics in Japan. The number of influenza cases per sentinel clinic was digitised and scaled to estimate the average number per GP nationwide (Ujike et al., 2011). H1N1 cases were confirmed based on random samples collected from the sentinel clinics and subjective samples in non-sentinel clinics according to the different stages of surveillance rolled out at different points in the pandemic (Ujike et al., 2011).

Republic of China, Taiwan: On 18 June 2009, Taiwan temporarily changed their original influenza surveillance system to an influenza pandemic clinical surveillance system which focuses on reporting the possible H1N1 cases (Chao et al., 2011). This ad hoc surveillance system during the 2009 pandemic provides the number of hospitalised H1N1 and confirmed H1N1 (Chao et al., 2011). Since Taiwan's surveillance is already extant, it is a natural node in our hypothetical network.

Singapore: The available data includes the number of H1N1 confirmed deaths, the number of outpatient ILI and the number of confirmed H1N1 cases. The timing of confirmed deaths were collated by reviewing news articles from Channel NewsAsia, AsiaOne, and TR Emeritus, an independent Singapore online news site, since the start of the pandemic until the end of 2009, 18 confirmed deaths were related to H1N1 (MOH, 2009) and cross checked against Ministry of Health (MOH) press releases for completeness (MOH, 2009).

Since June 2009, Singapore started a sentinel GP network with 23 participating GPs to report the number of ILI consultations on a daily basis (Ong et al., 2010). We digitise the graph showing the average number of ILI consultations from each GP and rescale by 2 138 times, the total number of GPs in Singapore in 2009 (Lee

et al., 2011). Cutter et al. (2010) presented the number of imported and local cases of H1N1 reported to MOH in a bar chart which we used as the number of confirmed H1N1 cases.

Southern Hemisphere: Countries in the southern hemisphere only experienced one pandemic wave in 2009 during their winter season and the countries explored includes Brazil, Peru, Bolivia, Australia, Chile, Argentina and New Zealand (Opatowski et al., 2011). Opatowshi et al (2011) have extracted the number of confirmed H1N1 cases and ILI cases from the respective countries' surveillance system websites or public reports. We digitise the dataset from their graphical analysis for our use.

They also used a Bayesian approach with MCMC to find the posterior sample for the parameters of interest, but their parameters were distinguished by age (Opatowski et al., 2011), while we assume the rate of infection and removal to be the same for all age group. An additional move to build a hierarchical model could easily exploit information from similar outbreaks from other areas. Our investigation is more in depth as we have more countries in our analysis, involving countries from most continents across the northern and southern hemispheres, as well as tropical and temperate countries. We also have more types of surveillance data to provide information to our model. We will also demonstrate that all these countries with different seasons can be modelled together using a hierarchical model for more accurate predictions, by exploiting information from similar outbreaks of the same disease in other countries.

4.3 Model

This project will explore the use of Bayesian hierarchical modelling on the 2009 H1N1 pandemic. MCMC techniques are used to sample the parameters and the hyper-parameters over the large parameter space (Gilks et al., 1996). Advantages of MCMC are that it allows model flexibility, for any distribution that suits the data can be used, while it allows analysis of all parameters at the same time (O'Neill, 2002).

The unknowns that cannot be observed exactly at time t in the cth country are the number of individuals infected with H1N1, $I_c(t)$, and the cumulative number of

recovered or dead individuals since the start of the observation, $R_c(t)$.

In this analysis, the longitudinal observed statistics at time t in the cth country includes subsets of: the cumulative number of reported confirmed deaths due to H1N1, $D_c(t)$, the number of patients diagnosed with H1N1, $Z_c(t)$, the number of ILI patients (Hospitalised and/or outpatient), $X_c(t)$, the number of outpatient ILI patients (Government clinic and registered GP), $W_c(t)$, and the number of hospitalised H1N1 patients, $V_c(t)$.

The observed statistics are either related to the number of infected individuals, $I_c(t)$, or the cumulative number of recovered or removed individuals, $R_c(t)$, for the cth country at time t. We show how evidence synthesis can be used to pool information from these sources within the model to infer unknown quantities, such as $I_c(t)$ and $R_c(t)$.

Parameters can inform about many severity estimates, like the Case Fatality Ratio (CFR), Hospital Fatality Ratio (HFR), Case Hospitalization Ratio (CHR) and Final Attack Rate (FAR). Because different types of data were collected, severity estimates can be used to assess these different metrics of burden on the healthcare system. Presanis et al. (2009) have demonstrated the estimation of CFR, CHR and the Case Intensive care Ratio (CIR) by Bayesian Evidence Synthesis Framework. Our approach can be seen as a generalisation of theirs, though excluding the severity estimate for CIR which would require the data for number of H1N1 patients who were admitted to the intensive care unit (ICU). The Bayesian Evidence Synthesis Framework applied by Presanis et al. (2009) only considers each of these severity ratios as probabilities of occurrences given symptomatic cases and using these probabilities on a binomial model. It was also for the United States only, whereas our approach used dynamic time series data, compartmental modelling and data from many different countries.

The software used for this analysis is R (R Core Team, 2013).

4.3.1 Cumulative H1N1 Confirmed Deaths, $D_c(t)$

We model the cumulative number of confirmed deaths due to H1N1 as a negative binomial distribution:

$$D_c(t) \quad \sim \quad \text{NB}(n_c(t), p_c(t)) \; \forall c, t. \tag{4.1}$$

A negative binomial distribution is preferred due to its support over the non-negative

whole numbers, which is desirable for count data. Its flexibility in the shape of its distribution, which is controlled by two parameters, $n_c(t)$ and $p_c(t)$, is also more appropriate for providing different variance at different time points to account for the magnitude difference within each of the collected data points. A Poisson distribution is also not appropriate because of its inflexibility in the shape of the distribution which is controlled by one parameter, λ. Similarly, a normal distribution is rejected due to its support over the real numbers.

In the context of a negative binomial distribution, $D_c(t)$ is the number of trials until the occurrence of $n_c(t)$ number of successes based on the success probability of $p_c(t)$. The parameters, $n_c(t)$ and $p_c(t)$, can be calculated from the mean, $\mu_c(t)$, and variance, $\sigma_c^2(t)$.

The mean of the negative binomial distribution can be represented as

$$\mu_c(t) = \frac{n_c(t)\,(1-p_c(t))}{p_c(t)}\ \forall c,t \tag{4.2}$$

and the variance as

$$\sigma_c^2(t) = \frac{n_c(t)\,(1-p_c(t))}{p_c^2(t)}\ \forall c,t. \tag{4.3}$$

Manipulating the above equations, with the condition that $0 < p_c(t) \leq 1$,

$$p_c(t) = \frac{\mu_c(t)}{\sigma_c^2(t)}\ \forall c,t \tag{4.4}$$

and $n > 0$

$$n_c(t) = \frac{\mu_c(t)p_c(t)}{1-p_c(t)}\ \forall c,t. \tag{4.5}$$

The mean is taken to be proportional to the modelled number of removals $R_c(t)$, which includes both recoveries and confirmed deaths,

$$\mu_c(t) = \theta_{D(c)}R_c(t)\ \forall c,t \tag{4.6}$$

and the variance is also related to $R_c(t)$ by another parameter, $\eta_{D(c)}$,

$$\sigma_c^2(t) = \eta_{D(c)}R_c(t)\ \forall c,t. \tag{4.7}$$

This parametrization ensures a manageable number of parameters while still capturing the relationship between the model for infection and the mortality data. The two additional parameters, $\theta_{D(c)}$ and $\eta_{D(c)}$, were initially allowed to differ between

countries to account for differences in each country's population and health system characteristics that might result in greater risk of adverse events: according to Barrau et al. (2012), H1N1 patients with diabetes, cardiac insufficiency and morbid obesity are more likely to become severe cases, requiring intensive care or even resulting in confirmed death. As the prevalence of such risk factors differs in different settings (Barrau et al., 2012) so too should in general the proportionality parameters differ.

In an emerging infectious disease outbreak, we are not sure what the proportionality parameters ought to be. Case fatality ratios for different infections were 2.5% (Spanish influenza (Taubenberger & Morens, 2006)), 14–33% (H5N1, estimated to date, (Li, Choi, Sly, & Pak, 2008)) and roughly 90% (Ebola Virus, (King & Markanday, 2003)). Because of this, we assign the proportionality parameters for confirmed death, $\theta_{D(c)}$ and $\eta_{D(c)}$, uninformative prior distributions. The uninformative prior for $\theta_{D(c)}$ is

$$\theta_{D(c)} \sim U(0,1)\ \forall c \tag{4.8}$$

because it is impossible to have more confirmed deaths than the actual number of removed cases $R_c(t)$. $\eta_{D(c)}$ should be positive as variance is always positive, so the prior distribution will be uniform over the non-negative range with an arbitrarily large upper limit

$$\eta_{D(c)} \sim U(0, 500\,000)\ \forall c. \tag{4.9}$$

Although $\theta_{D(c)}$ should take different values in different countries (see above), the severity of the virus is unlikely to vary *too* greatly, and so evidence from one country or setting should inform estimates for others. The parameters would ideally be modelled hierarchically, but as we had data on mortality from three locations in our hypothetical surveillance network (England, New York and Singapore), it proved impossible to obtain good estimates on these parameters. We therefore changed the mean to

$$\mu_c(t) = \theta_D R_c(t)\ \forall c, t \tag{4.10}$$

with the same prior for the proportionality parameter

$$\theta_D \sim U(0,1). \tag{4.11}$$

To account for the longer time period between infection and confirmed death than infection and recovery (Riley et al., 2003), an additional parameter δ is incorporated, which is assumed to be the same for all countries. The mean and variance are then related to the shifted number of removed individuals by

$$\mu_c(t) = \theta_D R_c(t+\delta) \ \forall c, t \tag{4.12}$$

$$\sigma_c^2(t) = \eta_{D(c)} R_i(t+\delta) \ \forall c, t. \tag{4.13}$$

Again, a non-informative prior for the latent period, δ, (a discrete uniform prior) is used:

$$\delta \sim U_d(1, 100).$$

4.3.2 Confirmed H1N1 Cases, $Z_c(t)$

A similar model can be built for the number of confirmed H1N1 cases, using a negative binomial distribution

$$Z_c(t) \sim \text{NB}(n_c(t), p_c(t)) \ \forall c, t. \tag{4.14}$$

The two parameters, $n_c(t)$ and $p_c(t)$, can be calculated from the mean, $\mu_c(t)$, and variance, $\sigma_c^2(t)$, as illustrated in subsection 4.3.1.

Analogously, another pair of proportionality parameters, $\theta_{Z(c)}$ and $\eta_{Z(c)}$, are used for each country c. They should not be the same from country to country due to differences in their testing regimes, partly due to different risk perceptions as explained earlier. Since the parameter depends on the coverage of the surveillance network more than biological factors, the knowledge from one country should not directly affect the inference about other countries. Thus, the reporting parameters for these data should be independent across locations. Although the testing paradigm can also change within a country during the course of the pandemic, we will assume that the proportionality parameter, $\theta_{Z(c)}$, will remain constant as time progresses. The mean and variance for $Z_c(t)$ are proportional to the number of infected individuals, $I_c(t)$ by

$$\mu_c(t) = \theta_{Z(c)} I_c(t) \ \forall c, t \tag{4.15}$$

$$\sigma_c^2(t) = \eta_{Z(c)} I_c(t) \ \forall c, t. \tag{4.16}$$

Again, uninformative prior distributions are used on the parameters, $\theta_{Z(c)}$ and $\eta_{Z(c)}$. The observed data should be less than the actual number of people infected with H1N1, and so the proportionality parameter $\theta_{Z(c)}$ should also take values only from 0 to 1

$$\theta_{Z(c)} \sim U(0,1)\ \forall c. \tag{4.17}$$

The prior distribution of $\eta_{Z(c)}$ should also be positive and have a large upper bound,

$$\eta_{Z(c)} \sim U(0, 500\,000)\ \forall c. \tag{4.18}$$

4.3.3 Reported ILI Cases, $X_c(t)$

The parameters governing the number of reported ILI, $X_c(t)$, as a fraction of the actual number of H1N1 infections is expected to be independent for each country as the number of doctors or healthcare organizations under the surveillance system will deviate greatly due to social, not biological, factors. For example, there are only 1300 volunteering GPs in France, in contrast to the 5000 sentinel clinics in Japan reporting the number of ILI cases (Sentinelles, 2012; Ujike et al., 2011).

Similarly, differences in risk perception or medical usage may affect the proportion visiting the doctor in different countries. The highest weekly number of ILI cases reported as of 1 October 2009 are about 136 thousands and 24 thousands for France and Japan, respectively (Sentinelles, 2012; Ujike et al., 2011). The obligation of the physicians to report the ILI cases, as well as the stability of the surveillance network may result in differences in the reported ILI cases. Since France started their surveillance system in 1984, whereas Japan only started their surveillance system in 1997, the consistency of the system and the engagement of the physicians will be different (Sentinelles, 2012; NIID, 1998). The proportionality parameter, $\theta_{X(c)}$, that governs the portion of reported ILI data out of all those infected with H1N1 cases, $I_c(t)$, is therefore taken to be different for each dataset. Using the same negative binomial distribution,

$$X_c(t) \sim \mathrm{NB}(n_c(t), p_c(t))\ \forall c, t. \tag{4.19}$$

The two parameters, $n_c(t)$, and $p_c(t)$, can be calculated from the mean $\mu_c(t)$ and variance $\sigma_c^2(t)$, as illustrated in subsection 4.3.1. In accordance with the above, the

mean $\mu_c(t)$ and variance $\sigma_c^2(t)$ are also related to the number of people currently infected with H1N1, $I_c(t)$

$$\mu_c(t) = \theta_{X(c)} I_c(t) \ \forall c, t \tag{4.20}$$

$$\sigma_c^2(t) = \eta_{X(c)} I_c(t) \ \forall c, t. \tag{4.21}$$

As the number of ILI consultations will be under-reporting the actual H1N1 patients, $\theta_{X(c)}$ will be less than one and the prior distribution for $\theta_{X(c)}$ is chosen as

$$\theta_{X(c)} \sim U(0, 1) \ \forall c. \tag{4.22}$$

Uninformative prior distributions with a large upper limit is also chosen for measuring the spread as

$$\eta_{X(c)} \sim U(0, 500\,000) \ \forall c. \tag{4.23}$$

4.3.4 Outpatient ILI Cases, $W_c(t)$

Surveillance systems in some countries only involved volunteer GPs to submit the number of ILI consultations while some countries' network relied on hospitals, government clinics, as well as GPs. Data that come from the former will be classified as outpatient ILI cases, $W_c(t)$, and those from the latter will be categorised as the reported ILI cases, $X_c(t)$, described in the last section. Utilizing the negative binomial distribution again,

$$W_c(t) \sim \text{NB}(n_c(t), p_c(t)) \ \forall c, t \tag{4.24}$$

where $n_c(t)$ and $p_c(t)$ are calculated from the mean $\mu_c(t)$ and variance $\sigma_c^2(t)$, as illustrated in subsection 4.3.1, where $\mu_c(t)$ and $\sigma_c^2(t)$ are related to the actual number of people infected with H1N1, $I_c(t)$, by

$$\mu_c(t) = \theta_{W(c)} I_c(t) \ \forall c, t \tag{4.25}$$

$$\sigma_c^2(t) = \eta_{W(c)} I_c(t) \ \forall c, t. \tag{4.26}$$

Similarly, flat prior distributions are used for the proportionality parameters $\theta_{W(c)}$ and $\eta_{W(c)}$ as

$$\theta_{W(c)} \sim U(0, 1) \ \forall c \tag{4.27}$$

$$\eta_{W(c)} \sim U(0, 500\,000) \ \forall c. \tag{4.28}$$

4.3.5 Hospitalised H1N1 Cases, $V_c(t)$

Countries with better healthcare systems may have more hospitalised H1N1 cases. Within each country, the number of hospitalised H1N1 cases is modelled under the negative binomial distribution,

$$V_c(t) \sim \text{NB}(n_c(t), p_c(t)) \ \forall c, t \tag{4.29}$$

while $n_c(t)$ and $p_c(t)$ can be calculated similarly from the mean $\mu_c(t)$ and variance $\sigma_c^2(t)$. The mean $\mu_c(t)$ and variance $\sigma_c^2(t)$ will be proportional to the number of individuals infected with H1N1, $I_c(t)$, as

$$\mu_c(t) = \theta_{V(c)} I_c(t) \ \forall c, t \tag{4.30}$$

$$\sigma_c^2(t) = \eta_{V(c)} I_c(t) \ \forall c, t. \tag{4.31}$$

Because countries have a different proportion of patients being admitted into hospital due to H1N1, they should have different proportionality parameter, $\theta_{V(c)}$, but we assume this parameter to be hierarchical as the differences between countries are likely to reflect biological differences and differences in health seeking behaviour and not differences in the coverage of the surveillance system, so that similarities between them can be measured and controlled by the hyper-parameters. The prior distribution that is chosen for the $\theta_{V(c)}$ for the hospitalised H1N1 statistics, $V_c(t)$, will follow beta distribution,

$$\theta_{V(c)} \sim \text{Beta}(a_v, b_v) \ \forall c \tag{4.32}$$

to ensure that the parameter will only take values between 0 and 1. This prior distribution will be governed by the two hyper-parameters, a_v and b_v, which represents the shape parameters of the distribution. As these shape parameters should take positive values, the hyper-prior is chosen to be

$$a_v \sim \text{Exp}(1) \tag{4.33}$$

$$b_v \sim \text{Exp}(1). \tag{4.34}$$

Choosing the exponential parameters to be 1 will ensure that the (marginal) prior distribution of $\theta_{V(c)}$ is uniform over 0 and 1, as the mean of both hyperparameters is 1.

There should also be a constraint of $\theta_{V(c)} > \theta_{D(c)}$ for country c if the number of hospitalised H1N1 $V_c(t)$ and H1N1 confirmed deaths $D_c(t)$ are both available. Logically, there should be more hospitalised cases than confirmed deaths, which we have verified against our datasets. To impose this, we reject parameter values that do not meet this condition, i.e. the prior distributions are slightly modified to be proportional to those described herein times the indicator function, $\mathbb{I}[\theta_{V(c)} > \theta_{D(c)}]$.

4.3.6 SIR Model

To impose a correlation between estimates of disease prevalence over successive time points for country c, in particular in the number of H1N1 infections, $I_c(t)$, and the number of removals, $R_c(t)$, at time t, a mathematical, compartmental model can be used. The SIR model is the standard for large scale, respiratory outbreaks: in this model, $S(t)$ is the number of susceptible individuals, $I(t)$ the number of sick individuals and $R(t)$ the number of recovered or dead individuals at time t.

4.3.6.1 Stochastic SIR Model

This model characterises the two most important epidemic changes to the population: susceptible hosts becoming infected and infected hosts recovering or dying. The parameters governing these changes are the rate of infection per susceptible-infected pair, β, and the rate of removal per infected individual, α. Stochastically, the number of individuals in the infected, $I(t)$, and removed, $R(t)$, state could be simulated using the corresponding rates, β and α for each time t. These events simulation will then be used in the likelihood calculation.

It is almost impossible for any surveillance teams to collect complete data of $I_c(t)$ and $R_c(t)$ from the whole population, including the exact event time point and the exact number of events. Censored data actually allow a window for all the events to occur within the interval of data collection Hence, in this context, we will make use of the representative subset as mentioned in the above data type and to synthesize evidence for $I_c(t)$ and $R_c(t)$ by proportion.

If we are examining a small population, the problem of heavily censored observations can be overcome using data augmentation, a common inferential approach for stochastic epidemic models (Cook, Gibson, Gottwald, & Gilligan, 2008; Cooper, Medley, Bradley, & Scott, 2008; McKinley et al., 2009). This method of inference regards the unknown events as parameters to be estimated alongside the other un-

knowns in the model, typically with an MCMC routine in which unknown events are changed at each iteration to explore the space of events that are consistent with the data—for instance, having non-negative sizes in each category at all times.

If the exact event times and event types are not available, the likelihood cannot be calculated directly. Data augmentation is a method to simplify the likelihood (O'Neill, Balding, Becker, Eerola, & Mollison, 2000) as it allows the likelihood to be replaced by the probability of the unobserved, augmented variables, which can be calculated for most epidemic models. With the observed data D, event times can be randomly generated based on the current parameter values $\boldsymbol{\theta}$ and conditional on these augmented data A, likelihood is available

$$f(D|\boldsymbol{\theta}) \quad = \quad \int_A f(D|A,\boldsymbol{\theta}) \cdot f(A|\boldsymbol{\theta}) dA. \tag{4.35}$$

Clearly, when the augmented events A do not agree with the observed data, $f(D|A,\boldsymbol{\theta}) = 0$, giving $f(D|\boldsymbol{\theta}) = 0$. Else, if augmented events agree with data, it forms a possible path for the epidemic, so $f(D|A,\boldsymbol{\theta}) = 1$. With the above, the posterior distribution of all the parameters can be obtained by putting the idea of MCMC and data augmentation together. Along with each set of parameter simulation using, for instance, normal proposal distributions, a new set of augmented data are also generated. They are checked for the consistency with the data. With a suitable set of augmented times, the acceptance probability P_{acc} is calculated using the log likelihood function $\log f(A|\theta)$ and log prior density of all the parameters $\boldsymbol{\theta}$, $\log f(\boldsymbol{\theta})$. Then, the newly proposed parameter values $\boldsymbol{\theta}^*$ will be accepted with probability P_{acc}.

Unfortunately, this approach is not feasible for an analysis of the H1N1 pandemic, which affected the whole population, leading to too many event times and event types to explore. According to the Population Reference Bureau 2009 World Population Data Sheet (2009), the world population was 6.8 billion, and all of these individuals' statuses would need to be explored by data augmentation. Most of the collected data have a weekly frequency, which allows for even more variations in the number of events that can take place. Moreover, it will be even more computationally extensive if the number of individuals infected with H1N1 is computed separately for each country. In this methodology, the trajectory can only be accomplished by including every events in the each country based on their rate of infection and removal at each

step of the MCMC iteration. Instead, we replaced the stochastic model described above with a deterministic analogue, formed by solving a series of ODEs.

4.3.6.2 Deterministic SIR Model

In contrast to the stochastic model described in the last subsection, a deterministic model treats the number of individuals in each state as a variable that can take on any value and whose changing values over time are characterised by ordinary differential equations (ODE). The structure of the ODEs for the SIR model for the H1N1 pandemic is described below:

Infection can only occur when a susceptible individual is in contact with an infected individual. There are $S_c(t) \times I_c(t)$ possible contacts that can result in infection at time t. Thus, at any time t, the rate of decrease of $S_c(t)$ for country i can be represented by the product of the rate of getting infected per SI pair in country c, β_c, and the number of SI pairs in country c, $S_c(t)I_c(t)$,

$$\frac{dS_c(t)}{dt} = -\beta_c S_c(t) I_c(t). \tag{4.36}$$

We assume that infection across countries is negligible compared to infection within countries and each country is a homogeneous population where the people in the same country will react similarly to the disease.

Despite evidence that risks do differ in different sub-segments of the population (Chen et al., 2010; Lim et al., 2011), the assumption of homogeneity simplifies analysis tremendously, while the additional variability caused by heterogeneity can be partially accounted for via the observation model.

Correspondingly, at any time t, the rate of increase of $R_c(t)$ for country c can be represented by the product of the rate of removal in country c, α_c, and the number of infected individuals in country c, $I_c(t)$,

$$\frac{dR_c(t)}{dt} = \alpha_c I_c(t). \tag{4.37}$$

Since a susceptible individual S will become an infected individual I when infected, the rate of decrease of $S_c(t)$ will translate into the rate of increase of $I_c(t)$. Likewise, an infected individual I will become a removed individual R when recovered or died, the rate of increase of $R_c(t)$ will be interpreted as the rate of decrease

of $I_c(t)$. Together, the rate of change of infected individuals is

$$\frac{dI_c(t)}{dt} = \beta_c S_c(t) I_c(t) - \alpha_c I_c(t). \tag{4.38}$$

With these ODEs, the trajectories of $S_c(t)$, $I_c(t)$ and $R_c(t)$ for any value of β_c and α_c can be calculated numerically (we used the R package, `odesolve` (Setzer, 2012)) given initial conditions. As the time taken to compute the trajectory of the model is non-trivial, it is important to design the inferential algorithm to be as efficient as possible, since this solution of the ODEs will be required in every step of MCMC for the calculation of likelihood. We will discuss how we achieved computational efficiency for all the countries.

Because the rates of infection and removal are highly dependent on the number of S and I, standard models which posit a constant incidence risk over the whole period are inappropriate, as, for example, the risk of infection is much lower at the start of a pandemic than at its peak when there are many infected individuals in the population. One alternative, when data are informative enough, is to use a semi-parametric model in which the per-capita rates of infection do not depend on the state of the epidemic but are left to be free parameters that change over time, an approach that multiplies the number of parameters to be estimated substantially. This method has been successfully used by Cauchemez and Ferguson (2008) to study measles transmission in London where the hazard changes fortnightly. The data used by Cauchemez and Ferguson (2008) was collected fortnightly from 1948 to 1964. The hazard rate for every fortnight will be used repeatedly for 16 years, allowing for sufficient information to inform about the parameters. But, in our context, the amount of information (over a few months and one epidemic wave) is insufficient to do likewise.

4.3.6.3 Technical Challenges for solving ODE for Different countries

Because the data were collected in countries or territories of varying sizes, the numbers in each compartment will vary substantially. One approach would be to solve the system of ODEs separately for different countries. However, this would increase the computation time of the algorithm as a whole. Hence, we worked with the proportion of people in each disease state instead. In particular, we set $S_c(t) = n_c s(t)$,

$I_c(t) = n_c i(t)$, and $R_c(t) = n_c r(t)$ where n_c is the population size of the country in year 2009 and the lower case variables are proportions. Most of the countries' sizes were taken from the World Population Policies 2009 (UN, 2010). The population sizes of England, New York and Taiwan cannot be found from the above-mentioned report as they are constituent parts of larger states (the United Kingdom, the United States, and China, respectively) and so were taken from other sources; England's population size in 2009 was taken from their Office for National Statistics (2009), New York's was taken from the paper where we got the number of H1N1 confirmed death (Lee et al., 2010) and Taiwan's population size in 2009 was taken from the 2009 World Population Data Sheet by the Population Reference Bureau (PRB, 2009).

It is assumed that the whole population is susceptible to the H1N1 disease before the pandemic, so that $s(0) = 1 - i(0) - r(0)$. (Although note that according to a serological test of elderly in Finland, some had antibodies against this virus due to the infection from previous influenza outbreaks due to a related virus, such as the Spanish influenza (Ikonen et al., 2010). As the proportion was low, this complication was omitted.)

We assumed an arbitrary small proportion of individual to be infected with H1N1 at the start of the pandemic, $i(0) = 0.000001$. The presumed small number of infections prior to the declaration of the H1N1 pandemic by the WHO are ignored, i.e. $r(0)$ is set to 0. Using these initial conditions and different sets of values for α and β, the trajectories of $i(t)$ and $r(t)$ at the indicated times $t = 1, 2, \ldots, 587$—the time from 1 Jan 2009 to 10 August 2010 when the WHO declared the end of the pandemic—can calculated using the `lsoda` function in the `odesolve` package and stored for use in the MCMC stage for all countries (Setzer, 2012) in an array, with a grid of values for α and β and a set of times t.

Due to numerical approximations, it is possible for the solution to the ODEs to take negative values towards the tail of the epidemic. To prevent this, we set all negative entries of $i(t)$ to 0.

As we store the solution of the ODEs to an array, there is a limit to the number of sets of α and β that we can explore before storage becomes prohibitively expensive. To allow other values to be used, apart from those stored, we used bilinear interpolation on the two dimension space of α and β for the values of $i(t)$ and $r(t)$ for any

values of α and β within a certain range. Both α and β are rates which should only be positive. The upper limit for α is chosen to be smaller than or equal to 1 because it is not biologically plausible that the number of days of being infective, represented by the reciprocal of α, be less than 1. Suppose there is only 1 infected host and m possible contacts with susceptible, the rate of infection per day is βm. If β is more than 1, this infected host is able to infect more susceptible than he is able to meet. Hence, we choose the upper limit for β to be less than 1.

Suppose the fifty values that we divide equally in the above range can be named as $\alpha(1), \alpha(2), \ldots, \alpha(50)$ and $\beta(1), \beta(2), \ldots, \beta(50)$. The solution of the ODE for the $i(t)$ trajectory with the simulated parameter α and β in the Metropolis-Hastings Step that is lying between $(\alpha(j), \alpha(j+1))$ and $(\beta(k), \beta(k+1))$ will be represented as $i(t, \alpha, \beta)$ in this bilinear interpolation context.

The first interpolation will be between the $\alpha(j)$ and $\alpha(j+1)$ while keeping β fixed at $\beta(k)$,

$$\begin{aligned} i(t, \alpha, \beta(k)) &= \frac{\alpha(j+1) - \alpha}{\alpha(j+1) - \alpha(j)} \cdot i(t, \alpha(j), \beta(k)) + \\ &\quad \frac{\alpha - \alpha(j)}{\alpha(j+1) - \alpha(j)} \cdot i(t, \alpha(j+1), \beta(k)). \end{aligned} \tag{4.39}$$

The same interpolation is done for $\alpha(j)$ and $\alpha(j+1)$ while keeping β fixed at $\beta(k+1)$,

$$\begin{aligned} i(t, \alpha, \beta(k+1)) &= \frac{\alpha(j+1) - \alpha}{\alpha(j+1) - \alpha(j)} \cdot i(t, \alpha(j), \beta(k+1)) + \\ &\quad \frac{\alpha - \alpha(j)}{\alpha(j+1) - \alpha(j)} \cdot i(t, \alpha(j+1), \beta(k+1)). \end{aligned} \tag{4.40}$$

With these two sets, an interpolation can be done between $\beta(k)$ and $\beta(k+1)$ to get the final interpolation done by

$$\begin{aligned} i(t, \alpha, \beta) &= \frac{\beta(k+1) - \beta}{\beta(k+1) - \beta(k)} \cdot i(t, \alpha, \beta(k)) + \\ &\quad \frac{\beta - \beta(k)}{\beta(k+1) - \beta(k)} \cdot i(t, \alpha, \beta(k+1)). \end{aligned} \tag{4.41}$$

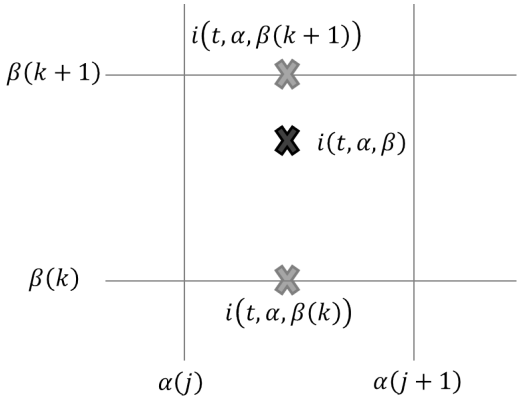

Figure 4.3: **Illustration of bilinear interpolation.**
First, the values at the two grey crosses are calculated by interpolation between $a(j)$ and $\alpha(j+1)$ while fixing at $\beta(k)$ and $\beta(k+1)$ respectively. The black cross can be computed by interpolating the values at the two grey crosses.

The above described bilinear interpolation is done for the trajectories for the number of infected $i(t, \alpha, \beta)$ or $r(t, \alpha, \beta)$ over the required time t. If the simulated α and β fall outside of the inspected range, we will use the `odesolve` to get the solution (Setzer, 2012).

The basic reproduction number, the expected number of secondary infections resulting from a single infected individual in a population otherwise susceptible (Heffernan, Smith, & Wahl, 2005), $R_0 = f(\alpha, \beta)$, is a threshold parameter which is able to inform us whether the pandemic has the potential to take off ($R_0 > 1$) or will die out quickly ($R_0 < 1$).

In this project, the historical data for the 2009 H1N1 pandemic will be analysed, and prediction of the trajectory of $S_c(t)$, $I_c(t)$ and $R_c(t)$ can be done and they could be scaled to useful data type as described in the previous section. Nonetheless, we focus on how hierarchical models can be implemented worldwide to improve estimates for any emerging pandemic.

4.4 Hierarchical Model

In hierarchical modelling, information from multiple sources (here countries) is pooled. In our context, there will be much variability between forecasts of the outbreaks in each country if each country is modelled independently. However, if parameters across all countries take the same parameter value, differences between countries' experiences cannot be accounted for, and estimates will be unjustifiably narrow. Hierarchical modelling can help to address this problem by introducing

hyper-parameters to measure the variability of parameters for the different countries. A hierarchical model also allows us to use information from countries which have more data for countries which have limited local surveillance of the pandemic.

It is clear that β_c should be different for each country c, since different countries have different policies to deal with the rate of disease spreading. As forecast by Kubiak and McLean (2012), if the control measure of school closure in England had not been implemented, the number of infections by the end of the first wave in England would have been much higher. To allow for such differences requires the infection rate be different between countries with different control measures, but if β_c is nevertheless sufficiently similar for different countries, a hierarchical model would be appropriate, because the infection rate per SI pair, β_c, for country c should still be exchangeable across different countries. Logarithmic transformation on the parameters allowed a normal distribution to be used which takes continuous, real values,

$$\log(\beta_c) \quad \sim \quad N(\mu_\beta, \sigma_\beta^2), \tag{4.42}$$

where the mean, and standard deviation can be characterised by the hyper-parameters, $(\mu_\beta, \sigma_\beta)$.

Although β_c must be positive, they can be any real numbers after taking logarithms. Therefore, the hyper-prior distribution of μ_β is flat and allowed to take any values over a large range,

$$\mu_\beta \quad \sim \quad U(-1\,000, 1\,000). \tag{4.43}$$

Because the standard deviation cannot be negative, a similar arbitrary range over small to large positive values is chosen as the hyper-prior distribution for σ_β,

$$\sigma_\beta \quad \sim \quad U(0, 1\,000). \tag{4.44}$$

We assume that all α_c are equal ($\alpha_c = \alpha$) because the rate of removal of the disease should be the same in all countries, representing as it does a purely biological phenomenon.

4.4.1 Informative prior for removal rate

In contrast to the rate of between-host transmission for a novel variant of influenza, the within host dynamics of (seasonal) influenza are well understood. This would

allow an informative prior for the removal rate, α, using data from previous studies. Carrat et al. (2008) did a detailed analysis on the duration of the course of virus infection for H1N1 from multiple published research studies. It was found that for (pre-pandemic) H1N1, the average duration for illness was 4.50 days, with the 95% confidence interval from 4.31 to 5.29 days.

The above conclusion was about how long the illness will last, but we are interested in how long is the infected individual infectious. For a better estimate of the length of infection period for our prior distribution for α, we looked into the sources that they cited for possible information that we can make use of. In the daily serological tests on the volunteers, the H1N1 antibodies titers were recorded. If the recorded titers are above certain threshold, they will be considered as infected individuals. Once it falls below again, that marks the end of the course of infection. We digitised six log mean viral titers plots for use (Barroso, Treanor, Gubareva, & Hayden, 2005; Fritz et al., 1999; Hayden et al., 1994, 1996, 1998; Treanor, Betts, Erb, Roth, & Dolin, 1987). These papers were studying the physical response of the volunteers to the use of placebo and drugs, and so we digitised only data from the placebo arms of these studies.

Other than the log mean of the viral titers at each day, the standard error (SE) and the number of volunteers (n) can also be found, either from the plot or stated in the paper. For each plot that we digitised, we calculate the standard deviation (σ) of the viral titers by $\sigma = \frac{\text{SE}}{\sqrt{n}}$. Assuming that the log viral titers for day k of the dataset l, T_{kl}, follow normal distributions,

$$T_{kl} \sim N(\mu_{kl}, \sigma_{kl}^2) \ \forall k, l, \tag{4.45}$$

the probability, p_{kl}, that the log viral titers of day k of dataset l will exceed a given threshold ε can be calculated for each day is

$$p_{kl} = \Pr(T_{kl} > \varepsilon) \ \forall k, l. \tag{4.46}$$

The expected duration of infection for the dataset l is

$$D_l = \sum_k p_{kl}. \tag{4.47}$$

The estimated duration of infection is estimated from the average and standard deviation of D_l from all the six datasets. We trialed several different thresholds for

infectivity, $\varepsilon = \{0, 1, 2\}$, selecting $\varepsilon = 2$ to be the most appropriate threshold based on the face validity of the resulting infectious period. Note that this rate is actually the recovery rate, but is used as a proxy for the removal rate, α, as the confirmed death rate is close enough to 0 to be ignored.

At $\varepsilon = 2$, the mean and standard deviation of the expected duration of infection for all the dataset are 2.53 and 0.714 respectively. These values become the information for the informative prior for the infectious period, $\frac{1}{\alpha}$,

$$\frac{1}{\alpha} \sim N(2.53, 0.714^2). \tag{4.48}$$

4.4.2 Modification to Overall Model on initial analysis

On fitting the model described above using MCMC, the routine would not converge despite many attempts. The countermeasures discussed here are very common solutions for MCMC non-convergence. MCMC was tried independently on each country to allow higher rate of acceptance of proposed parameters. This is because when a large number of parameters were proposed for the data for so many countries, the chances that they will all suit the available data, and yield a high likelihood, is low, which will then result in rejection of the proposal. We also tried to get better initial values for the Markov chain to reach convergence quickly, and experimented with proposal distributions of different covariances as well as mixtures of distributions.

4.4.2.1 Multiple waves

According to Borja-Aburto et al (2012), there have been four waves of H1N1 since 2009 and the virus has displaced the pre-pandemic H1N1 as one of three main seasonal influenza strains (including influenza A/H3N2 and influenza B) (Belshe, 2010). A territory which shows clearly that there is a change in pandemic trajectory is Hong Kong. As stated by news.gov.hk, an online news platform launched by the Government of the Hong Kong Special Administrative Region, Hong Kong changed from the containment phase to the mitigation phase on 12 June 2009.

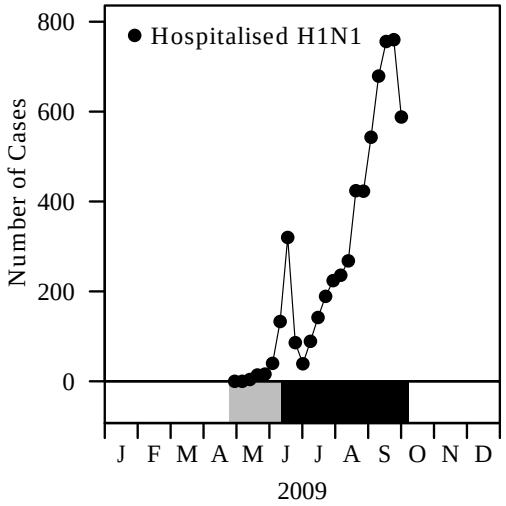

Figure 4.4: **Number of hospitalised H1N1 in Hong Kong Public Hospitals collated by Riley et al. (2011)**.
The number of hospitalised H1N1 in Hong Kong is double peaked in June 2009 and September 2009. The grey and black rectangles at the lower panel show the containment period and the mitigation period adopted by The Government of the Hong Kong Special Administrative region.

Figure 4.4 supports the claim by Wu et al (2010) that Hong Kong's policy changed on 30 June 2009 to the criteria for the admission to hospital for H1N1 patients to be based on medical needs, rather than for isolation. Other than Hong Kong, Finland and England also displayed two peaks for the number of confirmed H1N1 cases $Z_c(t)$ and number of outpatient ILI cases $W_c(t)$ respectively. Japan also had two peaks for both the number of reported ILI cases $X_c(t)$ and the number of confirmed H1N1 cases $Z_c(t)$. Thus, we decided to introduce a new parameter, τ_c, to identify the date of policy change which would cause a change in the shape of the trajectory for each country c. Note that the date which the policy change takes effect is not necessary the date of phase change.

The phase change in Hong Kong should involve the relaxing of H1N1 hospitalization policies. To account for the reduction in the numbers of admissions, the proportionality parameter, $\theta_{V(c)}$, should change accordingly.

This parameter should conceivably account for the time of either a second wave or a change in the country's surveillance policy which can affect any of the data type. Using the hospitalised H1N1 cases, $V_c(t)$, as an example, the reporting parameter, $\theta_{V(c)}$ was split into two, one to report before the change and the other to report after the change.

Recall that the model is

$$V_c(t) \sim N(\mu_c(t), \sigma_c^2(t)), \tag{4.49}$$

where the mean is now characterised by the two proportionality parameters $\theta_{1V(c)}$ and $\theta_{2V(c)}$

$$\mu_c(t) = \theta_{1V(c)} I_c(t) \mathbb{I}(t < \tau_c) + \theta_{1V(c)} \theta_{2V(c)} I_c(t) \mathbb{I}(t \geq \tau_c). \tag{4.50}$$

Before the effective date of changes, τ_c, the proportion of individuals infected with H1N1 should remain the same, and so the prior distribution for $\theta_{1V(c)}$ is still

$$\theta_{1V(c)} \sim U(0,1). \tag{4.51}$$

For a further reduction in the observations after τ_c, the prior distribution of $\theta_{2V(c)}$ should be

$$\theta_{2V(c)} \sim U(0,1). \tag{4.52}$$

To avoid over-parametrization, parameters that measure the spread of the data were kept the same as before

$$\sigma_c^2(t) = \eta_{V(c)} I_c(t), \tag{4.53}$$

where the prior distribution remains as

$$\eta_{V(c)} \sim U(0,1). \tag{4.54}$$

However, this analysis did not work as there was insufficient evidence from the datasets to inform (i) the degree of change and (ii) when change occurred. As a result, we focused only on the first, main wave of H1N1, and assumed constant reporting rates within countries across time.

4.4.2.2 Start dates

Because H1N1 virus has been circulating in other countries prior to the WHO announcement (Chao et al., 2011), a parameter is introduced to describe the start date for country c, $t_0(c)$. This parameter will shift the ODE solution down the time line and replace the gap from the first day, 1 Jan 2009, to day $t_0(c)$ by the initial values

of the trajectory, $I_c(0)$ and $R_c(0)$. It is beneficial to observe that there is a different connotation for τ_c and $t_0(c)$. The former changes the proportion of the number of cases but the latter only changes the time where the outbreak starts in the country c. Because there is no prior knowledge on the starting date for each country, a flat discrete uniform prior distribution is used on $t_0(c)$ from 1 Jan 2009 (day 1) to the day that WHO declared the end of pandemic on 10 August 2010 (day 587) (WHO, 2010).

4.4.2.3 Seasonality

We initially attempted to account for seasonality differences between countries, accounting for countries that are not near the equator having different seasonal patterns at different times of the year. The rate of infection, β_c, previously taken to be a parameter, was reformulated as a function of time, t. A sine function of time t was used to give the smooth oscillating effect to mimic the seasonal effect within a year. A scaling parameter for country c, κ_c, is used to adjust for the difference in the magnitude of the rate of infection in the different seasons. The exponential function ensures that $\beta_c(t)$ remains positive by transforming negative sine values to values between 0 and 1, which will reduce the value of β_c during the warmer season; it will also transform positive sine values to values more than 1, which will increase the value of β_c during the cold season. As there are 365 days in a year, there is approximately one complete cycle is time t is in terms of degree in the trigonometrical function. Either a cosine curve or a translated sine curve will coincide with the climatic patterns for a year, where the latter is chosen. The function of β_c with seasonal effect becomes

$$\beta_c(t) \quad = \quad \beta_c \exp\left(\kappa_c \sin\left((t + 90.5) \times \frac{\pi}{180}\right)\right). \tag{4.55}$$

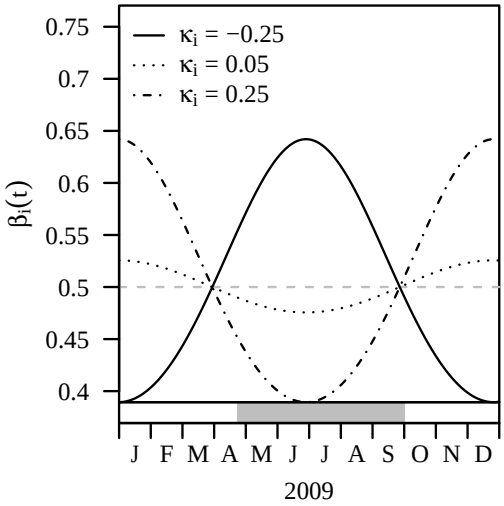

Figure 4.5: **Seasonality characterisation using a transformed sine function of time t.**
Different values of κ_c can affect $\beta_c(t)$. The grey dotted line is when β_c is fixed at 0.5 ($\kappa_c = 0$) for no variation in $\beta_c(t)$ against time t. When $\kappa_c < 0$, the black solid line shows the shape of the $\beta_c(t)$ function for the southern hemisphere and when $\kappa_c > 0$, the black dotted and dotdash lines shows the shape of $\beta_c(t)$ function for the northern hemisphere. The larger the value of κ, the more pronounced the variability in the curve will be. At the lower panel, the grey rectangular box shows the range of the days where the pandemic data is used if we do not want the seasonality to affect the infection rate.

Recall that the trajectory of the pandemic is derived using the ODE solution from an R package (Setzer, 2012). If β_c becomes a function which varies with time, we need to use a method such as Euler's method (Atkinson & Kendall, 2008) to solve the ODE. Similarly, Euler method can be a step-wise, deterministic way of finding the trajectory of a pandemic. Presumably, the rate of occurrence, $f'(x)$, is the gradient of the graph $y = f(x)$, represented by

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{4.56}$$

Suppose for a small time step, h, the rate of decrease of $S_c(t)$ is stated earlier to be $\beta_c S_c(t) I_c(t)$. If the rate of infection can vary with time and displaying using the above relation,

$$-\beta_c(t) S_c(t) I_c(t) \approx \frac{S_c(t + h) - S_c(t)}{h}. \tag{4.57}$$

The approximate number of susceptible in the next time step can be derived from the above relation,

$$S_c(t + h) \approx S_c(t) - \beta_c(t) S_c(t) I_c(t) \cdot h. \tag{4.58}$$

Similarly, the number of individuals infected with H1N1 can be calculated by

$$I_c(t+h) \approx I_c(t) + (\beta_c(t)S_c(t)I_c(t) - \alpha_c I_c(t)) \cdot h. \tag{4.59}$$

With the above relation for $S_c(t)$ and $I_c(t)$, the whole trajectory could be calculated from the initial values $S_c(0)$ and $I_c(0)$. Intuitively, larger h will lead to a coarser trajectory that is a greater approximation to the true trajectory implied by the model. This seasonal effect, coupled with the Euler's method to replace the ODE solution, was explored but was eventually abandoned because the data did not display sufficient seasonal effects for the estimation of the newly introduced parameters to work. We also realised that the seasonal effect with double humps in the $I_c(t)$ trajectory will only happen when the values of α and β are similar to each other: in other scenarios a single epidemic wave resulted and so the increase in complexity did not seem warranted.

4.4.2.4 Correlation between α and β_c

Previously, the posterior samples in the MCMC routine showed that α and β_c are correlated for certain countries. We will change to propose the basic reproduction number for country c, $R_0(c)$, and the rate of removal, α, and calculated the infection rate, β_c, based on those.

Because we assume that all individuals are susceptible at the start of the pandemic, i.e. $s_c(0) = 1$, at the start of the H1N1 outbreak in country c, if the rate of infection is $\beta_c s_c(0) i_c(0)$ and rate of removal is $\alpha i_c(0)$, the basic reproduction number can be expressed as a ratio of infection rate to removal rate,

$$\begin{aligned} R_0(c) &= \frac{\beta_c s_c(0) i_c(0)}{\alpha i_c(0)} \\ &= \frac{\beta_c}{\alpha}, \end{aligned} \tag{4.60}$$

where β_c can be calculated by $\beta_c = \alpha \cdot R_0(c)$. The original hierarchical model for (α, β_c) will be changed to model $R_0(c)$ of different countries,

$$R_0(c) \sim N(\mu, \sigma^2). \tag{4.61}$$

The hyper-parameters are reduced to μ and σ. As we have no prior information we wish to use, and since $R_0(c)$ should only be positive (and take values a little more

than 1), we set a uniform prior with an arbitrary large upper limit,

$$\mu \sim U(0, 100). \tag{4.62}$$

Similarly, σ should also take positive values,

$$\sigma \sim U(0, 100). \tag{4.63}$$

4.5 Model Fitting

As mentioned before, we fit a Bayesian hierarchical model to multiple data types relating to the H1N1 pandemic for a hypothetical network of countries. The parameter space for this model is large and the likelihood expensive to calculate. MCMC is a method to sample from the posterior distribution of parameter given the available observed data (Cauchemez et al., 2006) and that is well suited for our problem as it is a sampling methodology that efficiently draws from the actual posterior distribution using the Metropolis-Hastings algorithm to decide whether a proposed parameter should be included into the posterior sample.

4.5.1 Metropolis-Hastings Algorithm

For this particular project on H1N1, due to the complexity of the model, the implementation of the algorithm is non-trivial and hence summarised in the following:

1. We need to prepare an array of the proportion of infected individuals, $i(t)$, and proportion of removed individuals, $r(t)$, over time t after introduction of the virus, for values of infection rate, β, and removal rate, α, over a grid from 0 to 1. This array of information will be stored for use in all countries for the likelihood calculation. To make use of the `odesolve` package, the ODE have to be defined as a function. The initial values of the trajectories, as well as the vector of time steps for exploring, should also be specified with the values of β_c and α.

2. The initial value for each parameter is chosen to be in the vicinity of the posterior by trial-and-error, using graphical comparison of the model against the data. This makes the routine less computationally unwieldy, especially important due to the many parameters in this model.

3. The likelihood is calculated based on the current initial parameter values $\boldsymbol{\theta^0}$.

   a) Using the bilinear interpolation method, we can find the trajectory of the proportion of infected individuals $i_c(t)$ and the proportion of removed individuals $r_c(t)$ for each country c using the proposed (or initial) rates of infection and removal, β_c^0 and α^0. Recall that the number of H1N1 confirmed deaths, $D_c(t)$, will relate to the removed individuals and the other data types will relate to the infected individuals. Also, the rate of infection β_c^0 will be computed deterministically from the parameter value of $R_0^0(c)$ and α^0.

   b) We need to incorporate the latent period of confirmed death, δ^0, from the time when the individuals get infected by shifting the time of the trajectory for each of the country c relative to the H1N1 confirmed death data. The data time for the number of H1N1 confirmed deaths should be reduced by δ^0 days temporarily for this iteration and stored as a temporary confirmed death data,

$$D'_c(t) \quad = \quad D_c(t - \delta^0). \tag{4.64}$$

   c) Next we need to combine the delay in the trajectory for country c based on $t_0^0(c)$. The trajectory will only start for country c from day $t_0^0(c)$. We will insert a baseline of proportion of infected individuals before $t_0^0(c)$ to be $i_c(0) = 0.000001$ and proportion of removed individuals before $t_0^0(c)$ to be $r_c(0) = 0$.

   d) With the relevant proportions $\left(\theta^0_{D(c)}, \theta^0_{Z(c)}, \theta^0_{X(c)}, \theta^0_{W(c)}, \theta^0_{V(c)}\right)$ for the data type for each country c, the proportion trajectories can be converted into the actual numbers by multiplying with the population size of country c, n_c. For instance, the mean trajectory for the number of H1N1 confirmed death cases for country c will be related by proportion to the modified number of removed individuals by

$$\mu_c(t) \quad = \quad \theta^0_{D(c)} \times n_c \times r_c(t). \tag{4.65}$$

   Similarly, the trajectory for the number of confirmed H1N1 cases (or other data types) in country c will also be related by proportion to the modified

number of infected individuals by

$$\mu_c(t) = \theta^0_{Z(c)} \times n_c \times i_c(t). \tag{4.66}$$

e) The parameter that measures the spread of the data from the trajectory is computed. Recall the variance formula for the number of H1N1 confirmed death cases is

$$\sigma^2_c(t) = \eta^0_{D(c)} \times n_c \times r_c(t). \tag{4.67}$$

Similarly, for other data types, we relate the variance with the modified number of infected individuals.

f) With mean, $\mu_c(t)$, and variance, $\sigma^2_c(t)$, the parameters necessary for the negative binomial distribution can be calculated as

$$p_c(t) = \frac{\mu_c(t)}{\sigma^2_c(t)} \tag{4.68}$$

$$n_c(t) = \frac{\mu_c(t)p_c(t)}{1-p_c(t)}. \tag{4.69}$$

g) The likelihood for the H1N1 confirmed death data type for country c can be computed using a negative binomial distribution with the above parameters $p_c(t)$ and $n_c(t)$.

h) The likelihood for all the different data types of the different countries will be computed in a similar manner and the product of all these likelihood values will be the overall likelihood for the data given the all the parameters $\boldsymbol{\theta^0}$. Here, we assume that data from each data type is independent from the other data type, conditioned on the parameters that have been proposed. Similarly, the data from each country will be also assumed as independent conditioned on the parameters, as the amount of trans-border mixing of infectious hosts from one country to another will pale into insignificance relative to the number of infections acquired locally within the country. It will be easier to work with log-likelihood, which will then be the sum of all the log-likelihood contributions for each country and each data type.

4. Other than the log-likelihood, $\log f(D|\boldsymbol{\theta^0})$, we also need to calculate the log prior density $\log f(\boldsymbol{\theta^0})$ for all these initial parameters $\boldsymbol{\theta^0}$. The pseudo log posterior density $\log f(\boldsymbol{\theta^0}|D)$ can also be calculate by adding log-likelihood and log prior density, as the normalizing constant will cancel with the subsequent log posterior density.

5. In pilot rounds of the MCMC algorithm where our proposal distribution is not well tuned, the parameters will be proposed individually. The first parameter value, α^*, will be proposed using a normal proposal distribution centered at the initial parameter value, α^0 and an arbitrarily specified standard deviation, 0.0001,

$$\alpha^* \sim N(\alpha^0, 0.0001^2). \tag{4.70}$$

The choice of normal distribution is due to the preference of a symmetrical proposal distribution, $q(\alpha^* \to \alpha^0)$ for the simplification of acceptance probability calculation,

$$P_{\text{acc}} = \min\left(1, \frac{f(\alpha^*|D, \boldsymbol{\theta^0})}{f(\alpha^0|D, \boldsymbol{\theta^0})} \cdot \frac{q(\alpha^* \to \alpha^0)}{q(\alpha^0 \to \alpha^*)}\right) \tag{4.71}$$

$$= \min\left(1, \frac{f(\alpha^*|D, \boldsymbol{\theta^0})}{f(\alpha^0|D, \boldsymbol{\theta^0})}\right), \tag{4.72}$$

where D is the collected data and $\boldsymbol{\theta^0}$ is the initial values of the other parameters which were not yet proposed.

6. We will check the proposed parameter value of α^* based on the model conditions, i.e. $\alpha^* > 0$. If the conditions cannot be fulfilled, they can be rejected straight away without wasting time in their likelihood calculation.

7. If they satisfy the criteria, we will use likelihood procedure mentioned earlier in step 3 to find the value of the log-likelihood for the new proposal $\log f(D|\boldsymbol{\theta^*})$ where $\boldsymbol{\theta^*}$ represent the set of initial values with the newly proposed α^*.

8. The log prior density for the new set of parameter $\log f(\boldsymbol{\theta^*})$ can also be calculated.

9. As we subsequently will draw a uniform(0,1) variable and compare it to the acceptance probability, outlined below, we can neglect the requirement that

the acceptance probability be ≤ 1 for notational brevity. The log acceptance probability can be calculated as follows

$$P_{\text{acc}} = \frac{f(\boldsymbol{\theta}^*|D)}{f(\boldsymbol{\theta}^{\mathbf{0}}|D)} \tag{4.73}$$

$$\log P_{\text{acc}} = \log f(\boldsymbol{\theta}^*|D) - \log f(\boldsymbol{\theta}^{\mathbf{0}}|D) \tag{4.74}$$

$$= \log f(D|\boldsymbol{\theta}^*) + \log f(\boldsymbol{\theta}^*) - \log f(D|\boldsymbol{\theta}^{\mathbf{0}}) - \log f(\boldsymbol{\theta}^{\mathbf{0}}) \tag{4.75}$$

where $\log P_{\text{acc}}$ is essentially the difference between the pseudo log posterior density under $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{\mathbf{0}}$.

10. A random number can be generated from $r \sim U(0,1)$. The proposed parameter values $\boldsymbol{\theta}^*$ will be accepted with probability P_{acc}. If $\log r > \log P_{\text{acc}}$, we will reject the proposed α^* and $\alpha^1 = \alpha^0$, otherwise, we will accept the proposed α^* and update $\alpha^1 = \alpha^0$.

11. After updating the rate of removal, α, we can propose for the next parameter by repeating the step 5 to 10. The parameters change accordingly.

12. After all the parameters have been proposed and updated, we propose new values for all the hyper-parameters. Because the acceptance probability for hyper-parameters does not include the likelihood calculation, updating of the hyper-parameters is faster. Suppose $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ represent all the parameters and hyper-parameters respectively and D represents the data, the acceptance probability is

$$P_{acc} = \min\left(1, \frac{f(\boldsymbol{\theta}, \boldsymbol{\eta}^*|D)}{f(\boldsymbol{\theta}, \boldsymbol{\eta}|D)} \cdot \frac{q(\boldsymbol{\eta}^* \to \boldsymbol{\eta})}{q(\boldsymbol{\eta} \to \boldsymbol{\eta}^*)}\right). \tag{4.76}$$

The ratio of proposal density for the hyper-parameters equal to 1 because a normal proposal distribution centered at the current value is used. The ratio of the posterior densities can be simplified by

$$\frac{f(\boldsymbol{\theta}, \boldsymbol{\eta}^*|D)}{f(\boldsymbol{\theta}, \boldsymbol{\eta}|D)} = \frac{f(D|\boldsymbol{\theta}, \boldsymbol{\eta}^*) \cdot f(\boldsymbol{\theta}|\boldsymbol{\eta}^*) \cdot f(\boldsymbol{\eta}^*)}{f(D|\boldsymbol{\theta}, \boldsymbol{\eta}) \cdot f(\boldsymbol{\theta}|\boldsymbol{\eta}) \cdot f(\boldsymbol{\eta})} \tag{4.77}$$

$$= \frac{f(\boldsymbol{\theta}|\boldsymbol{\eta}^*) \cdot f(\boldsymbol{\eta}^*)}{f(\boldsymbol{\theta}|\boldsymbol{\eta}) \cdot f(\boldsymbol{\eta})} \tag{4.78}$$

because the likelihood, $f(D|\boldsymbol{\theta}, \boldsymbol{\eta})$, is not affected by the change of hyper-parameters. As such, the hyper-parameters will be updated 20 times for every complete round of parameter proposals.

13. The whole routine will have to be repeated a large number of times to give us the posterior sample for all the updated parameters.

4.5.2 Solution for Non Converging MCMC

On initial runs, the traceplots of the posterior samples did not show convergence. We first attempted to preserve the model and change the proposal distributions similar to the tuning of proposals in simple algorithms: If the proposal of the new parameters is always far from the posterior, because the variance of the proposal is too large, the proposed parameters will invariably be rejected. Likewise, if the proposed values are too close, because the variance is small, it will take too much steps to move around the posterior.

Moreover, the posterior distributions of many of the parameters are closely correlated. If one of the proposed parameters is inappropriate, the other parameters will also be affected. Although each parameter can be proposed separately and a different acceptance probability can be calculated for each parameter, it will become too computationally intensive as there were many countries involved and the trajectories have to be interpolated at every proposal for the calculation of the likelihood. So, we propose changes to all the parameters individually for a small amount of iterations, 2500, with 500 burn-in and thinning on every 10 iterations as described in the previous section. The burn-in is chosen to be small because suitable initial parameter values were used.

After completing this first trial round of proposal, we tuned the standard deviation of all the parameters and hyper-parameters proposal distributions with the respective standard deviation of the posterior samples. Following this update of the proposal standard deviation, the actual MCMC routine was conducted afresh with 100,000 iterations and no burn-in, as the routine will start with the parameters values from the previous step, itself assumed a draw from the posterior.

In this step of MCMC with better proposal distributions, the parameter, α, is proposed independently from the other parameters and hyper-parameters while the other parameters and hyper-parameters are proposed together in batches, within countries where appropriate. α and β (from R_0) are capable of affecting the proportional trajectories of all the countries. So, this proposal approach will only change

the trajectories twice at every round. Proposing all other parameters and hyper-parameters together will most likely result in rejection, so, we propose the remaining parameters by country, followed by the hyper-parameters.

4.5.2.1 Sequential Importance Sampling

Even the modified algorithm described in the previous section did not lead to convergence. We proceeded with an alternative solution, Sequential Importance Sampling. In Importance Sampling, parameters are simulated from a proposal distribution and weighted using the likelihood, prior and proposal density. If we can sample from the posterior distribution directly, this gives an unweighted sample, but otherwise the weights correct for sampling from an incorrect distribution. Although the previous results of the non-converging MCMC may not be the exact posterior distribution, it should still be close to the desired posterior distribution, and so by approximating the MCMC sample by a suitable multivariate distribution, we can generate samples from a distribution close to the target distribution.

Since it is vital to sample from a distribution that is close to the desired distribution, we will progressively improve the distribution that we sample from, using the weighted samples from the previous rounds. In the initial rounds, when the proposal distribution is still not so similar to the target distribution, we reduce the weight contributed by the likelihood density by an intensity constant, T. The value of T will gradually increase from 0.1 to 1 in the sequential steps to successively allow the weighted samples to inform about the target distribution.

The Sequential Importance Sampling algorithm is:

1. Using the mean and variance of the posterior sample for each parameter and hyper-parameter and assuming independence between all parameters (a conservative assumption), set up a multivariate normal proposal distribution with covariance matrix only having entries on the diagonals and zero elsewhere.

2. Sample 100 000 particles from this multivariate normal proposal distribution and calculate the weights for each particle. Weights are represented by the ratio of the posterior to the proposal density, where the posterior is proportional to the likelihood, prior and hyper-prior density. In importance sampling, the

weight for the lth particle is

$$w_l = \frac{f(D|\boldsymbol{\theta_l})f(\boldsymbol{\theta_l}|\boldsymbol{\eta_l})f(\boldsymbol{\eta_l})}{q(\boldsymbol{\theta_l},\boldsymbol{\eta_l})}. \tag{4.79}$$

Because the particles may not characterise the data well in the first round, the intensity of the likelihood could be scaled down by T and the weight for the lth particle becomes

$$w_l = \frac{(f(D|\boldsymbol{\theta_l}))^T f(\boldsymbol{\theta_l}|\boldsymbol{\eta_l})f(\boldsymbol{\eta_l})}{q(\boldsymbol{\theta_l},\boldsymbol{\eta_l})}. \tag{4.80}$$

We let $T = 0.1$ in this first round and gradually increase this amount to 1 when the proposal distribution has been improved to become close to the posterior distribution.

3. Taking logarithms, we can better see how the intensity T will be relating the likelihood to the weights

$$\begin{aligned}\log w_l &= T\log f(D|\boldsymbol{\theta_l}) + \log f(\boldsymbol{\theta_l}|\boldsymbol{\eta_l}) \\ &\quad + \log f(\boldsymbol{\eta_l}) - \log q(\boldsymbol{\theta_l},\boldsymbol{\eta_l}).\end{aligned} \tag{4.81}$$

If we take exponential transformation to convert $\log w_l$ back to w_l for lth particle, many of the values will become 0 if the $\log w_l$ is small. Hence, we overcome numerical overflow issues by transforming the $\log w_l$ to

$$(\log w_l)^* = \log w_l - \max(\log \boldsymbol{w}) \tag{4.82}$$

before exponentiating to get

$$w_l^* = \exp(\log w_l)^* \tag{4.83}$$

and rescaling all w_l^* by

$$w_l = \frac{w_l^*}{\sum w_l^*} \tag{4.84}$$

so that all the weights, w_l, sum to 1.

4. The particles, together with their respective weights, will be the information for the next round's multivariate normal proposal distribution, with weighted mean of the particles and weighted covariance matrix of the particles. Repeat step 2 to 3 and sequentially, increase the intensity, T, of the likelihood give a better representation of the weights for the description of the next proposal distribution.

The intensity, T, started small to prevent over-concentration of the few particles with good weights. As Sequential Importance Sampling progressed from the first round to the tenth round, the particles will gradually become a good realization of the posterior samples. In the final round, $T = 1$ and the routine above yields a weighted sample from the correct posterior distribution; only this round is used for analysis.

Some challenges arose in the importance sampling step. The covariance matrix that we calculated from the proposed samples was non-invertible, resulting in an invalid proposal distribution in the next round. This problem was reduced by reparameterising parameters by taking logarithms of rates and logit transformations of probabilities.

After the transformation, the covariance matrix might occasionally still be non-invertible. On investigation, we realised that for certain parameters, the weights concentrate the proposed particles for certain parameters to a single value. This is because the proposal distribution for that parameter became too focused due to the weights in the previous rounds. This would lead to zero variance for the proposal distribution in the next round, resulting in a singular covariance matrix. The solution for this problem was to increase the sampling size sequentially and reduce the stepwise increment of T, to ease the over-concentration at certain points.

This sequential importance sampling routine is repeated with the data available at four different time points (1 June 2009, 1 July 2009, 1 August 2009 and 1 September 2009) to show the improvement in prediction and severity estimation with increasing data in a real pandemic outbreak.

4.6 Results and Inference

This model for the outbreak of H1N1 at different territories has made used of some fixed parameters, as well as parameters which were modelled independently and hierarchically. To decide which parameters should be the same, one approach is to use expert opinion. If they should be the same, the parameter can be fixed for all the places, for example, the rate of removal, α, in the H1N1 project, due to the similar biological capability of the infected individual to recover from the disease. On the other hand, if the parameter should be different, for example, the proportionality constant, $\theta_{X(c)}$, which accounts for the fraction of reported ILI patients in country

c and that can be much affected by the capacity of the healthcare system in each territories. The other proportionality constant, $\theta_{Z(c)}$, which represents the portion of infected H1N1 who were confirmed by laboratory test for country c, would also be affected by the testing paradigm in the different countries,and would be better modelled independently.

In this case, we decide which parameter should be hierarchically modelled or fixed depending on the amount of information we had. For example, there were only three territories with mortality data, and so the information quantum to estimate mortality rate was insufficient to allow useful hierarchical estimates. The solution for the lack of data can be solved by using a fixed proportionality constant, θ_D, for all territories. In contrast, as there were many areas with hospitalized H1N1 data, the proportionality constant, $\theta_{V(c)}$, can be modelled hierarchically.

With a richer dataset, the ideal statistical approach is to model every parameter hierarchically. This would allow us to estimate how similar or different they were. The different ways of modelling can be compared against using the Deviance Information Criterion (DIC). Using the posterior samples achievable from the MCMC simulations (if convergence exists) or the sequential importance sampling (the alternative method if there is non-convergence in MCMC), the average log likelihood of a sample of parameters, $D(\bar{\theta})$, and the average log likelihood from every set of parameters, $\bar{D}$, can be calculated. The DIC is $p_D + \bar{D}$ or $D(\bar{\theta}) + 2p_D$ where $p_D = \bar{D} - D(\bar{\theta})$. The model with the smallest DIC should be preferred out of all the trials, and this would allow a statistically informed decision on whether a hierarchical model was needed.

The main motive of this model is to show that prediction of the severity of a new outbreak at both a per country and global level can be done in real time using a network of countries each providing their own outbreak information in real time. This section assesses the feasibility of this goal by applying the method to the 2009 influenza pandemic as a case study.

Observing the peak for at least some countries is necessary to inform the parameters appropriately as there were no informative priors for the basic reproduction number, R_0. By the end of July 2009, about three months from the start of the pandemic, when some but not all of the countries had experienced their peak, the data could give a good estimate to the parameters, which would lead to better prediction.

We will present the results of each country sorted by latitude.

4.6.1 Finland

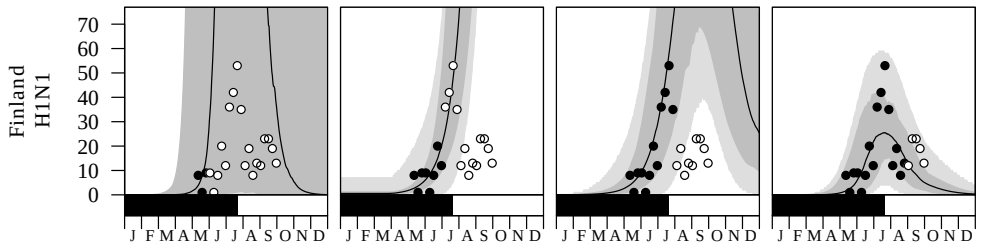

Figure 4.6: **Forecasts of pandemic H1N1 confirmed cases in Finland.** Solid circles indicate data available at the point the forecast is made, hollow circles indicate future data, black lines indicate best forecast, shaded regions indicate uncertainty (dark) and observation error (light). The black bar at the bottom of each panel shows the change in the countries policy to change from containment to mitigation phase. The four time points used are the beginning of June, July, August and September, 2009.

In the first panel, there were only three available data points for analysis. The forecast was vague because of the limited information. Despite this the shape of the best forecast of the pandemic was close to the actual shape as the hierarchical model is able to draw information from the other countries. However, the magnitude of the number of H1N1 cases were over predicted, as $\theta_{Z(c)}$ was independent for each country and the information on this Finland-specific parameter was insufficient at the time the first estimate was made to yield a good estimate of $\theta_{Z(c)}$.

In the next panel, data up to 1 July 2009 were used. As there were more points showing an upward trend, the pandemic was projected to spread by the increasing H1N1 cases. However, with the extra information collected over June, the uncertainty was greatly reduced compared to the first panel.

In the third panel, the available data were still showing upward trend, but other countries or territories—such as New York, Bolivia, Argentina and Chile—provided a fair amount of data by 1 August 2009 to show the epidemic had peaked and was waning, affecting the common removal rate, α, and the mean and variance of the basic reproduction number, $R_0(c)$. As a consequence, the model forecast that the number of H1N1 cases would fall, while fitting closely to the observed data available at that time, even when there was as yet no sign of the epidemic having peaked. The forecast does not characterise the data after the peak well, but the post-peak data

do not appear to be consistent with the data before the peak, or with the SIR model we are fitting. This may result from changes to the way H1N1 cases were tested as the country switched from the containment to the mitigation phase on 22 July 2009 (Saarinen, Järvinen, Haikala, & Ruutu, 2009).

By the beginning of September 2009, when most of the data were available for use, the model is able to provide a satisfactory prediction, by informing the H1N1 peak for Finland at the correct time point.

4.6.2 England

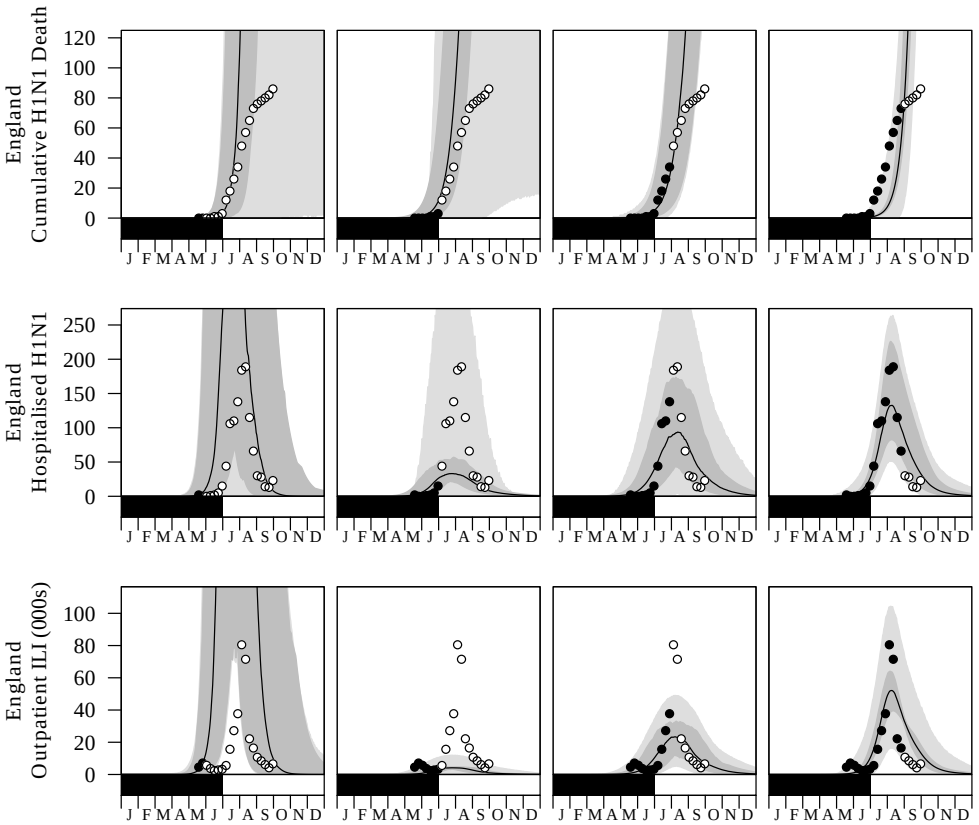

Figure 4.7: **Forecasts of pandemic H1N1 confirmed deaths (cumulative), H1N1 hospitalizations and ILI cases in England.**
Features in this figure are as in figure 4.6.

The worst prediction for England occurs when early data up to 1 July is used. Because of the change from containment to mitigation stage, there is a minor peak in the outpatient ILI data up to 1 July, which the model misinterpreted and used to predict a small magnitude pandemic that would end early (see the second column). The

prediction performance of England for the other time points are similar to that of Finland.

4.6.3 France

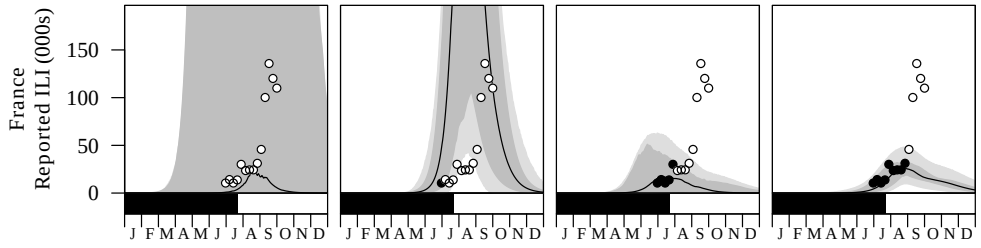

Figure 4.8: **Forecasts of ILI cases during the pandemic H1N1 in France.** Features in this figure are as in figure 4.6.

The prediction range for the first panel was as ambiguous as that in Finland because no data were available by the beginning of June 2009, and so the sole information was that pooled from the countries with available data. In the last panel, as the data available to the end of August are not consistent with the sudden rise that accompanied September, the forecast, though it fits the data until the end of August well, does not predict the data well thereafter. It is not likely that this is due to reporting biases as the surveillance data from France is considered robust (Sentinelles, 2012), and the rise may be due changes not present in the model, such as the end of the long August holidays in France and the return of children to school (Merler, Ajelli, Pugliese, & Ferguson, 2011).

4.6.4 New York

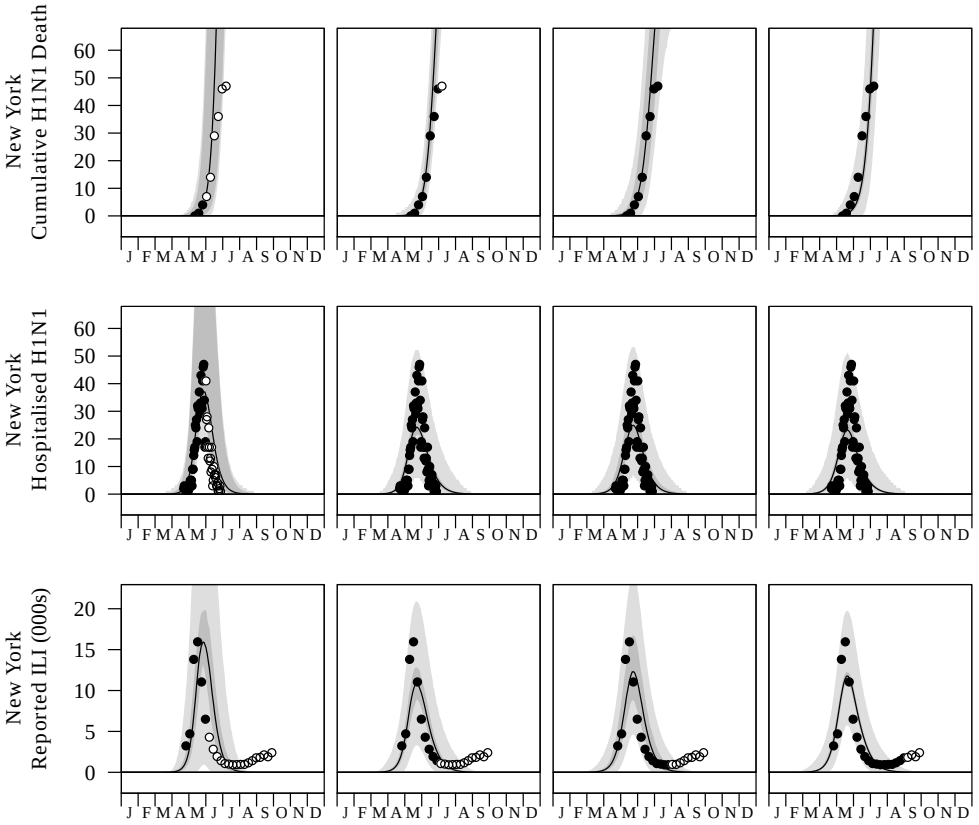

Figure 4.9: **Forecasts of pandemic H1N1 confirmed deaths (cumulative), H1N1 hospitalizations and ILI cases in New York.**
Features in this figure are as in figure 4.6. There is no grey bar at the bottom of each panel because New York started off with the mitigation phase (Nicoll & Coulombier, 2009).

New York provided a very informative dataset. By the beginning of July 2009, the available data were already sufficient to show that the pandemic was coming to an end. The trajectories for the three data types for New York were reliable and precise, with little noise. But we can still see a change in direction of the number of reported ILI cases after 1 August 2009 which should not be due to the change in pandemic phase. A possible reason for this mild increase in the number of reported ILI cases could be due to the other influenza viruses that were also circulating in New York.

4.6.5 Japan

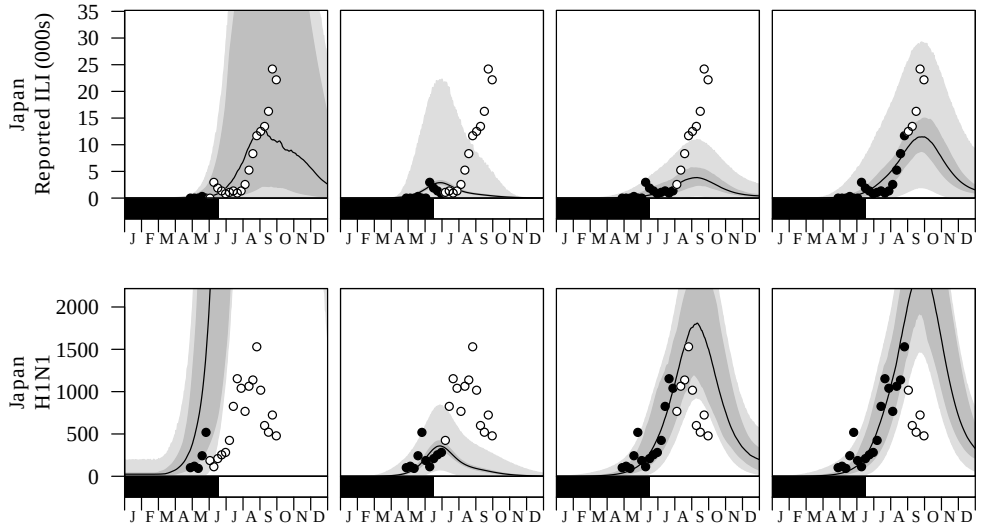

Figure 4.10: **Forecasts of ILI cases and pandemic H1N1 confirmed cases in Japan.**
Features in this figure are as in figure 4.6.

Japan also shows a small peak before mitigation and another major peak after the phase change in both available datasets. This resulted in a prediction that is not able to capture the major peak when we are at the second time point. However, by the third time point, the benefit of synthesizing evidence from multiple data type is demonstrated. Because the number of confirmed H1N1 cases which has yet to show an end to the pandemic, the trajectory of the pandemic was projected to peak at September. By 1 September 2009, the available data shows the epidemic was still growing but the hierarchical model provided information that the numbers should be decreasing after that. An interesting fact for Japan is while the number of H1N1 cases decreases starting from September, the number of reported ILI actually continued to increase. This may also be due to the circulation of other influenza viruses, as ILI is a syndrome caused by both influenza and non-influenza respiratory pathogens (Babcock, Merz, & Fraser, 2006).

4.6.6 Republic of China, Taiwan

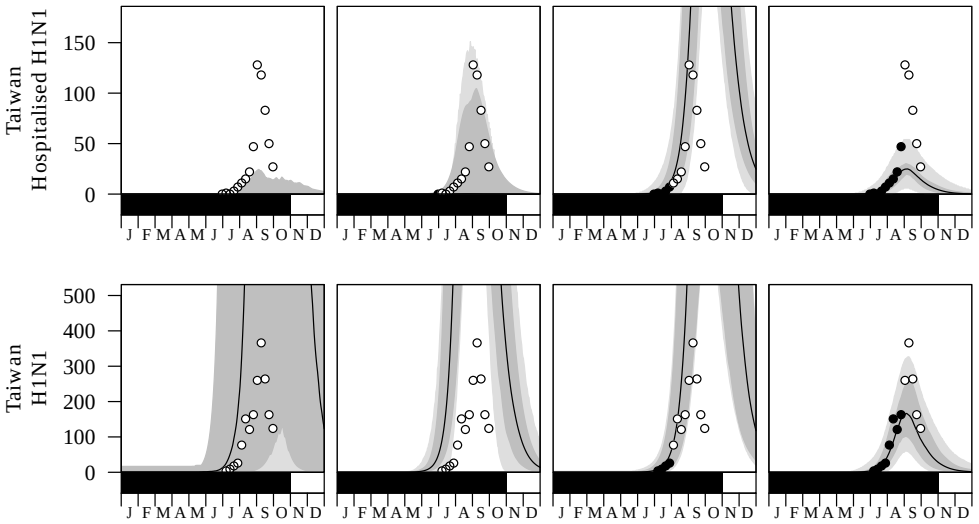

Figure 4.11: **Forecasts of pandemic H1N1 hospitalised and confirmed cases in Taiwan.**
Features in this figure are as in figure 4.6.

The spread of H1N1 in Taiwan started quite late and there was only one data point by 1 July 2009. This would be an example of how a country could benefit from the hierarchical model without even a single data point being collected. Despite the bad prediction of the *magnitude* of confirmed H1N1 cases, due to $\theta_{Z(c)}$ being independent between territories, the shape of the projection of the number of confirmed H1N1 cases was appropriate. Another advantage of modelling $\theta_{V(c)}$ hierarchically is being able to provide us with constructive projections for the number of hospitalised H1N1 cases in the first and second time point by pooling information from the other countries. By the third time point, the peak was not predicted at the right time using the limited data that is available by 1 August. By the fourth time point, there was sufficient information which resulted in a befitting shape for both type of data. The sudden jump in the number of hospitalised H1N1 cases in September could not be accommodated by the epidemic model used. It is not clear whether this is due to a change in the hospitalisation rate or due to a change in the virulence of the pathogen.

4.6.7 Singapore

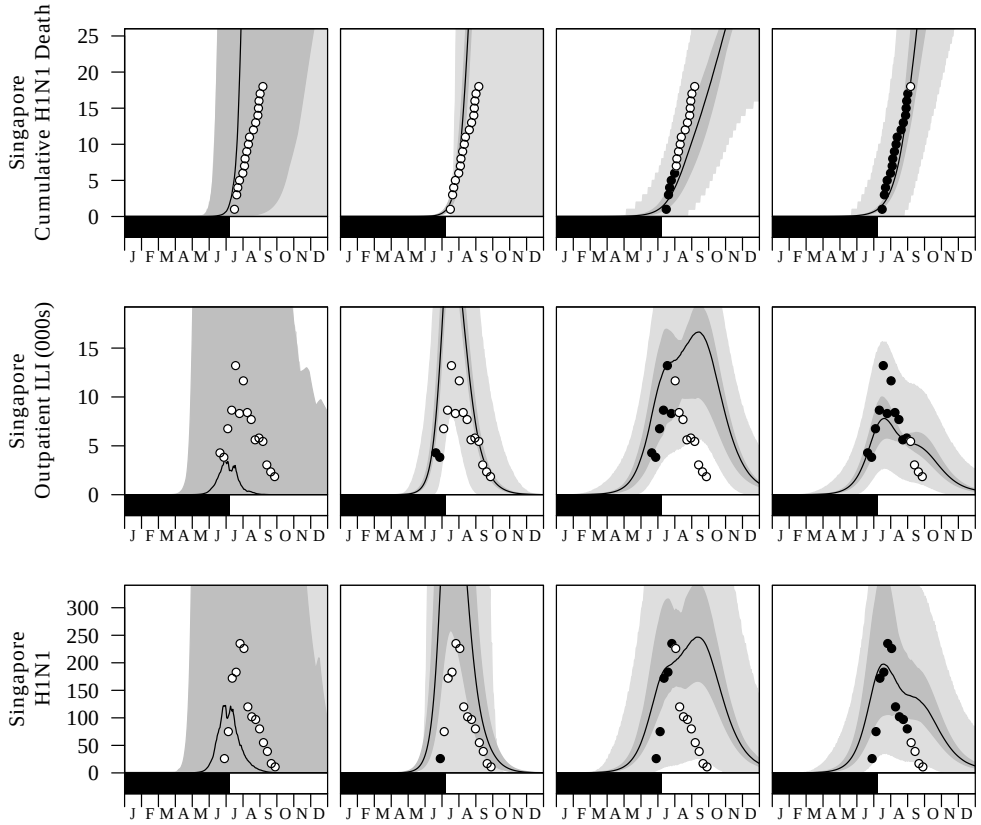

Figure 4.12: **Forecasts of pandemic H1N1 confirmed deaths (cumulative), ILI cases and H1N1 confirmed cases in Singapore.**
Features in this figure are as in figure 4.6.

Because the proportion of H1N1 confirmed deaths, θ_D, was treated as being constant across territories, the cumulative number of H1N1 confirmed deaths was modelled fairly well, even before there were available data on the mortality rate in the first two time points. In the first column, the shape for the number of outpatient ILI and H1N1 cases was appropriate due to the common removal rate, α, and the hierarchically modelled R_0, but the magnitudes of these two projections were not consistent with the subsequently observed data due to the independence of proportionality constants, $\theta_{W(c)}$ and $\theta_{Z(c)}$ between countries and the paucity of data to estimate those for Singapore at that time. Similarly, by 1 August 2009, the number of outpatient ILI and H1N1 cases were still increasing, the model forecast an imminent decline by borrowing information from other countries. However it was not able to predict the sharpness of the decline observable in the data. By 1 September 2009, the predictions for the number of outpatient ILI and H1N1 cases did not capture the

observed patterns in the data, which may result from the two datasets not being in synchrony. One possible explanation for this is that the outpatient ILI data was collected by a small network of 23 Singapore GPs and so may be unreliable, while the testing regime may have changed over time, leading to inconsistencies in the number of confirmed cases.

4.6.8 Brazil, Peru and Bolivia

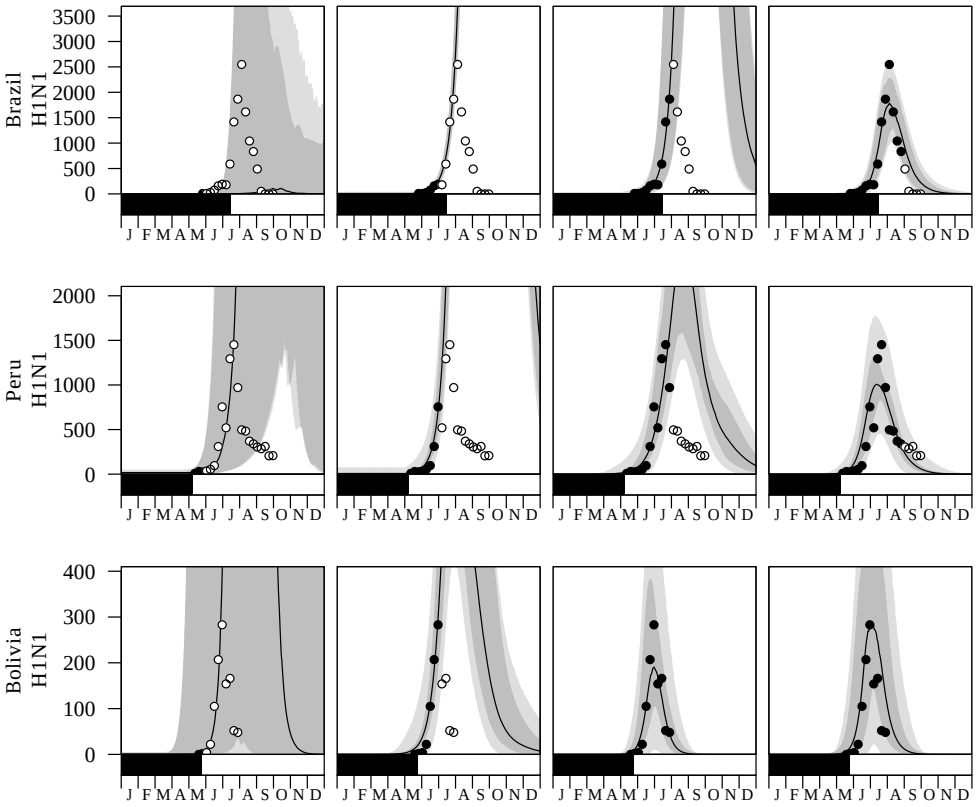

Figure 4.13: **Forecasts of pandemic H1N1 confirmed cases in Brazil, Peru and Bolivia.**
Features in this figure are as in figure 4.6.

These three countries in the southern hemisphere depicted similar patterns. When there is no information yet available for analysis, the prediction is wide reflecting the uncertainties in how the pandemic will evolve. If there is only an upward trajectory at the second time point, the model will predict that the pandemic will be severe in these countries. By 1 August 2009, only Bolivia's data showed an end to the outbreak and the model was able to model the shape of the trajectory well. With the almost complete data for analysis in the fourth time point, the projection became

very reliable.

4.6.9 Australia

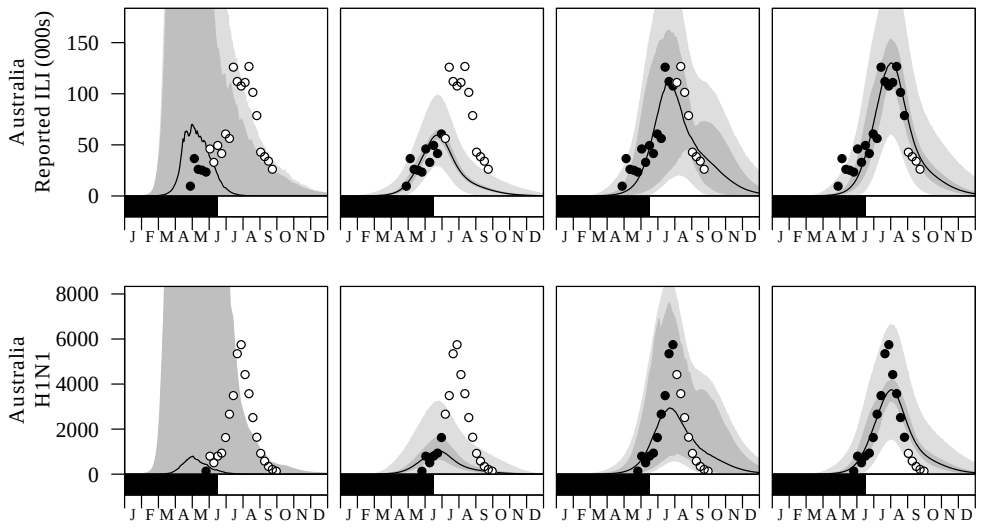

Figure 4.14: **Forecasts of ILI cases and pandemic H1N1 confirmed cases in Australia.**
Features in this figure are as in figure 4.6.

The first four data points for reported ILI for Australia provide a misleading impression that the pandemic had already peaked in early May, affecting the forecasts as a result. It is not clear whether this problem might have been caused by anomalous data collection methods or just stochasticity. Similarly, the shape of the ILI data for Australia do not follow the standard epidemic curve, which resembles a Gaussian, and it is not clear whether this is due to changes in data collection protocols, the merging of data from different outbreaks across this large country, or some other reason. Regardless of the reason, the sudden rise in the number of reported ILI in July and August caused a poor fit for the second time point. If a consistent protocol were available that stratified data spatially, this problem might have been averted.

4.6.10 Chile

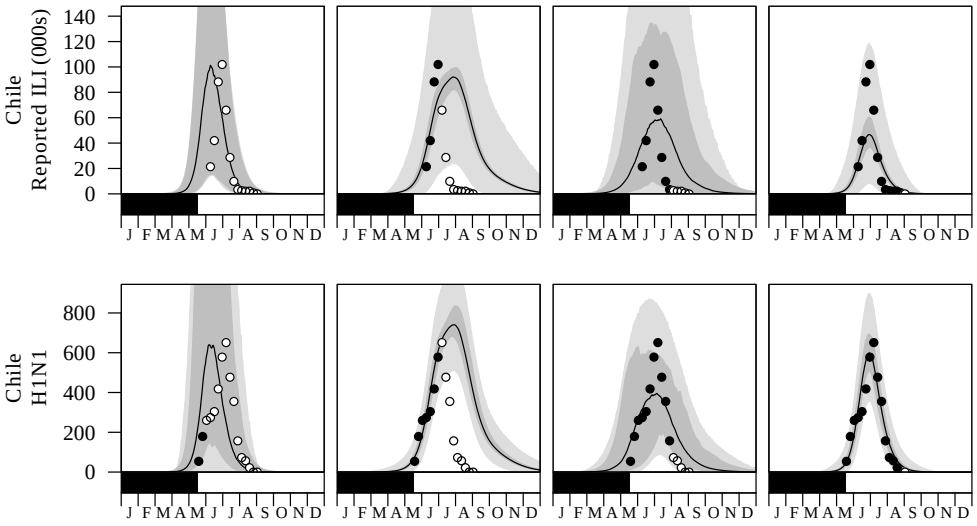

Figure 4.15: **Forecasts of ILI cases and pandemic H1N1 confirmed cases in Chile.**
Features in this figure are as in figure 4.6.

By 1 June 2009, the prediction was considerably good with evidence from two data points in number of H1N1 cases. The outbreak can be concluded by 1 August 2009 but the uncertainty in prediction was larger for the reported ILI than that of the number of H1N1 cases. The predicted trajectory 'tried' to accommodate the number of H1N1 cases more than the ILI data because there are more data points than the number of reported ILI.

4.6.11 Argentina

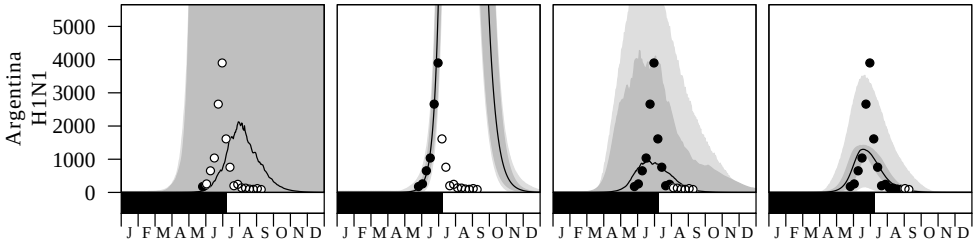

Figure 4.16: **Forecasts of ILI cases and pandemic H1N1 confirmed cases in Argentina.**
Features in this figure are as in figure 4.6.

Forecasts for Argentina suffer similar problems to most other Latin American countries, in that the H1N1 cases are predicted to grow rapidly at the second time point, and the model fits well only at the end of the epidemic.

4.6.12 New Zealand

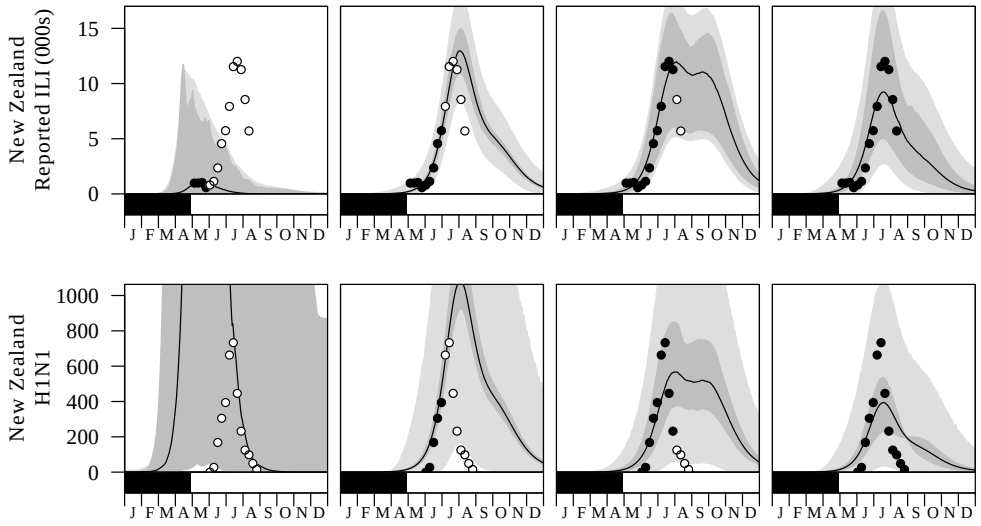

Figure 4.17: **Forecasts of ILI cases and pandemic H1N1 confirmed cases in New Zealand.**
Features in this figure are as in figure 4.6.

Similar to Australia, the number of reported ILI cases of New Zealand showed a minor dip at the end of May, which might have misguided the model to think that the pandemic started early and will be ending soon in the first time point analysis. In the later time points' analysis, it was shown that with sufficient data, the model fits the rest of the data well.

Other than predicting the number of cases, hospitalisations and confirmed deaths, have good, early severity estimates is essential when rolling out suitable changes to the intervention strategies against the pandemic. These severity estimates are described in the next subsection.

4.6.13 Case Hospitalization Ratio (CHR)

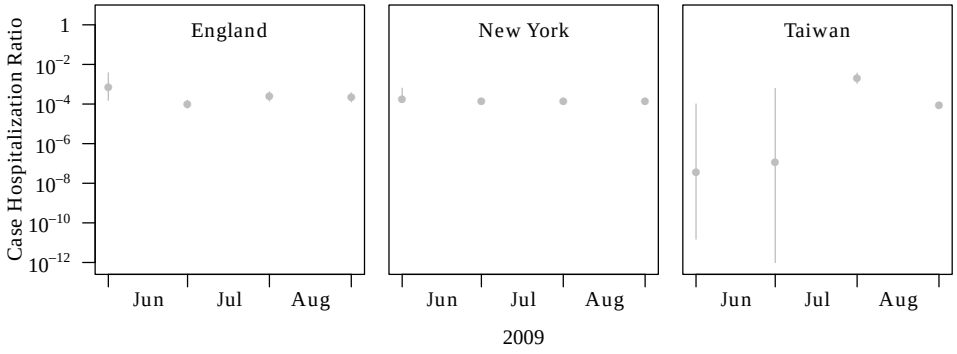

Figure 4.18: **Severity estimate of Case Hospitalization Ratio (CHR).** These are the real-time estimates if such a network had been established in 2009 for countries where hospitalised H1N1 cases are available. Dots represent posterior medians and lines 95% equal-tailed credible intervals. CHR is the number of hospitalizations due to H1N1 over the estimated total H1N1 cases which is represented by $\theta_{V(c)}$ in the model for the ith country.

The Case-Hospitalisation Ratio (CHR) is the ratio of the number of hospitalised H1N1 cases, $V_c(t)$, to the total number of individuals infected with H1N1, $I_c(t)$. Since the mean of $V_c(t)$ is $\theta_{V(c)}I_c(t)$, CHR is represented by $\theta_{V(c)}$. Estimates are available for three locations: England, New York and Taiwan.

The assessment of the CHR for England and New York are similar. At the last time point, median estimate of CHR for England was 0.000219, or approximately one hospitalised H1N1 case for every 5 000 H1N1 cases. For New York, the CHR was 0.000143, or one hospitalised H1N1 case for nearly every 7 000 H1N1 cases.

For Taiwan, wide credible intervals were found for the first two time points, due to data scarcity then. However, for all three territories, the estimate of CHR would have been precise by 1 August 2009.

4.6.14 Hospital Fatality Ratio (HFR)

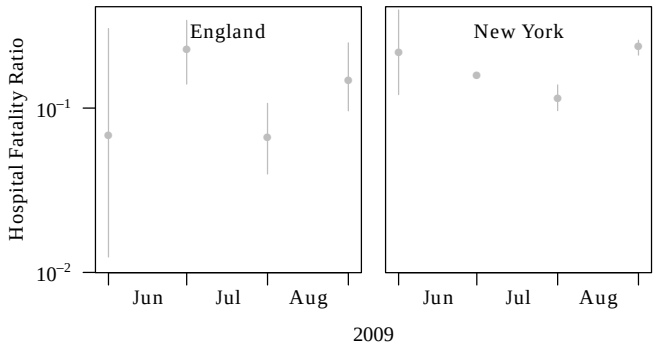

Figure 4.19: **Severity estimate of Hospital Fatality Ratio (HFR).** Features in this figure are as in figure 4.18.

The Hospitalisation Fatality Ratio (HFR) is the ratio of the estimated cumulated number of confirmed deaths due to H1N1 to the estimated daily cumulated number of hospitalization H1N1 cases. The cumulative H1N1 confirmed deaths is estimated by the ODE solution of $R(t)$ using posterior samples of $R_0(c)$ and α with the proportionality parameter, θ_D. The hospitalised H1N1 cases are found from the ODE solution of $I(t)$ and the proportionality parameter, $\theta_{V(c)}$ and the cumulated values form the denominator.

By the last time point, the median HFR estimates for England and New York are 0.165 and 0.235 respectively. These estimates can be understood as having a confirmed death for approximately every 6 and 4 hospitalised H1N1 cases for the two countries respectively.

In the earlier explanation for CHR, there was only about 1 hospitalised H1N1 case for every 7000 H1N1 cases in New York, which might possibly mean that only patients with severe conditions were admitted. This could have indirectly caused a much larger HFR for New York as compared to England.

4.6.15 Case Fatality Ratio (CFR)

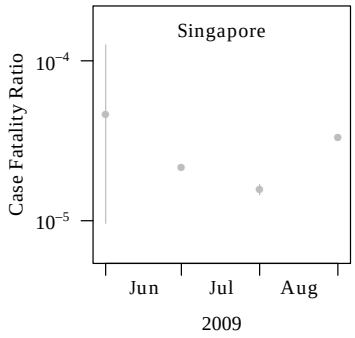

Figure 4.20: **Severity estimate of Case Fatality Ratio (CFR).**
Features in this figure are as in figure 4.18.

The real-time estimate of Case Fatality Ratio (CFR) is the cumulative number of confirmed deaths over the cumulative number of H1N1 infections. Because the mortality proportion, θ_D, was constant for all countries/territories, the severity estimate of CFR is expected to be similar across England, New York and Singapore (we will only present the estimates for Singapore in figure 4.20). At the last time point, the median estimate of CFR for all three territories was 0.000033. This corresponds to a fatal H1N1 case for approximately every 30 000 H1N1 cases.

4.6.16 Basic Reproduction Number (R_0)

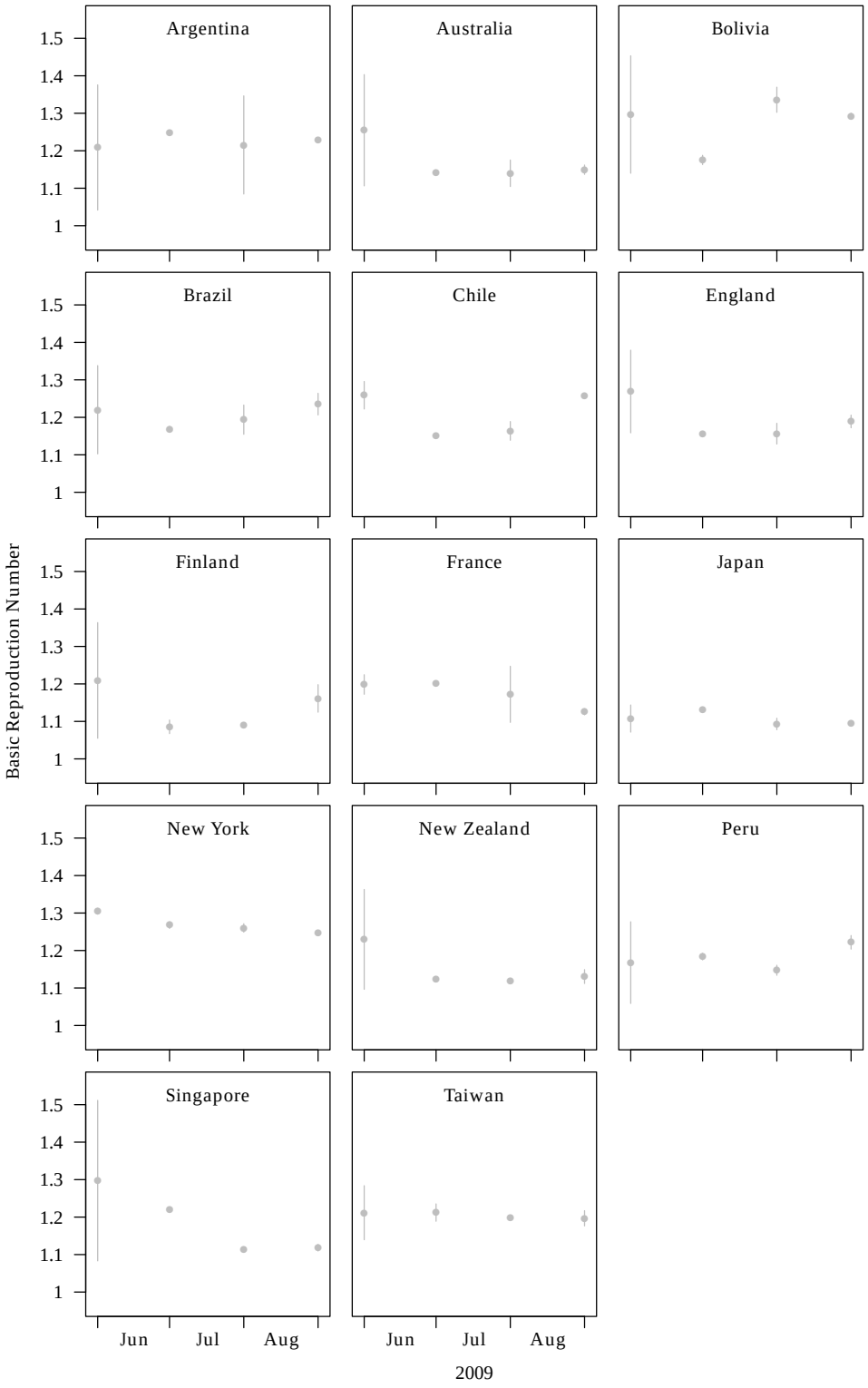

Figure 4.21: **Basic reproduction number, $R_0(c)$.**
Features in this figure are as in figure 4.18. Every country can benefit from this estimate as the model is formulated to synthesize evidence for the actual number of H1N1 $I_c(t)$, as well as the actual number of removed H1N1 cases $R_c(t)$.

Let us focus on the R_0 estimates (which represent the expected number of infections caused by a single infected individual during his or her infectious period) for two cities, New York and Singapore. New York exhibited a complete wave of the pandemic within the shortest time. Their estimate for R_0 was precise at around 1.25 for all four time points considered.

In contrast, in Singapore, where there were absolutely no information at the first time point, the R_0 estimate was to be pooled from the other countries, which resulted in an unduly wide credible interval from 1.08 to 1.51. However, when there were data available at the next time point, the uncertainty reduced substantially.

4.6.17 Final Attack Rate (FAR)

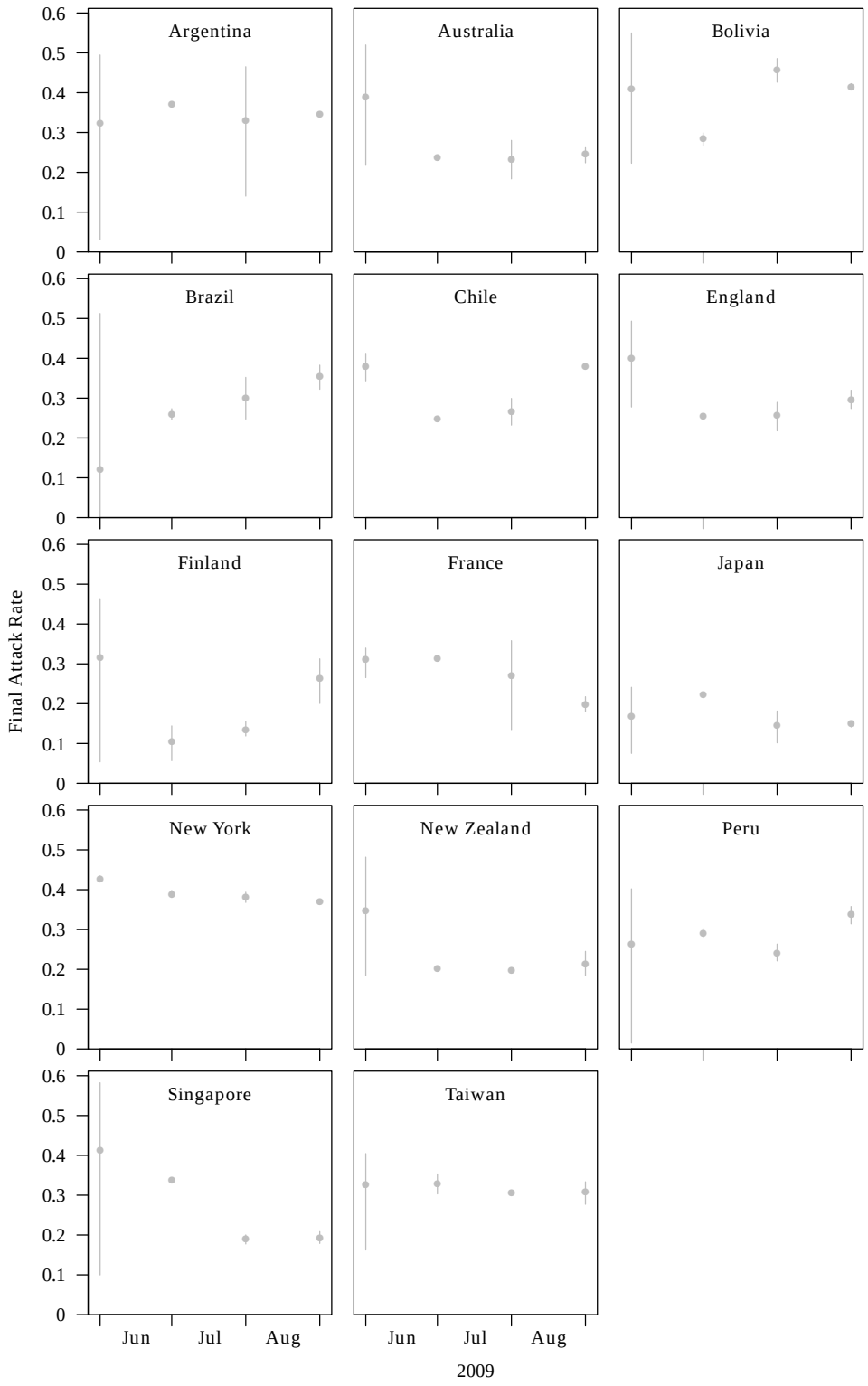

Figure 4.22: **Severity estimate of Final Attack Rate (FAR).**
Features in this figure are as in figure 4.18. The values were computed by dividing the predicted number of individuals in the removed epidemic class at the end of 2009 by the total population size.

Most of the countries have showed that approximately 30% of the people will have been infected by H1N1 by the end of 2009. We would like to draw the attention of this figure to England and New Zealand where the initial FAR was predicted to be higher at 0.398 and 0.358 and dropped to 0.299 and 0.218 at the last time point. The first figures result from the time points when there were limited data available and projections were over-predicted (Figure 4.7 and 4.17)

The higher peak in figure 4.7 for England may be due to the inaccurate proportionality estimate but it was shown in figure 4.18 that $\theta_{V(c)}$ estimates for England were similar across the four time points. Thus, a higher peak in figure 4.7 can be translated to more predicted infections which led to a higher forecast FAR.

In figure 4.17 for New Zealand, not only is the peak of H1N1 cases larger in the first time point than the other time points, the pandemic was also estimated to have started on the 26 Jan 2009 based on the posterior sample of $t_0(c)$ for New Zealand. If the prediction for New Zealand had started early, if would have also accumulated more removed individuals $R_c(t)$ by the end of 2009, resulting in a larger FAR.

By 1 August 2009, our model is able to estimate the median FAR of Singapore to be 0.191, which is comparable to the results of FAR for all adults by Lee et al. (2011) which they based on four different methods. This further suggests that our model is promising in providing appropriate estimations for this H1N1 pandemic.

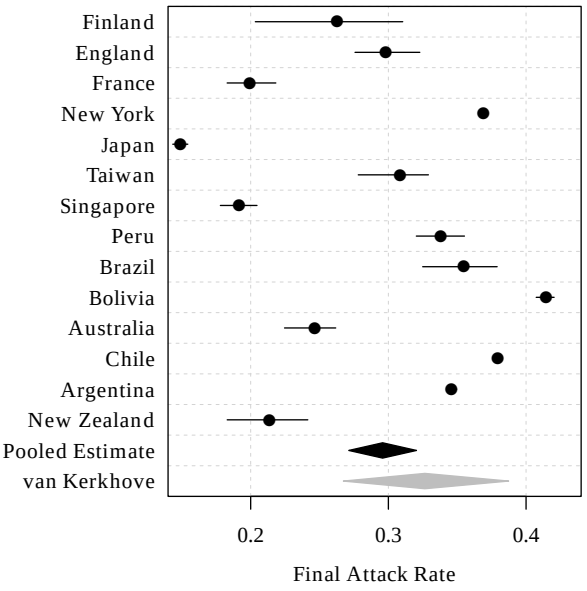

Figure 4.23: **Comparison of the estimated final attack rates for pandemic H1N1 for each country or territory considered by the end of 2009 with the estimate by van Kerkhove et al. (2013).**
Van Kerkhove et al. (2013) used a meta-analysis of seroepidemiological studies, which provide a proxy for the proportion infected and the result is represented by the grey diamond. Dots represent posterior medians and lines 95% equal tailed credible intervals. The black diamond represents the pooled estimate from all countries considered.

Our confidence interval, represented by the black diamond, is calculated by

$$\left(\bar{x} \pm t_{0.025,14-1}\frac{s}{\sqrt{14}}\right)$$

where $\bar{x}$ and s are the sample mean and sample standard deviation of the posterior medians from the 14 countries. Figure 4.23 demonstrates that our results (using data up to 1 September 2009) coincide with the result from the seroepidemiological studies by van Kerkhove et al. (2013), which could only be realised at the end of pandemic by testing the population subset (52 479 sera samples from 27 published/unpublished serological studies from 19 countries/territories) for seropositivity.

4.6.18 Worldwide confirmed death Estimation

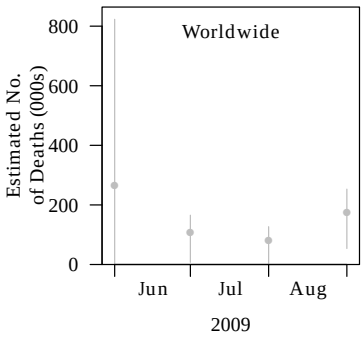

Figure 4.24: **Estimated number of confirmed deaths worldwide by the end of 2009.**
Dots represent posterior medians and lines 95% equal-tailed credible intervals. The estimated number of confirmed deaths worldwide is computed by multiplying the world population (6.8 billion) by θ_D and the proportion in the removed state by the end of 2009.

It is rewarding to see that the median estimated number of confirmed death by the end of 2009 in the fourth time point is 181 300 (95% credible interval 87 200 to 271 900) is comparable with the result found by Dawood et al. (2012) who concluded that the estimated global confirmed deaths should amount to 201 200 (95% confidence interval 151 700 to 575 400). (Contrast to the number of laboratory-confirmed confirmed deaths in 2009: 18 500.) Our forecast for the number of confirmed deaths worldwide are comparable to those of Dawood et al. (2012) even if only the data available by the end of June are used.

4.7 Considerations for Surveillance Network

In summary, although our model does not characterise all the epidemiological quirks of each country's data perfectly, our global estimates of attack rates and various severity measures are accurate by the end of June 2009, a time in which Singapore (as with many other countries) had seen just the beginning of community transmission and there remained great uncertainty about how severe the pandemic would be and what interventions to use. Our model succeeds in this regard by using a Bayesian hierarchical model for a simple epidemic model (the SIR model) and synthesizing information from various types of data that subsequently became available. Had

they been available in real time, the impact of the pandemic control measures on society and the economy might have been much less.

We believe that a network of surveillance data streams from a group of participating countries would be invaluable to provide information on future pandemic outbreaks. However, there may be some issues for considerations before this system can be put up. Wrong or inappropriate data collection would lead to mistaken conclusions, possibly resulting in inappropriate policies being implemented. It was mentioned by Chao et al. (2011) that there was transmission of H1N1 in other parts of the world before the confirmed case in Mexico, and so there is a real need for the network of different countries to contribute surveillance data continuously rather than merely reactively, so that any spread could be detected as early as possible. The pandemic occurred in the Summer in the northern hemisphere, when when most influenza surveillance programmes are on pause (they tend to run only during the Winter months), and to have detected it early— and to have had information on its severity—would require the data collection system to be in place year-round.

Currently, the numbers of cases reported from each country are based on their own authorities' criteria. Differences between countries' criteria may affect the accuracy of our model estimate and limit comparisons between countries. Hence, ideally criteria should be standardised for all countries involved in the surveillance network to follow in determining the classification of individuals and the coverage of each data stream (for instance, the proportion of clinics that contribute within each territory should be known and indicated in the network portal). This would ensure that the reported number of cases can be compared appropriately, in contrast to the ad hoc approach we were forced to use in extracting information from each country's studies.

On top of this, the surveillance network should be validated on routine outbreaks, such as seasonal influenza, before being used in a pandemic outbreak. Other than the department that is responsible for collecting the data, another team could be set up alongside with the involved organizations in each country at different times to collect the data at the same time. A comparison between the data collected from both units can be done and timely feedback could be given to the original data collection team for improvement. This could ensure consistency if the same team is used to validate the data collection from all the countries.

The most feasible way to set up the network would be a confederation of perhaps 20 cities located in as many continents as possible, with a convenient and straightforward online portal for data submission and consistent protocols for data collection and entry. This can boost the cooperation of the different organizations to share data for the network without affecting their daily work routine. We also foresee a great reduction in the time delay from data collection to data entry, permmitting more time for analysis and incorporation within policy.

With a convenient data submission procedure, there would also be security and confidential issues for consideration. We would need to keep the data secure to encourage participation from all the countries and healthcare organizations. Furthermore, in our model, we only need the daily or weekly number of cases from each organization, not data on individual cases to reduce the risk of breaking confidentiality and also ease the management of these large scale data, though such data would be valuable for analysis of clinical features.

Other than the confidentiality concerns, we should also be concerned about the ethical issues of data collection. The main objective should always be treating the patients and the questions or tests for classifying the patients into the various types should not be done at the expense of the patients' health condition or financing.

Although we believe that this proposed idea of a worldwide network of surveillance, we would still need the mutual effort of all participating countries and organizations to contribute to this data base. These collected data would belong to all the countries, hence each country and organization involved in this program should also be given easy access to the stored data. This will greatly encourage more countries' voluntary involvement to get access to real time data for their own research purposes.

While giving access to use the collated data, there arises a problem of ownership for all the data and who should be the provider of this network to hold the responsibility of managing the network. In our opinion, the best candidate for organizing this network would be the WHO, which has the relevant sophisticated technology and also the experience to lead all the participating countries in this long term surveillance project.

United in Tackling Epidemic Dengue (UNITEDengue) is an example of such long term surveillance network which shares dengue surveillance information between their

members (restricted to institution only), including Singapore, Malaysia and Indonesia (Unitedengue, 2014). Another transnational collaboration in the European Union (EU) is the European Centre for Disease Prevention and Control (ECDC) (ECDC, 2014). They will collect and share infectious diseases data between coordinating competent bodies, for example POLYMOD study for new and re-emerging epidemics (Mossong et al., 2008).

4.8 Future Work

With the above considerations for the surveillance network, the model described in this chapter can be used to fit real time data to predict any potential pandemic outbreak. However, future work can be done to improve the model against various assumptions used here.

It was shown by Rhim et al. (2012) that 94% of all the H1N1 cases, from a study conducted in one of the South Korea hospitals during the pandemic, were children and young adults, younger than 40 years old. Similar results have been found elsewhere, for instance in Singapore (Lee et al., 2011). Older people might have a lower risk because of prior exposure to previous pandemics which resulted in pre-existing immunity that are not detectable by cross-reactive antibodies (Dudareva et al., 2011).

If the assumption of homogeneous population were relaxed, the rate of infection should be set for each individual j in country c as

$$\beta_{cj} \quad = \quad \beta_c \phi_j, \tag{4.85}$$

where each individual can be differentiated by another risk score parameter ϕ_j which informs whether the individual j is more easily infected by $\phi_j > 1$. Without this homogeneity constraint, an enormous number of parameters would need to be explored. Moreover, the ODE structure will not hold if each individual is allowed for different risk. Instead, we might simulate this parameter from a log normal distribution,

$$\log \phi_j \quad \sim \quad N\left(1, \sigma^2\right) \ \forall j. \tag{4.86}$$

This distribution will guarantee a positive risk score, ϕ_j, that will either increase the rate of infection for individual j if $\phi_j > 1$ or decrease the rate of infection for

individual j if $\phi_j < 1$. There will be no difference in the infection rate if all the ϕ_j are equal to 1 when $\sigma^2 = 0$. As such, the estimated risk of the whole population will be $\beta_c \sum \phi_j$. From this approach, we will not be able to tell how the rate of infection will differ from each of the different groups of individuals, but we will be able to tell the discrepancy in the rate of infection across individuals by σ^2. Larger σ^2 implies more inconsistency in the rate of infection.

One alternative solution would be to use an age-structured model, rather than one model for the whole population or one rate per individual, with the application of different rates for each groups of people. However the absence of good age-specific data in many locations prohibit this approach, which would result in a substantial increase in the number of parameters to be explored but little additional information to do so. However, this would only be possible if the surveillance network is able to collect these relevant age-specific data to model it, and if data on contacts between people of different ages could be collected for more than the handful of high-income countries that have done so so far (Mossong et al., 2008).

Future work could also involve the exploration of the proportion of people within each country with immunity to counteract the weak assumption for initial susceptible proportion, $s(0) = 1$. This would require a substantial amount of laboratory tests to be conducted in these countries to estimate the required proportions before any pandemic outbreak. Better inference could be made if an appropriate initial number of susceptibles were used. But, we can foresee more technical challenges for solving the ODE as each country would have a different initial condition and separate calculation of ODE solutions is required for each country.

4.9 Conclusion

The results described in this chapter provide an enticing view of the use of hierarchical modelling for emerging infectious disease outbreaks. Not only can we pool information from countries with no or limited data, accurate severity estimates could have been achieved much earlier than was actually the case.

We have also showed how different data types can be integrated together by Bayesian evidence synthesis to capture the evolution of the spread of H1N1 in 2009.

Although MCMC proved difficult to converge due to the high dimensional parameter space, we have successfully incorporated the sequential importance sampling to estimate parameters, pandemic trajectory predictions and severity estimates.

The predicted trajectory for the various data type of all countries in our basket were also preferable after including the numerous factors, like the delay in trajectory for country i since the 1 Jan 2009, $t_0(c)$, and the latent period between the occurrence of the first confirmed death and the first removal event, δ. These estimated results compared well with those by other research teams, using other data, in particular the FAR (Lee et al., 2011) and the estimated worldwide confirmed death due to H1N1 (Dawood et al., 2012).

Chapter 5

Bayesian Optimal Design of Seroepidemiological Studies

5.1 Introduction

Seroepidemiological studies are important for judging population immunity to various infectious diseases based on factors like age. Time and money are needed to acquire and test sera from individuals and, as a result, studies should be designed efficiently: specifically to provide the maximum information within any fixed budget. Straight-forward or naïve designs may be inferior if sera are collected from unnecessary population groups. For instance, for a childhood disease with near 100% prevalence by age a, sampling additional children aged $a+1$ yields little useful information. The objective of this chapter, therefore, is to demonstrate how to find the best designed sample characteristics to estimate prevalence in different groups in the population.

We apply the methodology to design studies to estimate the age-specific prevalence of Enterovirus 71 (EV71), making use of similar studies in other settings using Bayesian hierarchical modelling and assuming the new study's setting is exchangeable with past studies'. This is achievable for this virus because many such serological studies have been carried out in Singapore (Ooi, Phoon, Ishak, & Chan, 2002; Ang et al., 2011) and other countries from East and South East Asia in which EV71 is a topical public health issue, including China (Yu et al., 2011; Zeng et al., 2012), Taiwan (Lu et al., 2002) and Vietnam (Tran et al., 2011).

EV71 is one of the main viruses capable of causing Hand, Foot, and Mouth Disease (HFMD), mostly affecting in children (Shih et al., 2004). The endemic caused by this virus is predominant in children below 5 years old (Ang et al., 2011) and as a result, most existing serological studies on EV71 have concentrated on young age groups.

Most seroepidemiological studies have collected purportedly 'random' samples from healthy individuals without specifying the sampling strategy (Yu et al., 2011; Lu et al., 2002; Tran et al., 2011). Only Zeng et al. (2012) sampled according to the age distribution of Shanghai and Ang et al. (2011) determined the sample size based on the estimated prevalence. None designed an optimal experiment.

Hierarchical models can be built on these data, collected from different countries at different years. So doing facilitates the search for an optimal design as it pools information from these different sources to represent better how the disease will affect the population in the upcoming studies (Chaloner & Verdinelli, 1995). Inferring a hyper-prior distribution for the parameters governing incidence/prevalence allows the resulting information to be fed into a Bayesian optimal design framework whose objective is to select the number of individuals in different demographic segments to maximise the precision in the estimates of the parameters in the future study (Atkinson, Donev, & Tobias, 2007). A well designed study, that in the context of estimating age-prevalence samples some age groups disproportionately, will allow more information to be obtained for the same cost, or the same amount of information at lower cost. In a previous study of the spatio-temporal spread of plant disease by Cook et al. (2007), 25% of the original observations can yield the same information content as a more intensively observed population, if they were arranged carefully. Although these optimal design methods are explored in other areas, especially in engineering (Hollister, Maddox, & Taboas, 2002), it requires considerable adaptation to make them work in the non-linear problems anticipated in seroepidemiology.

Serological studies are very expensive. The sequential serological study of about 2900 individuals from four different groups of people—namely the general population, military personnel, staff from an acute care hospital, as well as residents and staff from long-term care facilities—carried out during the 2009 influenza pandemic cost around S$1.1 million to conduct (Mark I-Cheng Chen, personal correspondence).

At such costs, if the experimental study designs can be more efficient in providing

more information per dollar spent, potentially the saving could be large. Even though optimal design methods would need to be applied differently for each virus under study, we predict that researchers will be able to adapt our methodology to their pathogen and obtain the optimal design that is appropriate to the illness that they are interested in.

5.2 Data from Past Studies on EV71

Data were extracted from previous serological studies performed in Singapore (Ooi et al., 2002; Ang et al., 2011), the People's Republic of China (Lu'an City (Yu et al., 2011) and Shanghai (Zeng et al., 2012)), the Republic of China (i.e. Taiwan) (Lu et al., 2002), and Viet Nam (Tran et al., 2011). They are collated from tables, text or figures found in these journal articles. Due to the irregular structure of the data, we standardised them before analysis. Since EV71 is dominant in young children, we limited attention to data corresponding to children up to the age of 12. As there may be carried over maternal immunity in newborns (Ooi et al., 2002), we also limit attention to sera collected from children at least one year old.

Singapore: In the 1997 Singapore study, convenience sampling was done at a paediatric clinic in the National University Hospital (NUH) (Ooi et al., 2002), among healthy individuals aged below 12 who were attending regular visits to the clinic for vaccination for which they gave a blood sample, residual sera of which was then used in the study. A consequence of the study design is that it may result in a biased sample as those who visit the NUH clinic may not be a good representation of the whole population.

Subsequently, in 2008 to 2010, Singapore's Ministry of Health administered a serology study which acquired samples from 1200 individuals aged below 17 who were at KK Women's and Children's Hospital and NUH for inpatient services or day surgery but not concurrently diagnosed with HFMD (Ang et al., 2011). Based on an estimated prevalence of 33%, they targeted a minimum of 340 sample size for each age group (1–6, 7–12, and 13–17 years) (Ang et al., 2011). The minimum sample size, n, was calculated by having 5% margin of error (ME), where z is 1.96,

$$ME = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{5.1}$$

Note that this was the minimum sample size to obtain a specified accuracy without

the smoothing of a model to borrow strength from nearby ages, and is not directly optimising the design. Furthermore, due to the use of residual sera, there remains the risk of unrepresentativeness.

People's Republic of China: Before the large EV71 epidemic outbreak in China in 2008, Yu et al. (2011) had collected 472 'random' serum samples from children below 15 years who stayed in Lu'an City and had no symptoms of HFMD. The number of samples were equally distributed across each age group (<1, 1, 2–4, 5–7, 8–11, and 12–15 years). Using the same sample size for all age group may result in unnecessary, or insufficient, sampling in certain age groups. After the outbreak, 83 additional serum samples were collected from healthy children below 15 years (Yu et al., 2011). For this, the sampling method was not mentioned in the paper and the sample sizes were no longer uniformly allocated across each age group.

Zeng et al (2011) was the only one amongst the past studies we identified to sample according to the age distribution of the target population (Shanghai). This stratified sampling design using proportional allocation is simple to apply and yet useful since it will sample more from the larger age group to get a better representation of the whole population. However, this might lead to wastage of resources if redundant samples were taken from a large age group. They collected a total of 614 samples from children below 5 years during their health check at the Children's Hospital of Fudan University (Zeng et al., 2012). Again, the samples might be biased due to the single location of serum collection and possibly due to a selection bias in favour of sicker children.

Republic of China: Serum samples were collected in Taipei City, Taiwan, in 1994, 1997 and 1999, the latter to compare the prevalence before and after the major outbreak in 1998 (Lu et al., 2002). There was no mention about how the decision was made for the experimental design in these three years. In 1994, 202 specimens were collected from those above 4 years whereas in 1997 the focus was on those below 4 years, among whom a total of 245 samples were taken (Lu et al., 2002). Participants lived in Taipei but there was no mention if the samples were collected during a routine health check or during a visit to the physician, for instance, due to illness (Lu et al., 2002). Additional effort to guard against biased sampling was made in 1999 by extending sampling to 1 258 participants from the city as well as nearby

non-urban areas (Lu et al., 2002).

Viet Nam: 794 samples were collected from the Hung Vuong Obstetric Hospital, Ho Chi Minh City, in 2007 (Tran et al., 2011). Again, there was no indication of how the design was set up and it may be convenience sampling from those who visited the hospital amounting to almost equal sample sizes in the three age groups.

As the age intervals of the samples vary across datasets, we represent the intervals by their respective start and end age in months. The unit is chosen to be months and not years because some datasets had age gaps in months, and so no information would be lost in this representation.

5.3 Hierarchical modelling of past studies

Leveraging on past experimental studies, we fit a hierarchical model to those data discussed in the previous section. Here, we use a discrete time survival analysis approach to transform the data into information that will later be used to design an optimal experiment using Bayesian decision theory.

T is defined as the age when an individual gets infected. In a continuous survival analysis, the hazard function, $h(t)$, is the instantaneous rate at which failure (here, infection) occurs at time t. In our discrete time version, $h(t)$ takes the form of a probability of infection within the 1-year interval starting at integer time t (where data are in fractions of a year, the hazard is assumed to be constant throughout the year, see later for details) conditional on non-infection to time $t-1$ (Singer & Willett, 1993), i.e.

$$h_t = \Pr(T = t | T > t - 1). \tag{5.2}$$

The infection risks are allowed to vary non-parametrically from one year to another, from one country to another. Since h_t is seen as a probability, the values are restricted to $0 \leq h_t \leq 1$.

A negative serology test implies that infection has not occurred for this individual by age i. The survival function, $S(t)$, is the probability of surviving (i.e. not experiencing the event) to time t. In the discrete case, S_t is the probability of not being infected by age t. By the conditional probability relationship, S_t can be related

to the hazard rate by

$$\begin{aligned} S_t &= \Pr(T > t) && (5.3)\\ &= \Pr(T > t | T > t-1) \Pr(T > t-1) && (5.4)\\ &= (1 - \Pr(T = t | T > t-1)) \Pr(T > t-1) && (5.5)\\ &= (1 - h_t) S_{t-1}. && (5.6) \end{aligned}$$

Iterating, the relationship between the survival function and hazard rate for a discrete time survival model is

$$S_t = \prod_{j=0}^{t-1} (1 - h_j). \quad (5.7)$$

Intuitively, survival to age t means not having been infected at any age before t, implying that the survival probability up to age t is the product of all the $(1 - h_j)$ up to age $t-1$, where $(1 - h_j)$ represents the probability of not being infected at age j.

The density (technically, mass) for the random variable T, represented by f_t, is the probability of an individual being infected at age t,

$$\begin{aligned} f_t &= \Pr(T = t) && (5.8)\\ &= \Pr(T = t | T > t-1) \Pr(T > t-1) && (5.9)\\ &= h_t S_{t-1} && (5.10)\\ &= h_t \prod_{j=0}^{t-1} (1 - h_j). && (5.11) \end{aligned}$$

Our data include the number of positive samples for individuals from age t to $t+1$ as $x_{(t,t+1)}$ and the total number of samples contributed by the individuals from age t to $t+1$ as $n_{(t,t+1)}$. We will model $x_{(t,t+1)}$ as a binomial distribution with $n_{(t,t+1)}$ trials and the probability of success as the probability of being identified as infected in the age interval $p_{(t,t+1)}$,

$$x_{(t,t+1)} \sim \text{Binomial}\left(n_{(t,t+1)}, p_{(t,t+1)}\right). \quad (5.12)$$

This may be appropriate if the number of individuals in the sample is small relative to the population as a whole, so that the effects of non-independence can be ignored. Being identified as infected at age t implies that the individual was infected

in age t or before, so p_t is represented by f_t as follows:

$$\begin{aligned} p_0 &= h_0 \\ &= f_0 \qquad (5.13) \\ p_1 &= (1-h_0)\,h_1 + h_0 \\ &= f_1 + f_0 \qquad (5.14) \\ p_t &= (1-h_0)\ldots(1-h_{t-1})\,h_t + \ldots + (1-h_0)\,h_1 + h_0 \\ &= f_t + \ldots + f_0. \qquad (5.15) \end{aligned}$$

Most data come reported to a yearly interval, but there were some cases of irregular intervals. In particular, we set $p_{(t,t+1)} = p_t$ for the small number of datasets with an age interval of less than a year, while if the interval is more than a year, we will take the average of the yearly probabilities of being seropositive, which assumes equal sampling across the age range. For example, the probability of being identified as infected in a three year age interval from 7 to 10 is $p_{(7,10)} = \frac{1}{3}(p_7 + p_8 + p_9)$.

Different datasets have information up to different maximum ages and in addition may lack data for some age groups within the age range. However, hazards for gaps in the data at earlier ages provide information at later ages by contributing to the density calculation f_t for the later age groups t. In addition, by taking a hierarchical approach (using MCMC to sample the resulting parameter space) the inference routine will pool information across different countries to fill the gaps for countries with missing data for those countries.

Due to the condition of $0 \leq h_{tj} \leq 1$ for hazard rate of age t in dataset j, the logit transformation of h_{tj},

$$\text{logit}(h_{tj}) = \log \frac{h_{tj}}{1-h_{tj}}, \qquad (5.16)$$

will lead to parameters with support on the real line. This allows a multivariate normal hyper-distribution for $\text{logit}(\boldsymbol{h_j})$ where $\boldsymbol{h_j} = (h_{0j}, h_{1j}, \ldots, h_{tj})$ will be governed by hyper-parameters characterizing the mean, standard deviation and correlation between the parameters,

$$\text{logit}\begin{pmatrix} h_{0j} \\ h_{1j} \\ \vdots \\ h_{tj} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_t \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{10} & \ldots & \sigma_{t0} \\ \sigma_{01} & \sigma_1^2 & \ldots & \sigma_{t1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{0t} & \sigma_{1t} & \ldots & \sigma_t^2 \end{pmatrix}\right) \qquad (5.17)$$

where $\sigma_{kl} = \rho_{k,l}\sigma_k\sigma_l$ and $\sigma_{kl} = \sigma_{lk}$.

As there is no prior knowledge for the hyper-parameters, non-informative hyper-prior distributions are used. Since μ_k can be any real numbers, a normal distribution with a wide range centred at 0 to allow for both positive and negative numbers is used,

$$\mu_k \sim N\left(0, 100^2\right). \tag{5.18}$$

Since the standard deviation should be positive, hyper-prior for σ_k is assigned an exponential distribution,

$$\sigma_k \sim Exp\,(1) \tag{5.19}$$

where a larger hyper-parameter σ_k will indicate a more differences in the hazard rate parameter between countries. Similarly, the correlation hyper-parameters $\rho_{k,l}$ govern the relationship between the parameters h_k and h_l. If they are positively correlated, $0 < \rho_{k,l} < 1$; if they are negatively correlated, $-1 < \rho_{k,l} < 0$. A flat hyper-prior is used on all the correlation hyper-parameter,

$$\rho_{k,l} \sim U\,(-1, 1)\,. \tag{5.20}$$

With the above model, we employ a Metropolis-Hastings algorithm with MCMC methodology to explore the parameters and hyper-parameters. We discard 1% of the sample as burn-in, select the choice of initial values to be close to the actual posterior (see below), allowing the chains to converge quickly. Thinning is used to reduce the correlation between subsequent stored values, by retaining only at every 10th iteration.

To make the initial values of the parameters suitable, we take a point estimate of the hazard rate in an interval i to be

$$h_i = \frac{d_i}{l_i\left(n_{i-1} - \frac{1}{2}d_i\right)}, \tag{5.21}$$

where the number of positive samples in the interval is d_i, the length of the interval, l_i, and the total number of samples is n_{i-1}.

Having selected the initial values for the parameters, we take the following steps in the MCMC routine:

1. Using the initial values for the hazard rates $\boldsymbol{h_j^0} = \left(h_{0j}^0, h_{1j}^0, \ldots, h_{ij}^0\right)$ for country j where the maximum age of the collected data is i and the randomly chosen initial values of the hyper-parameters $(\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0})$, calculate the log likelihood density, $\log f\left(D_j|\boldsymbol{h_j^0}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}\right)$, log prior density, $\log f\left(\boldsymbol{h_j^0}|\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}\right)$, log hyper-prior density, $\log f\left(\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}\right)$, as well as the log posterior density, $\log f\left(\boldsymbol{h_j^0}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}|D_j\right)$, for subsequent use.

   a) The log likelihood density is calculated using the built-in binomial distribution function, `dbinom`, in R, where the data for country j, D_j, will include the number of trials, $n_{(i,i+1)}$, and the actual number of infected, $x_{(i,i+1)}$, and the probability of being infected in age i, $f_{(i,i+1)}$, using the hazard rates $\boldsymbol{h_j^0} = \left(h_{0j}^0, h_{1j}^0, \ldots, h_{ij}^0\right)$.

   b) The log prior density is calculated using the multivariate normal distribution function, `dmvnorm`, from the `mvtnorm` package in R. Using the hyper-parameters, $\boldsymbol{\mu^0}$, as the mean vector, we also calculate the covariance matrix using the other hyper-parameters $(\boldsymbol{\sigma^0}, \boldsymbol{\rho^0})$ as stated in the model earlier. The covariance matrix is singular if there are non-positive eigenvalues, in which case we let the log prior density be $-999\,999$, an arbitrary negative number which approximates 0 on exponentiation. The hazard parameter is logit transformed before calculating the prior density.

   c) The log hyper-prior density is calculated as normal distributions for $\boldsymbol{\mu^0}$, exponential distributions for $\boldsymbol{\sigma^0}$ and uniform distributions for $\boldsymbol{\rho^0}$ using the build-in distribution function in R.

   d) The pseudo log posterior density (ignoring a constant) is computed by summing the three log densities. Note that posterior density is only proportional to the product of the three densities,

$$\begin{aligned} f(\boldsymbol{h_j^0}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}|D_j) \quad \propto \quad & f(D_j|\boldsymbol{h_j^0}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}) \cdot \\ & f(\boldsymbol{h_j^0}|\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}) \cdot f(\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}). \quad (5.22) \end{aligned}$$

   The proportionality constant that leads to the actual posterior density is not available but cancels in the step of calculating the acceptance probability.

2. Propose new hazard parameter values, $\boldsymbol{h_j^*}$, for country j from a multivariate normal distribution with mean $\boldsymbol{h_j^0}$. In pilot runs, the covariance matrix of the proposal distribution is initially a diagonal matrix, since the correlation of hazard rates across ages is unknown. This covariance matrix will be improved in subsequent rounds by the information from the posterior samples of previous, pilot rounds, leading to better proposals in the later rounds as the covariance matrix becomes a better representation of the actual posterior distribution of the hazard rates. After proposing changes, the legality of the move is assessed (i.e. that $0 < \boldsymbol{h_j^*} < 1$) and the proposal is rejected if the condition is not fulfilled. A multivariate normal proposal distribution is preferred due to its symmetrical property, in the Metropolis- Hastings algorithm, the acceptance probability will reduce to

$$P_{\text{acc}} = \min\left(1, \frac{f(\boldsymbol{h_j^*}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}|D_j)}{f(\boldsymbol{h_j^0}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}|D_j)}\right) \tag{5.23}$$

$$= \min\left(1, \frac{f(D_j|\boldsymbol{h_j^*}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}) \cdot f(\boldsymbol{h_j^*}|\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0})}{f(D_j|\boldsymbol{h_j^0}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}) \cdot f(\boldsymbol{h_j^0}|\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0})}\right). \tag{5.24}$$

With probability P_{acc}, the proposed $\boldsymbol{h_j^*}$ will be accepted and updated as $\boldsymbol{h_j^1}$, otherwise, the proposal is rejected and we let $\boldsymbol{h_j^1} = \boldsymbol{h_j^0}$. This step will be repeated for each dataset sequentially.

3. After updating all the hazard rates, we propose the hyper-parameters $(\boldsymbol{\mu^*}, \boldsymbol{\sigma^*}, \boldsymbol{\rho^*})$ using independent normal proposal distributions with arbitrary initial standard deviations for all the hyper-parameters. Proposed values should satisfy $\boldsymbol{\sigma^*} > 0$ and $-1 < \boldsymbol{\rho^*} < 1$, and are otherwise rejected. The acceptance probability can be determined by

$$P_{\text{acc}} = \min\left(1, \frac{f(\boldsymbol{h^1}, \boldsymbol{\mu^*}, \boldsymbol{\sigma^*}, \boldsymbol{\rho^*}|D_j)}{f(\boldsymbol{h^1}, \boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}|D_j)}\right) \tag{5.25}$$

$$= \min\left(1, \frac{f(\boldsymbol{h^1}|\boldsymbol{\mu^*}, \boldsymbol{\sigma^*}, \boldsymbol{\rho^*}) \cdot f(\boldsymbol{\mu^*}, \boldsymbol{\sigma^*}, \boldsymbol{\rho^*})}{f(\boldsymbol{h^1}|\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0}) \cdot f(\boldsymbol{\mu^0}, \boldsymbol{\sigma^0}, \boldsymbol{\rho^0})}\right). \tag{5.26}$$

This does not involve recalculating the likelihood and hence is computationally efficient, so the hyper-parameters are updated 50 times for every round of proposal for the hazard parameters.

4. Step 2 and 3 will be repeated 10 000 times and the covariance matrix and standard deviations of posterior samples will be used in subsequent proposal distributions.

5. The improvement of proposal distribution will be done for 10 rounds. The final round, with Markov chains that have apparently converged to the posterior distribution, is used to create samples of all the parameters and hyper-parameters, which are stored for the second stage of the analysis—designing a future study efficiently.

5.4 Optimal design of a future serological study

Instinctively, more information can be obtained as more samples are collected. However, as the number of serological tests varies proportionally with the study cost, and there is always a limit to the amount of money that can be spent on a single epidemiological study (experiment), suggesting only a fixed total number of serological tests can be performed. Naturally, a design for the experiment that gives the most information using a fixed expense is preferred to one giving less.

Atkinson (2001) argues that, in this context, we should fix the total number of serological tests while optimising over the characteristics of each participant to be tested. Treating each age as a cluster, in Atkinson's terminology, n_i represents the number of serological test needed for individuals of age i, which is subject to the restriction on the total number of tests, $n = \sum_i n_i$. (This assumes no difference in the cost for sampling in different age groups, which allows us to focus on the number in each cluster rather than the cost). In this stage of the analysis, we will explore different possible combinations of $(n_0, n_1, \ldots, n_{11})$, i.e. up to age 12, with the condition that they should sum up to a certain sample size.

Optimal designs are dependent on the precise choice of model (Rasch, Pilz, Verdooren, & Gebhardt, 2011). Later in this chapter, we show that the hierarchical model described earlier gives a good characterisation of EV71 prevalence for countries with the same characteristics as those we obtained data for.

In classical optimal design, in the situation where the optimising of the function that forms the design criterion relies on the exact values of the (hazard, in our case) parameters, Berger and Wong (2005) argue that point estimates of the parameters, i.e. from an extremely informative prior distribution, should be used to find the deterministic solution of the objective function, as this might more efficiently achieve the

best optimal design. In contrast, in Bayesian optimal design (Chaloner & Verdinelli, 1995), we typically sample parameter values from an informative prior, perhaps drawing on an analysis of past data, and simulate multiple datasets for each design, taking the average of a defined utility function over repeat Monte Carlo sampling for each point in the design space to obtain the optimal design, a more computationally intensive solution (Müller, Sansó, & De Iorio, 2004). The next subsection described classical experimental design, followed by a description of Bayesian optimal design.

5.4.1 Classical optimal design

Classical optimal design is categorised according to the objective function using so-called alphabetic optima. According to Atkinson (2007) , A-optimality is based on the criterion of minimizing the sum or average of all the variances of the parameter estimates (typically MLEs). The argument is that for an optimal design to provide us with the most information on all parameters, the resulting sampling variance should be small, which coincides with the A-optimality criterion.

If we only required a certain parameter to be precise, C_i-optimality will choose the best design based on the decision that can minimise the variance of parameter i (Rasch et al., 2011). A variant is E-optimality which minimises the variance of the parameter with the poorest precision (inverse variance) (Atkinson et al., 2007). This has a stronger benchmark than C_i-optimality criterion since it ensures that no particular parameters will have too much sampling error under the selected optimal design. It does however require that parameters have a common scale to facilitate sensible comparison.

The D-optimality criterion is the most popular classical design criterion, and motivates the optimal design in our project. According to the General Equivalence Theorem, the optimal design will minimise the imprecision or in other words, maximise the expected utility where utility is the determinant of the information matrix (Atkinson et al., 2007). As the expected utility for the design space may not be smooth, we will probe into different approaches for exploring an uneven design space.

The observed Fisher Information is the expectation of the negative second derivative of the log likelihood. To see why this is a sensible choice of objective function, consider the one dimensional case, where the reciprocal of the Fisher Information is the variance of the MLE (Efron & Htnkley, 1978), and so maximizing the determinant

of the information matrix will minimise the uncertainty in the parameter.

5.4.2 Bayesian optimal design

Classical experimental design makes use of point estimates for the parameters but does not account for the uncertainty in them. The Bayesian framework provides a more natural framework to overcome this problem by allowing the design to account for the uncertainties in the parameters using their posterior distribution.

In the Bayesian paradigm, information on the model's parameters is encapsulated by the prior or posterior distribution. In the context of a decision problem, the information *prior* to observing the outcome is relevant. If data are available to guide the decision, this 'prior' is the distribution *after* observing them, i.e. the posterior. The optimal decision, $D^\star$, is that which maximises the expected utility u (or objective function) over the uncertainty in the outcome X and in the parameters, θ (Cook, Gibson, & Gilligan, 2008), i.e.

$$D^\star \quad = \quad \arg\max \int \int u(X, D) p(X|\theta, D) p(\theta|D) d\theta dX. \tag{5.27}$$

In practice, typically the integration is done using Monte Carlo sampling, assuming the prior can be sampled, along with the data conditional on the prior. In the context of experimental design, the decision is the design (Chaloner & Verdinelli, 1995).

This approach requires defining a utility function that allows a good design that provides substantial information to correspond to a high expected utility value (Atkinson et al., 2007; Verdinelli & Kadane, 1992). According to Bernardo (1979), translated to the seroepidemiological scenario, when D is the decision of how to divide the sample size between age groups, $p(\theta|D)$ is the reported posterior density function of the parameters resulting from the experiment conducted under the decision D, the utility will be a function of the reported density function and the decision, $u(X, D)$.

We foresee that the design space will be complicated due to its high dimensionality, which has implications for how the design space is searched for the optimum. The method to explore the design space will be examined in the later subsections. As the estimated expected utility for the design space will not be smooth, we will probe different approaches for exploring the design space that account for this imperfect observation of the expected utility.

In simulating the utility for one realisation, for a specific design, we use the set of sample sizes allocated for each age group defined by the design,$(n_0, n_1, \ldots, n_{11})$, draw hazard rates from the hyper-posterior distribution of the hierarchical model, reflecting an assumption that the study we are designing is exchangeable with those previously analysed, and use these to simulate the number of positive samples, $(x_0, x_1, \ldots, x_{11})$, in each group. If the experimental data are to be analysed classically, as is frequently the case in epidemiological research, the appropriate objective function represents how much information will be captured in a classical analysis. To derive this, the likelihood function (of a model, either the generating model or a simpler model used for reporting) under this experiment can be used to compute the MLE, $\hat{\boldsymbol{h}}$. The utility function, inspired by the D-optimality criterion, is the determinant of the information matrix under the decision of D and the MLE $\hat{\boldsymbol{h}}$, where $L = \log p(\boldsymbol{h}|D)$ (Bernardo, 1979),

$$u(X, D) = \det\left(-\left(\begin{array}{cccc} \frac{d^2L}{dh_0^2} & \frac{d^2L}{dh_0 dh_1} & \cdots & \frac{d^2L}{dh_0 dh_{11}} \\ \frac{d^2L}{dh_1 dh_0} & \frac{d^2L}{dh_1^2} & \cdots & \frac{d^2L}{dh_1 dh_{11}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^2L}{dh_{11} dh_0} & \frac{d^2L}{dh_{11} dh_1} & \cdots & \frac{d^2L}{dh_{11}^2} \end{array} \right)\Bigg|_{\boldsymbol{h}=\check{\boldsymbol{h}}} \right). \quad (5.28)$$

The algorithm to maximise the expected utility over a set of designs is:

1. Set the first decision of sample sizes allocation, $D^1 = (n_0^1, n_1^1, \ldots, n_{11}^1)$.

2. By assumption the logit of the hazard rates follows a multivariate normal distribution governed by hyper-parameters. We sample a set of hyper-parameters from the hyper-posterior distributions and compute the covariance matrix. With the mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, simulate $\text{logit}(h_{ij})$ from the multivariate normal distribution and perform an inverse logit transformation to derive the set of true hazard rates under this experiment, using $h_{ij} = \frac{1}{1+\exp(-\text{logit}(h_{ij}))}$.

3. With this set of simulated true hazard rates, the probability of having a positive test at age i, p_i, can be calculated from the hazard rates directly. The probability of having a positive test can be compared with the propensity, ϕ (to be explained later) to simulate the experimental data, where $\phi_l < p_i$ will indicates

that individual l of age i has a positive test result. Tallying over individuals, we will get the data of this experiment, $\boldsymbol{x} = (x_0, x_1, \ldots, x_{11})$.

4. With the design, $\boldsymbol{n}$ and data, $\boldsymbol{x}$, the MLE, $\hat{\boldsymbol{h}}$, is found using a numerical method (see later).

5. The utility value for these simulated data is the determinant of the information matrix at the MLE.

6. Repeat step 2 to 5 for the same decision $D^1 = (n_0^1, n_1^1, \ldots, n_{11}^1)$ a large number of times (I use 100). The expected utility for this decision is the mean of the sampled utilities.

7. Repeat step 1 to 6 for the kth decisions D^k for $k = 2, 3, \ldots$ to acquire their respective expected utilities. The same hazard rates and propensities will be used for all decisions to reduce unnecessary variability in the expected utilities. Thenceforth, the Bayesian optimal design can be identified by maximum expected utilities.

Here, we shall demonstrate how to calculate the utility for a given dataset. The number of positive samples for age i , x_i, follows by assumption a binomial distribution with n_i number of individuals of age i at risk and p_i as the probability of being identified as infected by age i. The likelihood of the experimental data up to age 12 is

$$l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p}) = \prod_{i=0}^{11} \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i} . \tag{5.29}$$

Taking logarithms of the likelihood function and dropping the constant of proportionality that does not change with different values of p_i,

$$\log l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p}) = \sum_{i=0}^{11} (x_i \log p_i + (n_i - x_i) \log(1 - p_i)) . \tag{5.30}$$

The first derivative of the log likelihood function with respect to the hazard parameters h_j, where $j = 0, 1, 2, \ldots, 11$, is

$$\frac{d}{dh_j} (\log l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p})) = \sum_{i=0}^{11} \left(\frac{x_i}{p_i} \cdot \frac{dp_i}{dh_j} + \frac{n_i - x_i}{1 - p_i} \cdot \left(- \frac{dp_i}{dh_j} \right) \right) \tag{5.31}$$

$$= \sum_{i=0}^{11} \left(\frac{x_i}{p_i} - \frac{n_i - x_i}{1 - p_i} \right) \cdot \frac{dp_i}{dh_j} . \tag{5.32}$$

Similarly, we can find the second derivative of the log likelihood function with respect to the hazard parameter h_k, where $k = 0, 1, 2, \ldots, 11$

$$\begin{aligned}\frac{d}{dh_k}\left(\frac{d}{dh_j}\left(\log l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p})\right)\right) &= \sum_{i=0}^{11}\left(-\frac{x_i}{p_i^2}\left(\frac{dp_i}{dh_k}\right) - \frac{n_i - x_i}{(1-p_i)^2}\left(\frac{dp_i}{dh_k}\right)\right)\cdot\frac{dp_i}{dh_j} \\ &\quad + \sum_{i=0}^{11}\left(\frac{d}{dh_k}\left(\frac{dp_i}{dh_j}\right)\right)\left(\frac{x_i}{p_i} - \frac{n_i - x_i}{1-p_i}\right) \qquad (5.33) \\ &= \sum_{i=0}^{11}\left(-\frac{x_i}{p_i^2} - \frac{n_i - x_i}{(1-p_i)^2}\right)\cdot\frac{dp_i}{dh_j}\cdot\frac{dp_i}{dh_k} \\ &\quad + \sum_{i=0}^{11}\left(\frac{d^2 f_i}{dh_k dh_j}\right)\left(\frac{x_i}{p_i} - \frac{n_i - x_i}{1-p_i}\right). \qquad (5.34)\end{aligned}$$

We calculate the first and second derivatives of p_i with respect to h_j and h_k as a prelude to calculating the above second derivative of the log likelihood function. Since $p_i = \sum_{j=0}^{i} f_j$ and $f_i = h_i \prod_{j=0}^{i-1}(1-h_j)$, we let a pseudo function g_i to represent $\prod_{j=0}^{i-1}(1-h_j)$ for the subsequent calculation. Then, the probability of being identified as infected in age i is

$$\begin{aligned}\because p_{i-1} &= (1-h_0)\ldots(1-h_{i-2})\,h_{i-1} + \ldots + (1-h_0)\,h_1 + h_0 \qquad (5.35) \\ p_i &= (1-h_0)\ldots(1-h_{i-1})\,h_i + \ldots + (1-h_0)\,h_1 + h_0 \\ &= g_{i-1}h_i + (1-h_0)\ldots(1-h_{i-2})\,h_{i-1} + \ldots + (1-h_0)\,h_1 + h_0 \\ &= g_{i-1}h_i + p_{i-1}. \qquad (5.36)\end{aligned}$$

The first derivative can be obtained iteratively

$$\frac{dp_i}{dh_j} = \begin{cases} h_i\frac{dg_{i-1}}{dh_j} + \frac{dp_{i-1}}{dh_j} & j \le i \\ g_{i-1} & j = i \\ 0 & j > i \\ 1 & j = i = 0. \end{cases} \qquad (5.37)$$

Likewise, the iterative technique can be applied to the second derivative where

$$\frac{d}{dh_k}\left(\frac{dp_i}{dh_j}\right) = \begin{cases} h_i\cdot\frac{d}{dh_k}\left(\frac{dg_{i-1}}{dh_j}\right) + \frac{d}{dh_k}\left(\frac{dp_{i-1}}{dh_j}\right) & j \ne k, j \ne i, k \ne i \\ \frac{dg_{i-1}}{dh_j} + h_i\cdot\frac{d}{dh_k}\left(\frac{dg_{i-1}}{dh_j}\right) + \frac{d}{dh_k}\left(\frac{dp_{i-1}}{dh_j}\right) & k = i, j < i \\ \frac{dg_{i-1}}{dh_k} & j = i \ne 0, k < i \\ 0 & j = i = 0, j > i. \end{cases} \qquad (5.38)$$

This re-arrangement is beneficial because the derivative of the pseudo function g_i is simple. To elucidate this pattern, the first derivative of the function g_i with respect to h_j can be represented by

$$\frac{dg_i}{dh_j} = \begin{cases} -\frac{g_i}{1-h_j} & j \leq i \\ 0 & \text{otherwise.} \end{cases} \tag{5.39}$$

Correspondingly, the second derivative of the function g_i will result in the following two cases

$$\frac{d}{dh_k}\left(\frac{dg_i}{dh_j}\right) = \begin{cases} \frac{g_i}{(1-h_j)(1-h_k)} & j \neq k, j \leq i, k \leq i \\ 0 & \text{otherwise.} \end{cases} \tag{5.40}$$

Putting all of these together, we can calculate the second derivative of the log likelihood with respect to h_j and h_k. Inserting the MLE then gives the utility.

The stochasticity in the utility surface makes identification of a maximum difficult. To remedy this, we stabilised the information content between neighbouring design points as follows. Before starting the algorithm, we simulate hazard rates from the hyper-posterior sample to act as the true hazard rate, storing them and reusing them across the design space. This induces correlation between neighbouring points that makes identification of which is greater easier, for the same reason that paired t-tests typically have more power than two-sample t-tests. Furthermore, to reduce the variability in the information, we use fixed propensity scores for the individuals at each design.

A propensity score ϕ_l is associated to each individual l. The propensity scores, uniformly distributed from 0 to 1, are compared against the computed probability of being identified as infected at age i, p_i. Since each individual at age i will test positive with probability p_i, we can let the individual l of age i be tested positive if $\phi_l < p_i$ and contribute to the number of positive samples $\boldsymbol{x}$. Using the same set of propensity score to get the simulated datasets will induce a positive correlation between them and result in smoother utility surface, while maintaining the correct marginal distributions of the utility at any point, more computationally efficiently than using larger number of sets of simulated data.

Apart from the above set-up to simulate potential future data, $\boldsymbol{x}$, the MLE of the hazard parameters is also required. In the one dimensional case, the MLE is the

point where the likelihood is at its highest point, as evaluated numerically e.g. with a grid search, which is also the solution of the first derivative of likelihood function when equated to zero. But due to the complexity of the likelihood function, it is difficult to solve for the MLE (and more to the point, time consuming). Thus, we have explored several approaches to find the MLE.

5.4.3 MLE search using Newton-Raphson method

The Newton-Raphson method is a numerical approach to get to MLE through a deterministic series of iterative moves (Ypma, 1995). This method often works well for multidimensional problems. For demonstration, consider just two age groups to improve cognitive ease of visualization. Let the number of samples from age i be $n_i = 250$ where $i = 0, 1$, the true hazard rate is set at 0.1 and 0.05 respectively. The steps are as follows.

1. Using the true hazard rate, calculate the probability of being infected at each age and simulate the number of infected individuals using a binomial distribution. The initial parameters values are chosen to be the true hazard rate values, $h_0^0 = 0.1$ and $h_1^0 = 0.05$. Because the Newton-Raphson method is a stepwise procedure, we can get to the MLE faster if appropriate initial parameter values are chosen (Lauritzen, 2008). In this case, because the data were simulated, the true values are known.

2. Calculate the first derivative of the log likelihood function, $\frac{d}{dh_j}\left(\log l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p})\right)$ for $j = 0, 1$, and the second derivative of the log likelihood function $\frac{d}{dh_k}\left(\frac{d}{dh_j}\left(\log l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p})\right)\right)$ for $j = 0, 1$ and $k = 0, 1$ using the initial values of h_0^0 and h_1^0. These are commonly termed the score $S(\boldsymbol{h})$ and observed information $J(\boldsymbol{h})$ respectively.

3. The parameter values of the next iterative step, $\boldsymbol{h^1} = (h_0^1, h_1^1)$, are set to be

$$\boldsymbol{h^1} \quad = \quad \boldsymbol{h^0} + \lambda J(\boldsymbol{h^0})^{-1} S(\boldsymbol{h^0}), \tag{5.41}$$

where $0 < \lambda \leq 1$ is a constant that varies directly with the size of the steps.

4. Repeat step 2 and 3 until the sum of the absolute first derivative is less than a small preselect value $\epsilon = 0.0001$. If all the first derivatives are close to zero, the parameter values will be close to the MLE.

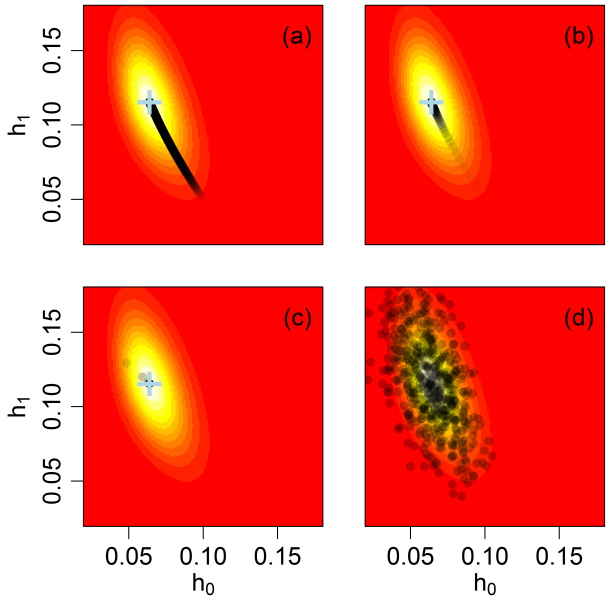

Figure 5.1: **Effects of different λ (controlling the size of steps) used in the Newton-Raphson method.**
Panel (a) shows the stepwise moves when $\lambda = 0.01$, (b) is when $\lambda = 0.1$ and (c) is when $\lambda = 0.9$. Panel (d) shows the hazard distribution when we simulate 500 points from multivariate normal distribution centred at the MLE parameter values and covariance matrix from the observed information based on the MLE parameter values. In the background image plots in all the panels yellow corresponds to a high magnitude of the likelihood, and red low. The grey dots represent the stepwise movement of the particles, with increasing intensity of darkness, whereas the light blue cross shows the position of the MLE.

In figure 5.1, we explored the values of λ that control the size of the stepwise movement of the parameter values. Because of the small value of 0.01 used in panel (a), the steps are so small that it required 1 373 steps to reach the MLE. The moves of the dots were so close that they overlapped. This is inefficient as computer time will be wasted to compute the score statistics $S(\boldsymbol{h})$ and observed information $J(\boldsymbol{h})$ at every step when there is not much change in every subsequent iterations.

In panel (b), $\lambda = 0.1$ only required 132 reasonable steps to reach MLE. Although λ is increased tenfold, it is clear that the path is effectively the same as that in panel (a). Panel (c) shows that if λ is larger, $\lambda = 0.9$, it only required 9 steps to reach

MLE. But the big steps had a high chance of moving immediately out of the region of high likelihood. Propitiously, in the subsequent steps, the algorithm is still able to bring the point back to the MLE parameter values as long as the routine has not stop. However, we experienced problems in the high dimensional case as the routine did not recover from overly big steps.

In panel (d), we tried to validate the complicated calculation of the second derivative of the log likelihood function. The inverse of the observed information matrix will represent the covariance matrix. The black dots represent $\tilde{\boldsymbol{h}}$ which are simulated from the multivariate normal distribution, centred at MLE $\hat{\boldsymbol{h}}$ and covariance matrix Σ calculated based on $\hat{\boldsymbol{h}}$,

$$\tilde{\boldsymbol{h}} \sim MVN\left(\hat{\boldsymbol{h}}, \Sigma\right), \tag{5.42}$$

where $\Sigma = J(\hat{\boldsymbol{h}})^{-1}$ and $L = \log l(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{p})$

$$\Sigma = \left(- \begin{pmatrix} \frac{d^2 L}{dh_0^2} & \frac{d}{dh_0}\left(\frac{dL}{dh_1}\right) \\ \frac{d}{dh_1}\left(\frac{dL}{dh_0}\right) & \frac{d^2 L}{dh_1^2} \end{pmatrix} \right)^{-1}. \tag{5.43}$$

Since the simulated $\tilde{\boldsymbol{h}}$ are located around the region of high likelihood, the information matrix calculation using the expressions that we have derived is correct. This empirical demonstration is important for confirming the accuracy of the information matrix, an important factor in determining the optimal design, and the implementation of the computer algorithm.

5.4.4 MLE search using Cross Entropy

Unlike the deterministic approach in Newton-Raphson method, which exploits the gradient of the (log) likelihood to determine good subsequent parameter values to search, Cross Entropy is a stochastic, and hence more volatile, method of optimising functions. Implementing Cross Entropy is straightforward even in high dimensional problems. The steps are adopted from De Boer et al. (2005) and applied to the same two dimensional problem where the number of samples of age i is $n_i = 250$ where $i = 0, 1$, and the true hazard rate is set at 0.1 and 0.05 respectively.

1. Since the MLE should be close to the true hazard rate, we let the initial hazard rate be $h_0^0 = 0.1$ and $h_1^0 = 0.05$ (for this simulation, these are the true param-

eters). The log likelihood for this initial set of hazard rates is calculated and stored as l^0.

2. Simulate n_{part} particles containing $\boldsymbol{h}^*$, the hazard rates, from independent normal distribution, centred at the current set of hazard rate $\boldsymbol{h^0} = (h_0^0, h_1^0)$ and variance $\sigma^2 = 0.01^2$,

$$h_i^* \quad \sim \quad N\left(h_i^0, \sigma^2\right) \forall i = 0, 1. \tag{5.44}$$

3. For each particle, calculate the log likelihood based on $\boldsymbol{h}^*$. Amongst all these points, the mean hazard rates of the top n_{top} particles with the highest log likelihood will be calculated and stored as $\boldsymbol{h^1}$.

4. If the maximum log likelihood, l^*, of all these n_{part} sets of particles is larger than the current log likelihood l^0, we will update the log likelihood for the next stage $l^1 = l^*$ and also update the current best hazard rate $\check{\boldsymbol{h}}^{\boldsymbol{1}}$ with the values of the particle that corresponds to the maximum log likelihood. Otherwise, $l^1 = l^0$ and $\check{\boldsymbol{h}}^{\boldsymbol{1}} = \boldsymbol{h^0}$.

5. The Cross Entropy procedure is terminated if the consecutive hazard values differ by less than ϵ,

$$\left|h_i^1 - h_i^0\right| \quad : \quad \begin{cases} < \epsilon \forall i & \text{Stop} \\ \geq \epsilon \forall i & \text{Continue.} \end{cases}$$

6. Repeat step 2 to 5 until the routine stops. The estimated MLE will be the particle with the maximum log likelihood, $\check{\boldsymbol{h}}$.

In round j of particles simulations, the normal distribution is centred at $\boldsymbol{h^j}$, the mean of the top n_{top} particles with the highest log likelihood, and not the current best hazard rate $\check{\boldsymbol{h}}^{\boldsymbol{j}}$. Because the best hazard rate $\check{\boldsymbol{h}}^{\boldsymbol{j}}$ will only be updated if the maximum log likelihood of the current round is more than the maximum log likelihood of all the previous rounds, $\check{\boldsymbol{h}}^{\boldsymbol{j}}$ may stay the same for several rounds, resulting in inefficiency in exploring the whole parameter space.

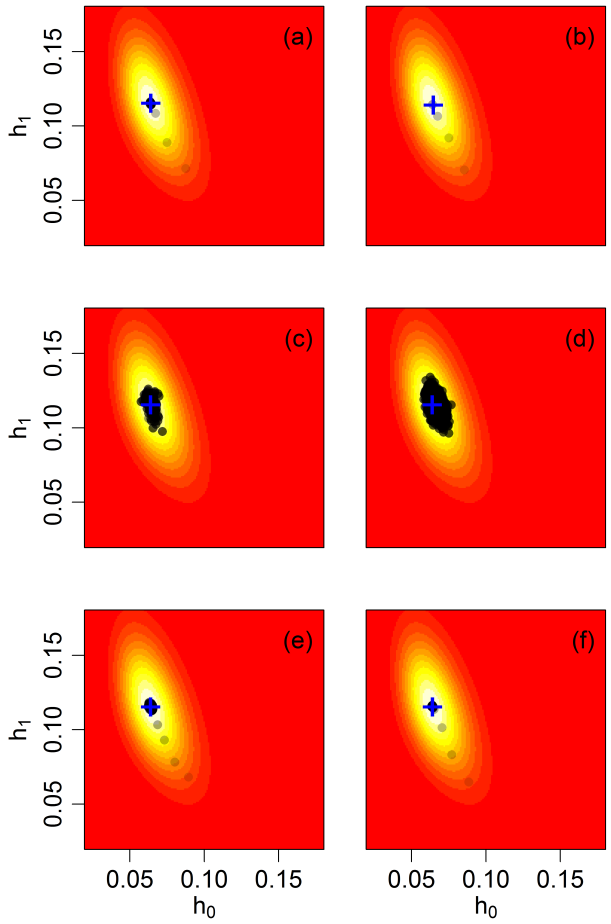

Figure 5.2: **Effects of different argument values in the Cross Entropy method.**
Image plots are as in figure 5.1, but the grey dots represents the stepwise movement of the mean of the top n_{top} particles, $\boldsymbol{h^j}$. Panel (a) is the result of $\left(\sigma = 0.01, n_{\text{part}} = 500, n_{\text{top}} = 10, \epsilon = 0.0001\right)$, (b) is the result of changing ϵ in (a) to 0.01, (c) is the result of changing σ in (a) to 0.1, (d) is the result of changing n_{top} in (c) to 25, (e) is the result of changing n_{part} in (a) to 100 and (f) is the result of changing n_{top} in (a) to 25.

The arguments $\left(\sigma, n_{\text{part}}, n_{\text{top}}, \epsilon\right)$ will also affect the efficiency of the procedure in finding the MLE. In panel (a) of figure 5.2, we tried more particles, $n_{\text{part}} = 500$, while keeping the other arguments small $\sigma = 0.01$, $n_{\text{top}} = 10$, $\epsilon = 0.0001$. This trial required 21 rounds of simulations. This is computer-time-consuming because the log likelihood will be calculated 500 times but only the top ten particles' information will be used. From the plot, the points around the MLE are very dark, symbolizing $\boldsymbol{h^j}$ was at the similar location for many rounds. Because the top ten particles are used, the variability between the means $\boldsymbol{h^j}$ at each round will be large, resulting in difficulty for the stopping condition $\left|\boldsymbol{h^j} - \boldsymbol{h^{j-1}}\right| < \epsilon$ to be fulfilled.

Then, we increased ϵ to 0.01 so that the routine can stop once it is close to the actual MLE. Results presented in panel (b) only required six rounds of simulations to attain the MLE. But the size of ϵ affects the precision of MLE. To achieve accurate MLE, it is still advisable to keep ϵ small.

In panel (c), we raised σ to 0.1 for the proposed values to be more dispersed. The routine is able to reach the region of high likelihood immediately. Because $n_{\text{top}} = 10$, the inconsistency in consecutive mean values resulted in 1992 simulation rounds before termination.

Panel (d) demonstrated the improvement in performance when n_{top} is increased to 25, while keeping the standard deviation $\sigma = 0.1$. As expected, the routine required many (308) more simulation rounds near the MLE than that in panel (c) to stop. Therefore, a larger n_{top} will draw more information from the simulated particles and be more efficient in getting to the MLE.

In panel (e), we reduced the number of simulations, n_{part}, to 100, while the other arguments stayed the same as that in panel (a). This reduced the number of log likelihood calculation at every round. Because fewer particles were simulated at every round, the chances of them having a high likelihood are reduced. This is justified by the smaller initial steps in panel (e) as compared to those in the earlier trials. Considering the small and cautious steps taken, this set of arguments required 221 steps before coming to a stop. It can be observed that the routine got near to the MLE within 5 steps but could not terminate because the condition was not met easily. This problem is likely to escalate when the dimension of the parameter space increases. So, we stay with the initial choice of $n_{\text{part}} = 500$.

In the last panel, we increased n_{top} from panel (a) to 25 to leverage on the information provided by the 500 particles at each round. When comparing panels (e) and (f), steps became larger due to the amount of information extracted from the proposed particles. Although it is faster, it still required 20 steps before the routine stopped. This performance in panel (f) is similar to that in panel (a), but, the steps in (f) appeared more regular. Thus, the argument in (f) was preferred since we wanted the search to have regular steps and terminate without wasting too many computation rounds, as well as not losing precision of the MLE result.

The details of the six trials may differ slightly due to the randomness in Cross En-

tropy algorithm, but we expect the qualitative interpretations of different arguments should still be valid.

The Cross Entropy method only works for uni-modal surfaces. If the log likelihood is multi-modal, the routine might stop at one of the local maxima which need not be the global maximum. Conservatively, then, Cross Entropy should be initialised at several starting points to verify that the same or similar $\check{\boldsymbol{h}}$ results.

5.4.5 MLE search using Monte Carlo Method

In Bayesian inference, Monte Carlo is a methodology which samples from the posterior distribution as discussed in Chapter 2. We can reduce the size of the parameter space if we simulate from the posterior distribution, rather than simulating over its whole support, in most of which the MLE is unlikely to be (for instance the same hazard example, if we only know $0 < h_i < 1$, we might sample uniformly over 0 and 1 even if the posterior is focused around 0.1). Although this is not equivalent to finding the MLE, if the posterior sample is large, the MLE can be estimated by the draw from the posterior with maximum (log) likelihood.

MCMC is a typical variant of Monte Carlo, but it is implausible to do MCMC on every simulated dataset to get the posterior sample, as for every dataset, there will be a different posterior distribution for the parameters, and the MCMC routine may need tuning to each. However, we could sample from a pseudo-'posterior', an appropriately selected Beta distribution (with support over [0,1]) where the shape parameters are formulated to focus on plausible hazard functions based on the simulated data.

Using the same two dimensional problem, where the number of samples of age i is $n_i = 250$ for $i = 0, 1$, the true hazard rate is set at 0.1 and 0.05 respectively. We assume a pseudo posterior distribution for the hazard rate for age i to be

$$h_i \quad \sim \quad Beta\left(\alpha_i, \beta_i\right), \tag{5.45}$$

where the shape parameters will relate its mean at the estimated hazard rate

$$\frac{\alpha_i}{\alpha_i + \beta_i} \quad = \quad \hat{h_i}. \tag{5.46}$$

Recall that hazard rate can be related to the probability of being identified as infected $p_i = g_{i-1}h_i + p_{i-1}$ where $g_i = \prod_{j=0}^{i-1}(1 - h_j)$. The estimated hazard rate can be

computed iteratively by

$$\hat{h}_i = \frac{\hat{p}_i - \hat{p}_{i-1}}{\hat{g}_{i-1}}, \tag{5.47}$$

where p_i can be estimated by the ratio of the number of positive samples to the total number of samples, $\frac{x_i}{n_i}$. Using the data of (x_i, n_i), the relation can be formulated for α_i

$$\alpha_i = \begin{cases} x_o & i = 0 \\ \frac{n_0 x_1}{n_1} - x_0 & i = 1 \\ \left(\prod_{j=0}^{i-1} n_j\right) \frac{x_i}{n_i} - \left(\prod_{j=0}^{i-2} n_j\right) x_{i-1} & i > 1 \end{cases} \tag{5.48}$$

and β_i

$$\beta_i = \begin{cases} n_0 - x_0 & i = 0 \\ \prod_{j=0}^{i-1} (n_j - x_j) - \alpha_i & i > 0. \end{cases} \tag{5.49}$$

Certain combinations of n_i and x_i may result in non-positive α_i and β_i, violating the condition of the shape parameters of a beta distribution. If this happens, we apply $\alpha_i = \beta_i = 1$ for the beta distribution to transform into a uniform distribution. This implies that if an informative posterior distribution does not exist, a vague distribution is used as a replacement.

We can simulate a large number of hazard rate particles from the pseudo posterior distributions and calculate the log likelihood value based on the corresponding dataset. The MLE could be identified as the particle with the largest log likelihood.

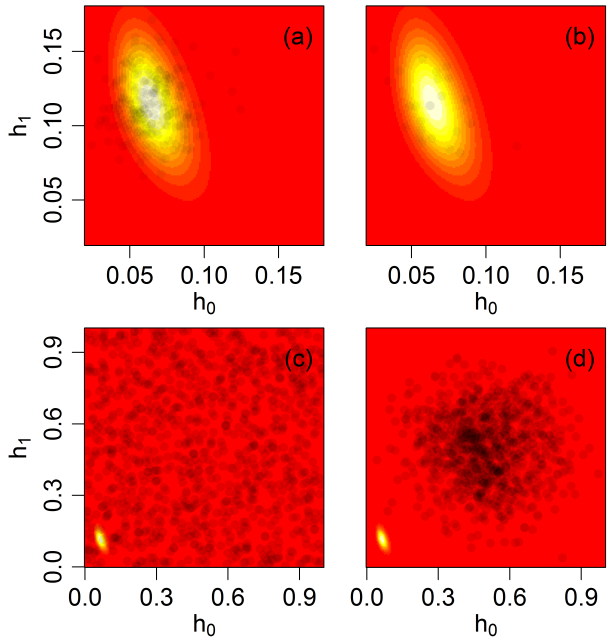

Figure 5.3: **Demonstration of how different proposal distributions affect the MLE search by Monte Carlo method.**
The image plot is as in figure 5.1. The simulated points are represented by the grey points. In panel (a), we sample 250 particles from the beta distribution, described in equations 5.48 and 5.49, where most particles are located at the region of high likelihood. In panel (b), we sample 25 particles from the same beta distribution, there were fewer points in the yellow region. To illustrate for other distributions in panel (c) and (d), we increase the range of the plot. In panel (c), we do a larger sample of 1000 particles from uninformative uniform distribution ranging from 0 to 1. In panel (d), we do the same large sample of 1 000 particles from a wrongly focused distribution: $N\left(0.5, 0.15^2\right)$.

In figure 5.3, we demonstrate the importance of proposing particles from a distribution that is akin to the likelihood. In the first panel (a), we only need 250 particles to obtain an estimated MLE that is consistent with the numerically calculated one. Comparing with panel (b) with 25 particles, having more particles will increase the chance of getting an accurate MLE. If more particles are preferred, computation for likelihood should be vectorised to reduce processing time.

The distribution in panel (c) can explore the whole parameter space but the chances of the simulated particles getting high likelihood are low. This will result in a poor MLE despite using 1 000 particles. The problem is analogous to that in panel (d) where the informative proposal distribution is targeted at the wrong part of the parameter space. For the same number of particles, the possibility of capturing a point with high likelihood is even lower.

In conclusion, with the consideration of computational time, we will search for

MLE using this Monte Carlo method. For higher dimensional problems, like in the original problem, we will need to increase the number of particles to be simulated to prevent a problem analogous to that faced in panel (b).

5.4.6 Design search using Grid Search

An experiment with three age groups and total sample size of $n = 500$ has $\frac{500(500+1)}{2} = 125\,250$ possible decisions. It will be very computational expensive to compute expected utilities for all decisions and we can expect the number of decisions to magnify with more age groups.

Atkinson et al. (2007) suggested searching the decision space by adding and subtracting 1 to the sample sizes of two randomly chosen age groups, but this mechanism took a long time to explore the decision space when tried on a simplistic three age group example. Instead, we use an iterative grid search, where initially a coarse grid is formed and the expected utility assessed on each point, and then the design space is restricted to high utility parts of the grid with increasingly fine grids overlaid.

Assuming the expected utility is uni-modal and smooth over the entire decision space, we form a grid of sample sizes for the first two age groups. The third age group will depend on the first two groups to satisfy the total sample size condition.

1. Setting the gap between each decision points to be 50 for the initial grid, the expected utilities of the decisions lying on the grid will be calculated.

2. With the expected utilities of the decisions on the grid, the decision with the highest expected utility is identified.

3. The next grid is re-centred at that decision with neighbouring points as limits and with the gap between new decision points reduced by $\frac{2}{3}$. This allows us to sequentially focus at the point where we believe the maximum expected utility will be.

4. Repeat step 2 and 3 until the gap is 1.

This is only feasible if the number of age group is small. For increased number of age groups, there will be too many decision points on the grid even if the gap is big. An optimal design might be missed in large grid gap when the mode is not close enough to the preselected decision points. It is also risky to use this method due to the assumption of uni-modal expected utility.

5.4.7 Design search using Cross Entropy

We use the Cross Entropy idea brought up in section 5.4.4 on the design search in 12 age groups, where the ith age group will have n_i samples.

1. To simulate n_{part} equally likely design points, a Dirichlet distribution is considered. The k dimensional Dirichlet distribution is multivariate and has support over the unit simplex, i.e. $[0,1]^k$ while ensuring that the simulated entries sum to 1. If the Dirichlet's concentration parameters, $\boldsymbol{\alpha}$, are the same, the mean for each dimension is $\frac{1}{k}$, with the magnitude of the α determining the spread. We multiply the total sample size, $n = 500$, to the simulated values from Dirichlet distribution. After rounding each entry to the nearest whole number, we alter the last entry such that they total to $n = 500$.

2. After calculating the expected utility for each design, we identify the point with the maximum expected utility, u^1, as $\check{\boldsymbol{n}}^{\mathbf{1}}$. The mean of the top n_{top} designs with the largest expected utility is stored as the current best design $\boldsymbol{n}^{\mathbf{1}}$.

3. We simulate the n_{part} new particles, $\boldsymbol{n}^*$, from the algorithm in step 1 using Dirichlet distribution centred at the current best design $\boldsymbol{n}^{\mathbf{1}}$,

$$\frac{\boldsymbol{n}^*}{n} \quad \sim \quad Dir(\boldsymbol{\alpha}), \tag{5.50}$$

where $\boldsymbol{\alpha}$ is the concentration parameters which relates the centre location of the distribution to the current best design by

$$\boldsymbol{\alpha} \quad = \quad \theta \times \boldsymbol{n}^{\mathbf{1}}, \tag{5.51}$$

and θ controls the dispersion of particles from $\boldsymbol{n}^{\mathbf{1}}$. The bigger the value of θ, the more concentrated the simulated $\boldsymbol{n}^*$ will be from $\boldsymbol{n}^{\mathbf{1}}$.

4. We calculate the expected utility for all new designs. If the maximum expected utility of these designs is larger than the current maximum expected utility u^1, it will be updated as u^2 and its corresponding design as $\check{\boldsymbol{n}}^{\boldsymbol{2}}$, otherwise, $u^2 = u^1$ and $\check{\boldsymbol{n}}^{\boldsymbol{2}} = \check{\boldsymbol{n}}^{\boldsymbol{1}}$. The current best design, $\boldsymbol{n}^{\boldsymbol{2}}$, is updated with the mean of the top n_{top} designs.

5. The routine stops if the absolute distances between subsequent best design points for all age groups are less than ϵ,

$$\left|n_i^2 - n_i^1\right| \quad : \quad \begin{cases} < \epsilon \forall i & \text{Stop} \\ \geq \epsilon \forall i & \text{Continue.} \end{cases}$$

6. Repeat step 3 to 5 until the routine stops. The optimal design experiment is the design with the maximum expected utility, $\check{\boldsymbol{n}}$.

Unfortunately this methodology arrived at completely different design points for three different seeds and was therefore not considered further.

5.4.8 Changes to Optimization Criterion

Currently, the optimization criterion is to maximise the estimated expected utility, which is the determinant of the information matrix under the decision of D and the MLE $\hat{\boldsymbol{h}}$. However, utilities calculated based on the MLE of the hazard rate demonstrated a high volatility due to the dimensionality of the parameter vector $\boldsymbol{h}$. In the optimal design search, the 'best' design is the one which corresponds to the maximum expected utility. If the expected utility is unstable, the optimal design will be difficult to estimate accurately. We therefore sought a less unstable objective function.

In changing the optimization criterion, we aim to reduce the irregularity in the estimation of the expected utility for each decision, D. Instead of depending on the larger number of parameters in $\boldsymbol{h}$, we fit a parametric survival regression model to the simulated data, characterised by only two parameters. This can readily be effected using built-in functions in R using the `survival` package (Therneau, 2013). This reduction in the number of parameters will alleviate the problem of volatility in the utility calculation.

Fitting survival regression models in R requires creating a 'survival object' which indicates the type of censoring for each individual, the start and end of event time. If an individual is seropositive at age i, infection has happened between birth and age i when the serum was collected. We categorise this event as left-censored. The likelihood for such individuals will be accounted by the density of the lower tail, $P(T \leq i)$. Otherwise, it is right-censored if the individual is seronegative at age i, as infection has not taken place. In such cases, the upper tail, $P(T > t)$, will be required since the event might take place after the specimen was collected.

In the survival regression model, we do not regress on any covariates and the survival times are assumed to follow a Weibull distribution, whose support is over the non-negative range, and whose distribution is characterised by the shape parameter, κ, and scale parameter, λ, both positive. These parameter estimates can be derived from the scale estimate in survival regression model, b, and intercept, a,

$$b = \frac{1}{\kappa} \tag{5.52}$$

$$a = \log \lambda. \tag{5.53}$$

The Weibull distribution parameters are less uncertain if the chosen design experiment can provide much information. If the parameters are dispersed, the determinant of the variance-covariance matrix will be large, symbolising a large parameter area. Thus, we change the optimization criterion to be based on a utility that is the reciprocal of the absolute determinant of the variance-covariance matrix. Maximising the utility value will correspond to the best design point that leads to the least dispersed Weibull distribution parameters.

The problem of randomness in the survival regression model could be reduced by using the same idea of propensity and prevalence to simulate the data. In this survival regression model, the Weibull distribution parameters and the corresponding variance-covariance matrix will be estimated deterministically and classically for each set of simulated data.

5.4.9 Design search using Monte Carlo Method

Since both the grid search and Cross Entropy were not feasible or reliable from one run to the next in our high dimensional (and possibly multi-modal space), we used the Monte Carlo method to search the whole design space for the best design.

This algorithm took ca. 20 days to explore ca. 40 000 design points on a standard desktop computer with four processors, and yielded in three independent runs 'optimal' designs that were rather consistent with each other. Note that to allocate 500 individuals into 12 groups, we need to choose 11 partitions. These partitions are considered as individuals to be added to the total 500 individuals. Hence, it was not feasible to explore all $\binom{500+11}{11} = 1.4 \times 10^{22}$ designs.

The algorithm is as follows:

1. A large number of design points are simulated for exploration. The same description can be found in section 5.4.7 step 1.

2. For each design, compare a set of propensity ϕ and the prevalence p for all the n individuals. If $\phi_l < p_l$, the lth individual is seropositive. Set up the survival object for these simulated data by creating appropriate censoring and event times according to the test results.

3. Run the survival regression model for the survival times and compute the reciprocal of the absolute determinant of the variance-covariance matrix as the utility for this simulated data.

4. Repeat step 2 and 3 for every set of propensity and prevalence and represent the expected utility of this design point with the average of the utilities.

5. Repeat step 2 to 4 for each design points simulated in step 1. The design point with the maximum expected utility is the optimal design based on this criterion.

This time-consuming approach could be improved by running the routine in parallel on a server.

5.5 Result and Discussion

The optimal design experiment is an interesting mixture of methods Bayesian

and Classical. In the first step, we use hierarchical modelling to synthesize data from historical studies; in the second step, we used Bayesian decision making with a Frequentist objective function to seek for the optimal design.

Results from the hierarchical model are presented in figure 5.4. Within country estimates are provided in the first 8 panels, with the hierarchical model fits in blue and empirical confidence intervals (based on the usual $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ formula) in red. The goodness of fit can be readily observed. The hierarchical model also allows the posterior distribution of the true prevalence for a subsequent study, such as the one being designed, in a similar country using the hyper-parameters. This may be found in the green panel in figure 5.4. Note that although we are concerned about the hazard rate from age 1 to 12, i.e. the pre- and primary school years, not all datasets in our literature review have the necessary length. Hierarchical modelling has been shown to draw information from available data to provide information for gaps such as this, or other potential datasets, based on the assumption that the populations are similar (i.e. exchangeable) across these countries and that seropositivity has not changed much across the time period 1994–2011.

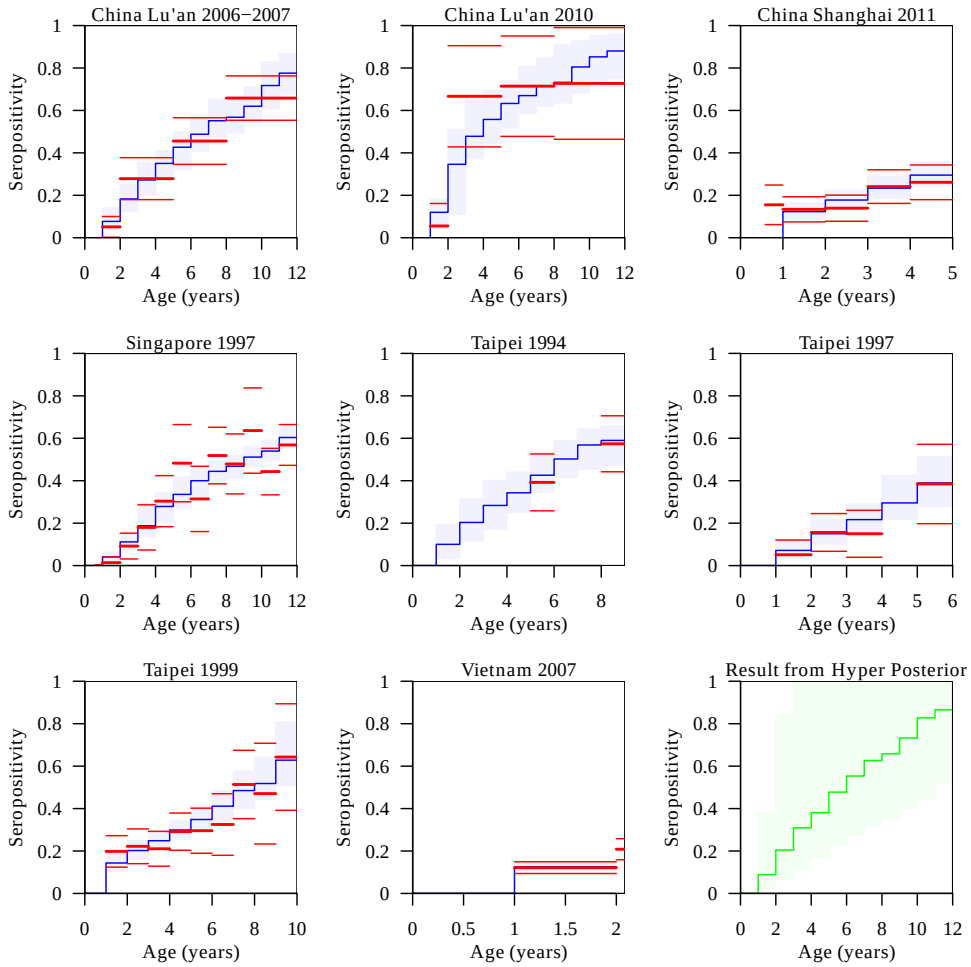

Figure 5.4: **Seropositivity of the eight datasets, as well as the projection using the hyper posterior from the hierarchical model.**
The light blue shades are the projection of the seropositivity using the posterior samples of the hazard rates h_i for age $0 < i < 12$ where the median of the projection is symbolised by the blue lines. The thick red lines represent the empirical mean seropositivity calculated from the dataset and the thinner red lines represent the confidence interval. The last plot of green shades shows the seropositivity for any randomly chosen country, simulated from the hyper-posterior sample, where the median is also illustrated by the green line.

The simulated seropositivity is narrow for age below 2 because of the abundance of information below age 2. If less information is collected for certain ages, the prediction will be vague. There were only three datasets with information up to age 12. Almost 90% of the older children from China Lu'an in 2010 were projected to be seropositive by age 12, in contrast to the estimates in the same location at 2006–2007 and Singapore in 1997 which were about 80% and 60% respectively. Due to this variability, the prediction from the hierarchical model using the hyper posterior samples is less precise, as shown by the wide green prediction for older children in the last panel.

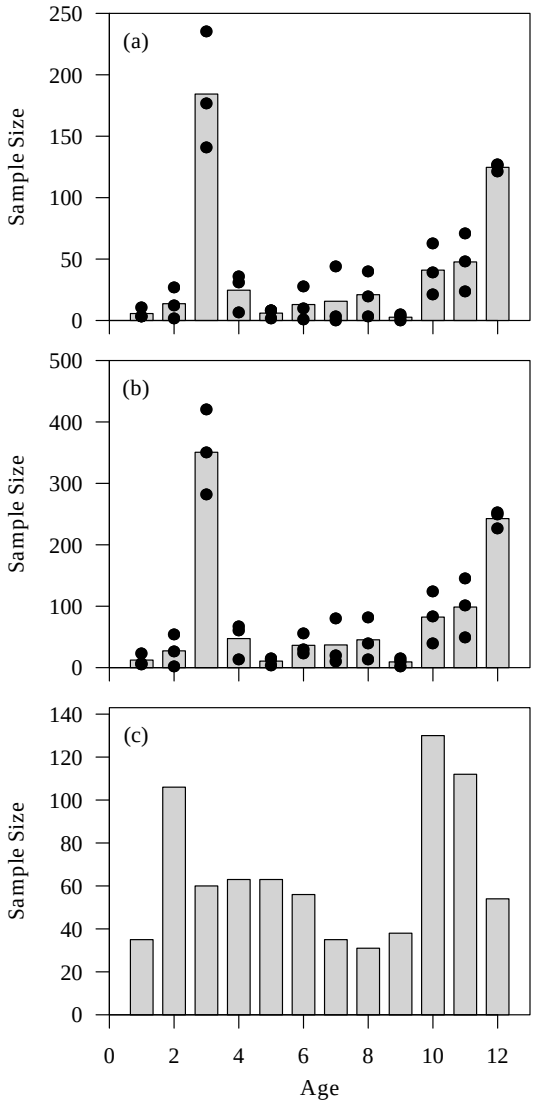

Figure 5.5: **Optimal sample sizes for each age.**
The black dots represent the optimal designs from each of the three runs and the grey bars represent the mean from the three optimal designs. Panel (a) and (b) are the results when maximum sample size is set at 500 and 1000 respectively. Panel (c) are the sample sizes used in Singapore from 2008 to 2010.

Both sample sizes considered (500 and 1 000) are consistent and have a 'U' shaped optimal design with approximately the same proportions in each age group. In the 2008–2010 study by Ang et al (2011), the design had a minimum of 340 sample size for each age group (1–6, 7–12, and 13–17 years), where we dropped the information above age 12 for proper comparison with our optimal design. Fortuitously, the Ang et al. (2011) study had a large number of samples at age 2 (close to the spike age 3 in the optimal design) and again around 10–12 (close to the second spike in the optimal design), and as a result, we expect the dataset from Singapore in 2008-2010 to be very informative.

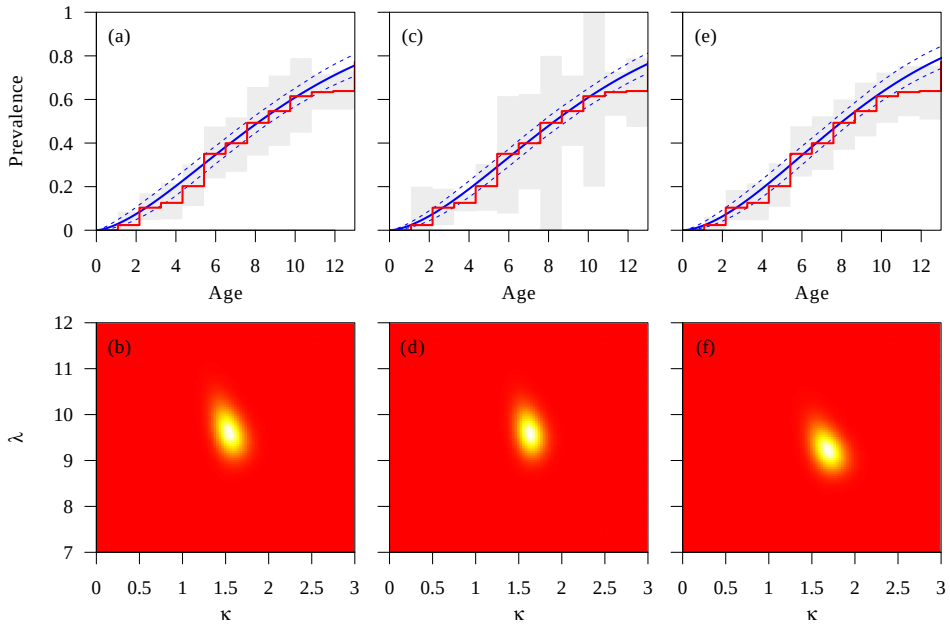

Figure 5.6: **Comparison of the performance of 3 different experimental designs.**
The result of sample sizes used by Singapore in 2008–2010 is presented in panel (a) and (b); one of our optimal designs scaled up to the same total as that in Singapore (729 sera samples) in panel (c) and (d), and equal sample size for all ages in panel (e) and (f). Assuming the underlying prevalence is coming from our hierarchical model, the first row shows the plot of prevalence against age. The grey shades represent the prediction interval of the prevalence for each age. The red line is the underlying prevalence that is simulated from our hierarchical model. The blue lines are the result of the survival regression where the solid lines are computed from the parameters estimated from the model and the dotted blue lines are the 95% confidence interval of the computed prevalence using the Weibull parameters simulated from multivariate normal distribution with mean and variance-covariance from the estimates in the regression model. The second row is showing heat map of the Weibull parameters κ and λ. Yellow represents the point where likelihood is the highest and red when it is the lowest.

The grey shades in figure 5.6, panel (a), (c) and (e), are the 95% prediction interval of prevalence computed by dividing the simulated number of infected based on the underlying prevalence and the chosen sample sizes using a binomial distribution, by the sample sizes.

As expected, the grey shades for the sample size used by Singapore in 2008–2010 show a nice fit in panel (a). The wider grey shades for optimal design in panel (b) is due to the poorer prevalence estimation for older children. Due to an equal sample size for all ages used in panel (e), the size of the prediction intervals were consistent across ages.

The dotted blue lines in figure 5.6 are prevalence estimates computed using the

result of survival regression. They are narrowest for our optimal design in panel (c), followed by Singapore's design in panel (a) and the widest in equal sample size in panel (e).

The heat map in figure 5.6 is based on the likelihood that is computed from the densities of all the survival times from the samples based on a Weibull distribution.

Panel (d) with the optimal design has the narrowest spread of the Weibull parameters, followed by panel (b) from Singapore 2008–2010 samples and the least precise parameter estimates from panel (f) of equal sample size.

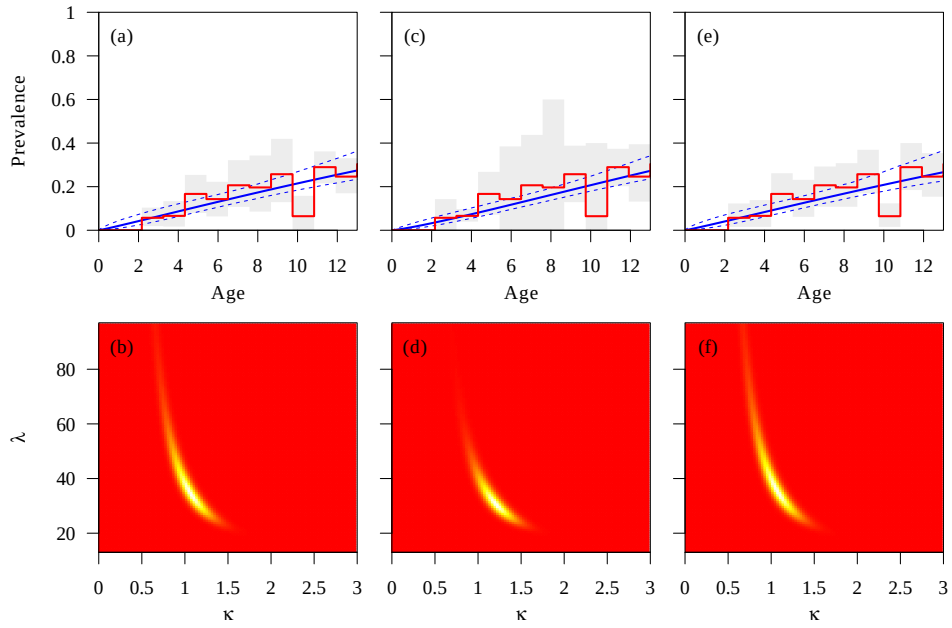

Figure 5.7: **Comparison of the performance of 3 different experimental designs using a different underlying prevalence.**
This is done by assuming the underlying prevalence comes from the Singapore 2008–2010 dataset. The features in this figure is the same as that in figure 5.6.

Although the grey shades are still wider for our optimal design in panel (c) as compared to the design used by Singapore in 2008–2010 in panel (a) and the equal sample size design used in panel (e), the Weibull parameters estimates for the survival time is still the sharpest for our optimal design as seen in panel (d). Our optimal design is still the one that provides the best information about the parameters that can inform about the survival time.

The ability of having the focused Weibull parameter estimates supports the criterion that was used to achieve this optimal design. Recall that our optimal designs were identified as the design with the maximum expected utility, where utility is

the reciprocal of the absolute determinant of the variance-covariance matrix of the Weibull parameter estimates derived from the survival regression model.

5.6 Conclusion

In this example, different seropositivity estimates in different Asian populations will lead to a wide seropositivity projection for any future seroepidemological studies. Because the optimal design is solely based on the efficiency in the seropositivity projection, the performance of the optimal design may deteriorate if the past studies cannot properly represent the population that the optimal design experiment is constructed for. To boost the performance of our optimal design might require more datasets with information on older children from more directly comparable studies be fitted in the hierarchical model. This will narrow the seropositivity estimates for the prevalence in the country being investigated which will lead to a more appropriately tailored design.

One discovery, not anticipated when we started this chapter, was the importance of the specific choice of optimality criterion. Clearly, optimal designs under different criteria will differ. We anticipated that the results using an objective function based on a parametric survival analysis would provide good estimates even under a more complex non-parametric model, but while our optimal designs did minimise the width of confidence/credible intervals for the survival function conditional on a Weibull model, the prevalence estimates using the non-parametric approach were sometimes broader under the optimal design than under an equal sampling approach. Thus, it is important to understand the motivation for the new serology experiment and choose the appropriate optimal criterion accordingly.

Chapter 6

Conclusion and Future Work

6.1 Summary

We have demonstrated that hierarchical models can boost the accuracy of parameters estimates and prediction in the area of infectious diseases. We applied the method to clinical data from patients with Dengue and Chikungunya, epidemiological data from multiple countries on pandemic influenza A H1N1 and seroepidemiological data on EV71 to design optimal future studies. Hierarchical models were fitted to the available data flexibly and the algorithms have been illustrated in this thesis.

In chapter 3, we encountered the separation problem in data from patients with Dengue or Chikungunya when first presented at Tan Tock Seng Hospital, Singapore's main referral centre for infectious diseases. In a previous publication, we resolved this problem using Firth's penalised likelihood logistic regression method (Firth, 1993). Without using some form of penalty, whether classical or Bayesian (via an informative or semi-informative prior), the significantly *predictive* variable of Platelet counts would have been excluded before the regression model was developed as being not statistically significantly different from 0. A hierarchical model was established for characterizing the time course of laboratory and clinical measurements of Dengue and Chikungunya patients. The precise analysis of the trend of disease course can facilitate the diagnosis and treatment of the patients of these two diseases with otherwise similar symptoms.

Subsequently, a hierarchical compartmental model was developed for data from the 2009 H1N1 pandemic for a basket of countries that would have allowed early and accurate severity estimates. Several factors of this worldwide pandemic outbreak

were encompassed in the model formation. Different forms of data type collected by independent research groups or government agencies in different territories were used to synthesize evidence for the unobservable components in the SIR model. This involved extending the methodology used when the MCMC algorithm would not lead to convergence. The technique of importance sampling solved this problem, providing necessary parameter information for updating the severity estimates of the pandemic in real time.

In chapter 5, we explored how hierarchical models can be used to improve the accuracy of the posterior estimates for hazard rates of EV71, using previously collected serological data from several Asian countries. The search for an optimal experiment design for any Asian country was done based on the hierarchical model, assuming exchangeability between past and future epidemic conditions. The idea is to sample from the appropriate population group to get the most desirable age effect estimates in a serological study, rather than using convenience sampling which may be wasting resources by over-sampling age groups that are no longer at risk of infection.

6.2 Future Work

Here we discuss possible extensions of our work.

In the hierarchical modelling for trends in observation for Dengue and Chikungunya patients, future work can involve integration of the hierarchical model with real time data. Importance sampling can be done by using the current weighted samples as a hyperprior distribution and new weights can be calculated based on new and incoming patients' data. This allows the project to be ongoing and the weighted samples can be constantly refined. However, we should manually exclude patients with abnormal observations. These type of observations might give higher weights to inappropriate parameter values, leading to distortion of the parameters samples.

Hierarchical models could be integrated with risk calculators which can be developed to predict risk of complications in dengue (risk of developing Dengue Hemorrhagic Fever, and risk of requiring to the Intensive Care Unit or of death). This could be done if data collection could be improved by being more organised. Patients' observations were recorded daily but at irregular times. The model could be more precise by replacing the discrete time model with a continuous time model to

leverage on the information of the measurement times.

As mentioned in chapter 4, the data collection from each country was based on different criteria. If a surveillance network can be set up for each country to submit the data counts based on a standardised criteria conveniently, real-time analysis can be done and data can be fitted better to the SIR trajectory. For analysis in real-time, a different approach should be done to replace the sequential importance sampling which will take a long computational time.

The earlier importance sampling routine was inefficient and not able to analyse the pandemic in real-time. Ong et al. (2010) have demonstrated in the context of Singapore how real-time analysis can be done by sequential importance sampling for a stochastic model. Parameters' particles could be simulated independently for each country and weight could be calculated daily based on the likelihood and proposal density, presuming that data from every country was entered into the system daily. For the period with no data input, there will be no addition to the weights. The cumulated weights can be representative for each particle and the credible interval for each parameter at each day can be achieved by resampling the particles based on the cumulative weights. Since each country can be analysed in parallel at this stage, computation time could be reduced. For each country, the same number of set of particles is resampled and a set of hyper-parameters are simulated to pair up with the particles from all countries. Weights could be calculated daily based on the hierarchical model and the sets of parameters and hyper-parameters can be resampled based on this weight. The result of this can be used for trajectory projection and severity estimation and we anticipate that this development into a two stage process will greatly reduce the computation time and the ability to provide daily credible interval and estimations will be more beneficial than having a monthly prediction in a real pandemic outbreak.

If such a network were established, age specific estimates of attack rates and severity indices could be established if the age of patients were also collected automatically. This would require extension to an age structured model in which case estimates of mixing between different age groups would be useful (Mossong et al., 2008). At the moment, these are lacking in most countries especially low to middle income ones.

In the last analysis for EV71, the countries that we have chosen to fit the hierarchical model have relatively different prevalence for older children. This resulted in a wide prediction interval for prevalence in any future serological study. This might be addressed by introducing additional structure in the hierarchical model where cities have random errors about their country's mean, while all countries still remain to be modelled by the hierarchical model. Alternatively, hazard rates can be regressed based on each country's risk factors before being combined with the hierarchical model. The risk factors which account for the difference in the prevalence for older kids can be explored/identified based on the school attendance (i.e. age of attending pre-school). This would allow better designs in a country with the risk factors being accounted for.

Other than the above applications of hierarchical model to the context of infectious diseases, other works could involve assessing the differences in outbreaks of Hand, Foot, and Mouth Disease in preschools across Singapore or other small, closed populations like army camps. Each of these small, closed populations could be governed by parameters which would be modelled hierarchically. Dengue cluster outbreak with more than 2 cases within 2 weeks of onset and 150m radius (usually from patients' homes) might also be modelled hierarchically if the data, collected by the Singapore Government, could be made available for research purposes. The analysis of dengue infection forms the basis for vector control operations in different regions of Singapore, allowing prioritisation of mosquito control, especially during a large outbreak like that in 2013. By more sophisticated application of statistical methodologies, the impact of diseases such as dengue, influenza and hand, foot and mouth disease could be mitigated.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Leukocyte functions and percentage breakdown.* New York: Garland Science.

Ang, L., Phoon, M.-C., Wu, Y., Cutter, J., James, L., & Chow, V. T. (2011). The changing seroepidemiology of enterovirus 71 infection among children and adolescents in singapore. *BMC Infectious Diseases*, *11*, 270.

Atkinson, A. (2001). *Optimum design 2000* (P. Pardalos, Ed.). Kluwer Academic Publishers.

Atkinson, A., Donev, A., & Tobias, R. (2007). *Optimal experimental designs, with SAS.* Oxford Statistical Science Series.

Atkinson, A., & Kendall, E. (2008). *An introduction to numerical analysis.* John Wiley & Sons.

Babcock, H. M., Merz, L. R., & Fraser, V. J. (2006). Is influenza an influenza-like illness? clinical presentation of influenza in hospitalized patients. *Infection control and hospital epidemiology*, *27*(3), 266–270.

Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., et al. (2009). Seasonal transmission potential and activity peaks of the new influenza a (H1N1): a monte carlo likelihood analysis based on human mobility. *BMC medicine*, *7*(1), 45.

Balter, S., Gupta, L. S., Lim, S., Fu, J., & Perlman, S. E. (2010). Pandemic (H1N1) 2009 surveillance for severe illness, new york, new york, usa, april - july 2009. *Emerging Infectious Diseases*, *16*, 1259-1264.

Barbour, A. D. (1974). On a functional central limit theorem for markov population processes. *Advances in Applied Probability*, *6*, 21–39.

Barrau, M., Larrieu, S., Cassadou, S., Chappert, J.-L., Dussart, P., Najioullah, F., et al. (2012). Hospitalized cases of influenza A(H1N1)pdm09 in the french territories of the americas, july 2009-march 2010. *Rev Panam Salud Publication*, *32*, 124-130.

Barroso, L., Treanor, J., Gubareva, L., & Hayden, F. G. (2005). Efficacy and tolerability of the oral neuraminidase inhibitor peramivir in experimental human influenza: randomized, controlled trials for prophylaxis and treatment. *Antiviral Therapy*, *10*, 901-910.

Belshe, R. B. (2010). The need for quadrivalent vaccine against seasonal influenza. *Vaccine*, *28*, D45–D53.

Berger, M. P., & Wong, W.-K. (2005). *Applied optimal designs.* John Wiley & Sons.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, *7*, 686-690.

Blum, M. G., & Tran, V. C. (2010). HIV with contact tracing: a case study in approximate bayesian computation. *Biostatistics*, *11*(4), 644–660.

Bolstad, W. M. (2004). *Introduction to bayesian statistics.* John Wiley and Sons, Inc.

Borja-Aburto, V. H., Chowell, G., Viboud, C., Simonsen, L., Miller, M. A., Grajales-Muñiz, C., et al. (2012). Epidemiological characterization of a fourth wave of pandemic A/H1N1 influenza in mexico, winter 2011 - 2012: Age shift and severity. *Archives of Medical Research*, *43*, 563-570.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, *5*(7), e151.

Carrat, F., Vergu, E., Ferguson, N. M., Lemaitre, M., Cauchemez, S., Leach, S., et al. (2008). Time lines of infection and disease in human influenza: A review of volunteer challenge studies. *American Journal of Epidemiology*, *167*, 775-785.

Cauchemez, S., & Ferguson, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london. *Journal of the Royal Society Interface*, *5*, 885-897.

Cauchemez, S., Temime, L., Guillemot, D., Varon, E., Valleron, A.-J., Thomas, G., et al. (2006). Investigating heterogeneity in pneumococcal transmission: A bayesian mcmc approach applied to a follow-up of schools. *Journal of the American Statistical Association*, *101*, 946-958.

CDC. (2009). *Seasonal influenza (flu) - flu activity & surveillance.* Interactive Website. 1600 Clifton Rd. Atlanta, GA 30333, USA. Retrieved from `http://www.cdc.gov/flu/weekly/fluactivitysurv.htm`

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304.

Chang, M., Southard, C., & Sullivan, M. (2010, April). *Learning from the 2009*

H1N1 influenza pandemic. Website. Retrieved from `http://www.rms.com/publications/H1N1_2009_SpecialReport.pdf`

Chao, D. Y., Cheng, K. F., Li, T. C., Wu, T. N., Chen, C. Y., Tsai, C. A., et al. (2011). Serological evidence of subclinical transmission of the 2009 pandemic H1N1 influenza virus outside of mexico. *PLoS ONE*, *6*, e14555.

Chen, M. I. C., Lee, V. J. M., Lim, W.-Y., Barr, I. G., Lin, R. T. P., Koh, G. C. H., et al. (2010). 2009 influenza A(H1N1) seroconversion rates and risk factors among distinct adult cohorts in singapore. *Journal of the American Medical Association*, *303*, 1383-1391.

Chowell, G., Zuno, S. E., Viboud, C., Simonsen, L., Tamerius, J., Miller, M. A., et al. (2011). Characterizing the epidemiology of the 2009 influenza A/H1N1 pandemic in mexico. *PLoS Medicine*, *8*, e1000436.

Chua, G. (2009, July). *5 honoured for excellence in medical work.* Newspaper. Retrieved from `http://www.nuh.com.sg/wbn/slot/u1753/Patients%20and%20Visitors/Media%20Articles/Jul%2009/23rd%20ST%205%20Honoured.pdf`

Cohen, J., & Cohen, P. (1984). *Applied multiple regression/correlation analysis for the behavioral sciences.* Psychology Press.

Congdon, P. (2001). *Bayesian statistical modelling.* John Wiley and Sons, Ltd.

Cook, A. R., Gibson, G. J., & Gilligan, C. A. (2008). Optimal observation times in experimental epidemic processes. *Biometrics*, *64*(3), 860–868.

Cook, A. R., Gibson, G. J., Gottwald, T. R., & Gilligan, C. A. (2008). Constructing the effect of alternative intervention strategies on historic epidemics. *Journal of the Royal Society Interface*, *5*, 1203-1213.

Cook, A. R., Otten, W., Marion, G., Gibson, G. J., & Gilligan, C. A. (2007). Estimation of multiple transmission rates for epidemics in heterogeneous populations. *Proceedings of the National Academy of Sciences (PNAS)*, *104, no.51*, 20392-20397.

Cooper, B. S., Medley, G. F., Bradley, S. J., & Scott, G. M. (2008). An augmented data method for the analysis of nosocomial infection data. *American Journal of Epidemiology*, *168*, 548-557.

Cutter, J. L., Ang, L. W., Lai, F. Y. L., Subramony, H., & Ma, . J. L., S. (2010).

Outbreak of pandemic influenza A (H1N1-2009) in singapore, may to september 2009. *Annals Academy of Medicine*, *39*, 273-282.

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P.-Y., et al. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet Infectious Diseases*, *12*, 687-695.

Debing, Y., Jochmans, D., & Neyts, J. (2013). Intervention strategies for emerging viruses: use of antivirals. *Current opinion in virology*, *3*, 217-224.

De Boer, P.-T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of operations research*, *134*(1), 19–67.

Domínguez-Cherit, G., Lapinsky, S. E., Macias, A. E., Pinto, R., Espinosa-Perez, L., Torre, A. de la, et al. (2009). Critically ill patients with 2009 influenza A (H1N1) in mexico. *Jama*, *302*(17), 1880–1887.

Dudareva, S., Schweiger, B., Thamm, M., Höhle, M., Stark, K., Krause, G., et al. (2011). Prevalence of antibodies to 2009 pandemic influenza A (H1N1) virus in german adult population in pre- and post-pandemic period. *PLoS ONE*, *6*, e21340.

ECDC. (2014, March). *European centre for disease prevention and control (ecdc) @ONLINE.* Retrieved from `http://www.ecdc.europa.eu/en/Pages/home.aspx`

Efron, B., & Htnkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, *65*, 457-487.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27-38.

Fox, J. L. (2009). Testing for H1N1 flu during surge stresses many U.S. clinical labs. *American Society for Microbiology*, *4*, 405-410.

Fritz, R. S., Hayden, F. G., Calfee, D. P., Cass, L. M. R., Peng, A. W., Alvord, W. G., et al. (1999). Nasal cytokine and chemokine responses in experimental influenza a virus infection: Results of a placebo-controlled trial of intravenous zanamivir treatment. *The Journal of Infectious Diseases*, *180*, 586-593.

Fuhrman, C., Bonmarin, I., Bitar, D., Cardoso, T., Duport, N., Herida, M., et al.

(2011). Adult intensive-care patients with 2009 pandemic influenza A (H1N1) infection. *Epidemiology and Infection*, *139*, 1202-1209.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Chapman and Hall.

Gibson, G. J., & Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using markov chain methods. *Mathematical Medicine and Biology*, *15*(1), 19–40.

Gilks, W. R., Spiegelhalter, D. J., & Richardson, S. (1996). *Markov chain monte carlo in practice*. Chapman and Hall.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, *81*(25), 2340–2361.

Goodwin, R., Haque, S., Neto, F., & Myers, L. B. (2009). Initial psychological responses to influenza A, H1N1. *BMC Infectious Diseases*, *9*(1), 166.

Government, A. (2011). *Review of australia's health sector response to pandemic (H1N1) 2009. lessons identified* (Tech. Rep.). Australian Government Department of Health and Ageing.

Halder, N., Kelso, J. K., & Milne, G. J. (2010). Analysis of the effectiveness of interventions used during the 2009 A/H1N1 influenza pandemic. *BMC Public Health*, *10*, 168.

Hammond, B. J., & Tyrrell, D. A. J. (1971). A mathematical model of common-cold epidemics on tristan da cunha. *Journal of Hygiene*, *69*, 423-433.

Hayden, F. G., Fritz, R. S., Lobo, M. C., Alvord, W. G., Strober, W., , et al. (1998). Local and systemic cytokine responses during experimental human influenza a virus infection relation to symptom formation and host defense. *The Journal of Clinical Investigation*, *101*, 643-649.

Hayden, F. G., Treanor, J. J., Betts, R. F., Lobo, M., Esinhart, J. D., & Hussey, E. K. (1996). Safety and efficacy of the neuraminidase inhibitor GG167 in experimental human influenza. *The Journal of the American Medical Association*, *275*, 295-299.

Hayden, F. G., Tunkel, A. R., Treanor, J. J., Betts, R. F., Allerheiligen, S., & Harris, J. (1994). Oral LY217896 for prevention of experimental influenza A virus infection and illness in humans. *Antimicrobial Agents and Chemotherapy*, *38*,

1178-1181.

Heffernan, J., Smith, R., & Wahl, L. (2005). Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, *2*, 281-293.

Heinze, G., & Ploner, M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine*, *71*, 181-187.

Heinze, G., Ploner, M., Dunkler, D., & Southworth, H. (2013). logistf: Firth's bias reduced logistic regression [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=logistf` (R package version 1.21)

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*, 2409-2419.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, *42*(4), 599–653.

Hollister, S., Maddox, R., & Taboas, J. (2002). Optimal design and fabrication of scaffolds to mimic tissue properties and satisfy biological constraints. *Biomaterials*, *23*, 4095-4103.

Ikonen, N., Haanpaa, M., Ronkko, E., Lyytikainen, O., Kuusi, M., Ruutu, P., et al. (2010). Genetic diversity of the 2009 pandemic influenza A(H1N1) viruses in finland. *PLoS ONE*, *5*, e13329.

ISD, I. S. D. (2009, June). *Strategy for contact tracing of flight passengers adjusted.* News Article. Retrieved from `http://www.info.gov.hk/gia/general/200906/12/P200906120350.htm`

Keeling, M. J., & Ross, J. V. (2008). On methods for studying stochastic disease dynamics. *Journal of The Royal Society Interface*, *5*(19), 171–181.

Kerkhove, M. D., Hirve, S., Koukounari, A., & Mounts, A. W. (2013). Estimating age-specific cumulative incidence for the 2009 influenza pandemic: a meta-analysis of A (H1N1) pdm09 serological studies from 19 countries. *Influenza and other respiratory viruses*, *7*(5), 872–886.

King, J. W., & Markanday, A. (2003). Ebola virus. *EMedicine. Medscape.*

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*, *4*, 865-877.

Kubiak, R. J., & McLean, A. R. (2012). Why was the 2009 influenza pandemic in england so small? *PLoS ONE*, *7*, e30223.

Lau, L. L., Nishiura, H., Kelly, H., Ip, D. K., Leung, G. M., & Cowling, B. J. (2012). Household transmission of 2009 pandemic influenza A (H1N1): a systematic review and meta-analysis. *Epidemiology (Cambridge, Mass.)*, *23*(4), 531.

Lauritzen, S. (2008, Jan). *Newton-raphson iteration and the method of scoring.* Online Lecture Notes. Retrieved from `http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/scoring.pdf`

Lee, E. H., Wu, C., Lee, E. U., Stoute, A., Hanson, H., Cook, H. A., et al. (2010). Fatalities associated with the 2009 H1N1 influenza A virus in new york city. *Emerging Infections*, *50*, 1498-1504.

Lee, J., & McKibbin, W. J. (2004). Estimating the global economic costs of SARS. In *Learning from sars: Preparing for the next disease outbreak–workshop summary* (p. 92).

Lee, P. (1997). *Bayesian statistics - an introduction.* Arnold.

Lee, V., Tan, C. H., Yap, J., Cook, A. R., Ting, P.-J., Loh, J.-P., et al. (2011). Effectiveness of pandemic h1n1-2009 vaccination in reducing laboratory confirmed influenza infections among military recruits in tropical singapore. *PloS ONE*, *6*, e26572.

Lee, V. J., Chen, M. I., Yap, J., Ong, J., Lim, W.-Y., Lin, R. T. P., et al. (2011). Comparability of different methods for estimating influenza infection rates over a single epidemic wave. *American Journal of Epidemiology*, *174*, 468-478.

Lee, V. J., Chow, A., Zheng, X., Carrasco, L. R., Cook, A. R., Lye, D. C., et al. (2012). Simple clinical and laboratory predictors of chikungunya versus dengue infections in adults. *PLOS Neglected Tropical Diseases*, *6*, 1-9.

Li, F., Choi, B., Sly, T., & Pak, A. (2008). Finding the real case-fatality rate of H5N1 avian influenza. *Journal of epidemiology and community health*, *62*(6), 555–559.

Lim, W.-Y., Chen, C. H., Ma, Y., Chen, M. I., Lee, V. J., Cook, A. R., et al. (2011). Risk factors for pandemic (H1N1) 2009 seroconversion among adults, singapore, 2009. *Emerging infectious diseases*, *17*(8).

Lu, C.-Y., Lee, C.-Y., Kao, C.-L., Shao, W.-Y., Lee, P.-I., Twu, S.-J., et al. (2002).

Incidence and case-fatality rates resulting from the 1998 enterovirus 71 outbreak in taiwan. *Journal of Medical Virology*, *67*, 217-223.

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *100*(26), 15324–15328.

Mawudeku, A., & Blench, M. (2006). Global public health intelligence network (GPHIN). In *7th conference of the association for machine translation in the americas* (pp. 8–12).

McKinley, T., Cook, A. R., & Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, *5*(1), 24.

McLean, E., & Paterson, B. (2010). *HPA weekly national influenza report - summary of UK surveillance of influenza and other seasonal respiratory illnesses* (Tech. Rep.). Health Protection Agency.

McLean, E., Pebody, R., Chamberland, M., Paterson, B., Smyth, B., Kearns, C., et al. (2010). *Epidemiological report of pandemic (H1N1) 2009 in the UK* (Tech. Rep.). Health Protection Agency.

Meltzer, M. I., Cox, N. J., Fukuda, K., et al. (1999). The economic impact of pandemic influenza in the united states: priorities for intervention. *Emerging infectious diseases*, *5*, 659–671.

Merler, S., Ajelli, M., Pugliese, A., & Ferguson, N. M. (2011). Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in europe: implications for real-time modelling. *PLoS computational biology*, *7*(9), e1002205.

MOH. (2009, December). *Ministry of health press room.* Website. Retrieved from `http://www.moh.gov.sg/content/moh_web/home/pressRoom/pressRoomItemRelease/2009.html`

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, *5*(3), e74.

Müller, P., Sansó, B., & De Iorio, M. (2004). Optimal bayesian design by inhomogeneous markov chain simulation. *Journal of the American Statistical Association*, *99*(467), 788–798.

Naylor, C. D., Chantler, C., & Griffiths, S. (2004). Learning from SARS in hong

kong and toronto. *JAMA*, *291*(20), 2483–2487.

Neumann, G., Noda, T., & Kawaoka, Y. (2009). Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature*, *459*(7249), 931–939.

Ng, L.-C., Tan, L.-K., Tan, C.-H., Tan, S. S., Hapuarachchi, H. C., Pok, K.-Y., et al. (2009). Entomologic and virologic investigation of chikungunya, singapore. *Emerging infectious diseases*, *15*(8), 1243.

Nicoll, A., & Coulombier, D. (2009). Europe's initial experience with pandemic (H1N1) 2009 - mitigation and delaying policies and practices. *Eurosurveillance*, *14*, 19279.

NIID. (1998). *National institute of infectious diseases outline - organization.* Website. Retrieved from `http://www.nih.go.jp/niid/en/aboutniid-2.html`

Nuraini, N., & Tasman, H. (2012). Simulation model for dengue infection. *International Journal of Basic & Applied Sciences*, *12*, 26-30.

O'Neill, P. D. (2002). A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical Biosciences*, *180*, 103-114.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M., & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by markov chain monte carlo methods. *Journal of the Royal Statistical Society*, *49*, 517-542.

O'Neill, P. D., & Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society*, *162*, 121-129.

Ong, J. B. S., Chen, M. I.-C., Cook, A. R., Lee, H. C., Lee, V. J., Lin, R. T. P., et al. (2010). Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in singapore. *PloS ONE*, *5*, e10036.

ONS. (2010, June). *Population estimates for uk, england and wales, scotland and northern ireland, mid 2009* [Website]. Website. Retrieved from `http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-213645`

Ooi, E.-E., Phoon, M.-C., Ishak, B., & Chan, S.-H. (2002). Seroepidemiology of human enterovirus 71, singapore. *Emerging infectious diseases*, *8*(9), 995–997.

Opatowski, L., Fraser, C., Griffin, J., Silva, E. de, Kerkhove, M. D. V., Lyons, E. J., et

al. (2011). Transmission characteristics of the 2009 H1N1 influenza pandemic: Comparison of 8 southern hemisphere countries. *PLoS Pathogens*, *7*, e1002225.

Pasha, G. R. (2002). Selection of variables in multiple regression using stepwise regression. *Journal of Research (Science)*, *13*, 119-127.

Plummer, M. (2013). rjags: Bayesian graphical models using MCMC [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=rjags` (R package version 3-11)

Poggensee, G., Gilsdorf, A., Buda, S., Eckmanns, T., Claus, H., Altmann, D., et al. (2010). The first wave of pandemic influenza (H1N1) 2009 in germany: From initiation to acceleration. *BMC Infectious Diseases*, *10*, 155.

PRB. (2009). *Population reference bureau 2009 world population data sheet.* Periodical. Retrieved from `http://www.prb.org/pdf09/09wpds_eng.pdf`

Presanis, A. M., Angelis, D. D., Team, T. N. Y. C. S. F. I., Hagy, A., Reed, C., Riley, S., et al. (2009). The severity of pandemic H1N1 influenza in the united states, from april to july 2009: A bayesian analysis. *PLoS Medicine*, *6*, e1000207.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rasch, D., Pilz, J., Verdooren, R., & Gebhardt, A. (2011). *Optimal experimental design with R* (S. Almgren, Ed.). Chapman and Hall.

Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., et al. (2009). Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in thailand. *New England Journal of Medicine*, *361*(23), 2209–2220.

Rhim, J. W., Go, E. J., Lee, K. Y., Youn, Y. S., Kim, M. S., Park, S. H., et al. (2012). Pandemic 2009 H1N1 virus infection in children and adults: A cohort study at a single hospital throughout the epidemic. *International Archives of Medicine*, *5*, 13.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., et al. (2003). Transmission dynamics of the etiological agent of SARS in hong kong: impact of public health interventions. *Science*, *300*(5627), 1961–1966.

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence

in approximate bayesian computation model choice. *PNAS*, *108*, 15112-15117.

Ross, J. V., Taimre, T., & Pollett, P. K. (2006). On parameter estimation in population models. *Theoretical Population Biology*, *70*, 498-510.

Saarinen, M., Järvinen, P., Haikala, O., & Ruutu, P. (2009, July). *The finnish health care prepares for an extensive epidemic of influenza A (H1N1).* Press Release. Retrieved from `http://www.thl.fi/en_US/web/en/pressrelease?id=15437?&print=true`

Schervish, M. J. (1995). *Theory of statistics.* Springer-Verlag, Inc.

Sentinelles. (2012, May). *French GPs sentinelles network.* Website. Retrieved from `http://www.sentiweb.org/`

Setzer, R. W. (2012). odesolve: Solvers for ordinary differential equations [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=odesolve` (R package version 0.9-9)

Shen, J., & Gao, S. (2008). A solution to separation and multicollinearity in multiple logistic regression. *Journal of Data Science*, *6*, 515-531.

Shih, S.-R., Stollar, V., Lin, J.-Y., Chang, S.-C., Chen, G.-W., & Li, M.-L. (2004). Identification of genes involved in the host response to enterovirus 71 infection. *Journal of NeuroVirology*, *10*, 293-304.

Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, *18*, 155-195.

Taubenberger, J. K., & Morens, D. M. (2006). 1918 influenza: the mother of all pandemics. *Rev Biomed*, *17*, 69–79.

Therneau, T. M. (2013). A package for survival analysis in s [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=survival` (R package version 2.37-4)

TNN. (2010, October). *Confusion prevails over diagnostic tests for dengue.* The Times of India Online News. Retrieved from `http://articles.timesofindia.indiatimes.com/2010-10-06/lucknow/28251822_1_dengue-cases-dengue-patient-health-directorate`

Tran, C. B. N., Nguyen, H. T., Phan, H. T. T., Tran, N. V., Wills, B., Farrar, J., et al. (2011). The seroprevalence and seroincidence of enterovirus71 infection in

infants and children in ho chi minh city, vietnam. *PLoS ONE*, *6*, e21116.

Treanor, J. J., Betts, R. F., Erb, S. M., Roth, F. K., & Dolin, R. (1987). Intranasally administered interferon as prophylaxis against experimentally induced influenza a virus infection in humans. *The Journal of Infectious Diseases*, *156*, 379-393.

Trifonov, V., Khiabanian, H., & Rabadan, R. (2009). Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus. *The New England Journal of Medicine*, *361*, 115-119.

Truscott, J., Fraser, C., Cauchemez, S., Meeyai, A., Hinsley, W., Donnelly, C. A., et al. (2012). Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal influenza. *Journal of The Royal Society Interface*, *9*(67), 304–312.

Ujike, M., Ejima, M., Anraku, A., Shimabukuro, K., Obuchi, M., Kishida, N., et al. (2011). Monitoring and characterization of oseltamivir-resistant pandemic (H1N1) 2009 virus, japan. *Emerging Infectious Diseases*, *17*, 470-479.

UN. (2010). *World population policies 2009* (Tech. Rep.). United Nations Department of Economic and Social Affairs/Population Division. Retrieved from `http://www.un.org/esa/population/publications/wpp2009/Publication_complete.pdf`

Unitedengue. (2014, March). *United in tackling epidemic dengue (unitedengue) @ONLINE.* Retrieved from `https://www.unitedengue.org/index.html`

Venzon, D., & Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, *37*, 87-94.

Verdinelli, I., & Kadane, J. B. (1992). Bayesian designs for maximizing information and outcome. *Journal of the American Statistical Association*, *87*(418), 510–515.

Victor, L. Y., & Madoff, L. C. (2004a). Promed-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, *39*(2), 227–232.

Victor, L. Y., & Madoff, L. C. (2004b). Promed-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, *39*(2), 227–232.

WHO. (2008, March). *Chikungunya.* Website. Retrieved from `http://www.who.int/mediacentre/factsheets/fs327/en/index.html` (Fact Sheet number 327)

WHO. (2010, August). *Pandemic (H1N1) 2009.* Website. Retrieved from `http://www.who.int/csr/disease/swineflu/en/`

WHO. (2013, September). *Dengue and severe dengue.* Website. Retrieved from `http://www.who.int/mediacentre/factsheets/fs117/en/index.html` (Fact Sheet number 117)

Wilkinson, R. D. (2013). Approximate bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, *12*, 129-141.

Wu, J. T., Ma, E. S. K., Lee, C. K., Chu, D. K. W., Ho, P.-L., Shen, A. L., et al. (2010). The infection attack rate and severity of 2009 pandemic H1N1 influenza in hong kong. *Journal of the Royal Society of Tropical Medical and Hygiene*, *51*, 1184-1191.

Yong, E. (2012). Trials at the ready: preparing for the next pandemic. *BMJ: British Medical Journal*, *344*, e2982.

Ypma, T. J. (1995). Historical development of the newton-raphson method. *Society for Industrial and Applied Mathematics*, *37*, 531-551.

Yu, H., Wang, M., Chang, H., Lu, J., Lu, B., Li, J., et al. (2011). Prevalence of antibodies against enterovirus 71 in children from lu'an city in central china. *Japanese journal of infectious diseases*, *64*, 528-532.

Zeng, M., Khatib, N. F. E., Tu, S., Ren, P., Xu, S., Zhu, Q., et al. (2012). Seroepidemiology of enterovirus 71 infection prior to the 2011 season in children in shanghai. *Journal of Clinical Virology*, *53*, 285-289.

Zuno, S. E., Aranguré, J. M. M., Obeso, A. J. M., Muñiz, C. G., Pérez, E. R., León, M. G., et al. (2009). Infection and death from influenza A H1N1 virus in mexico: a retrospective analysis. *The Lancet*, *374*, 2072-2079.