PREDICTION OF INFECTIOUS DISEASE OUTBREAKS BASED ON LIMITED INFORMATION

VINCENT-ANTHONY MARMARÀ

Doctor of Philosophy

Mathematics

University of Stirling

September 2016

DECLARATION

I, Vincent-Anthony Marmarà, confirm that the work presented in this thesis is, to the best of my knowledge, original. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Vincent-Anthony Marmarà

September 2016

ABSTRACT

The last two decades have seen several large-scale epidemics of international impact, including human, animal and plant epidemics. Policy makers face health challenges that require epidemic predictions based on limited information. There is therefore a pressing need to construct models that allow us to frame all available information to predict an emerging outbreak and to control it in a timely manner.

The aim of this thesis is to develop an early-warning modelling approach that can predict emerging disease outbreaks. Based on Bayesian techniques ideally suited to combine information from different sources into a single modelling and estimation framework, I developed a suite of approaches to epidemiological data that can deal with data from different sources and of varying quality. The SEIR model, particle filter algorithm and a number of influenza-related datasets were utilised to examine various models and methodologies to predict influenza outbreaks. The data included a combination of consultations and diagnosed influenza-like illness (ILI) cases for five influenza seasons.

I showed that for the pandemic season, different proxies lead to similar behaviour of the effective reproduction number. For influenza datasets, there exists a strong relationship between consultations and diagnosed datasets, especially when considering time-dependent models. Individual parameters for different influenza seasons provided similar values, thereby offering an opportunity to utilise such information in future outbreaks. Moreover, my findings showed that when the temperature drops below 14°C, this triggers the first substantial rise in the number of ILI cases, highlighting that temperature data is an important signal to trigger the start of the influenza epidemic. Further probing was carried out among Maltese citizens and estimates on the under-reporting rate of the seasonal influenza were established. Based on these findings, a new epidemiological model and framework were developed, providing accurate real-time forecasts with a clear early warning signal to the influenza outbreak.

This research utilised a combination of novel data sources to predict influenza outbreaks. Such information is beneficial for health authorities to plan health strategies and control epidemics.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Professor Adam Kleczkowski for his support and guidance throughout this journey, for his helpful advice, professional support, encouragement, and invaluable feedback that has shaped my way of thinking. Adam has helped me during this challenging journey with his patience, inspiration and expertise, which have been fruitful to my professional growth.

I am grateful to the Malta Health Promotion Department, namely Dr. Charmaine Gauci, Dr. Tanya Mellilo and Dr. Jackie Mellilo, who provided the data for this study and who took the time to provide prompt replies to my questions. Thanks also go to the Malta Airport Meteorological Services for providing the Maltese temperature data.

I would like to extend my gratitude to Professor Alex Cook for allowing me to use his particle filter algorithm R code. Furthermore, I would like to thank him for reviewing my research paper.

I would like to acknowledge the Maltese key health officials who took an interest in my research and made time to discuss the outcomes of this thesis. Thanks go to Dr. Renzo Degabriele (Chief Executive Officer, Primary Health Care Department), Dr. Neville Calleja (Director, Health Information and Research) and Mr. Mike Farrugia (Ministry Advisor). Particular gratitude goes to the Minister for Health in Malta, Honourable Mr. Chris Fearne for taking the time to discuss my findings.

I would also like to thank all the secretaries at the Department of Mathematics and Computing Science at the University of Stirling, who gave me secretarial and administrative support, and to my fellow PhD students, in particular Mr. Paul McMenemy, for being a source of personal support, for the teas and coffees, for the lifts to the airport, and for the regular dinners at Stirling's Molly's.

I extend my acknowledgements to the 'Times of Malta' newspaper for reporting my research findings (Appendix J).

I could not have done this without the support and encouragement of my wonderful family, my parents, Josephine and Charlie, my sisters, Fiona and Olivia-Ann, my nephews Matthew, Gabriel and Andre, and my in-laws. Thank you for always believing in me and for being there when I needed you most throughout this PhD.

Lastly, I would like to thank my wife, Danika, for her unwavering support and patience, for being an excellent sounding board at the end of my thesis, and for her rational influence on me throughout this journey. Thank you for encouraging me to stick at it and for always believing that I could do this.

Understanding the facts that no one can see…

This thesis is dedicated to my wife, Danika, who was my pillar of strength, for her love, understanding, and continuous support.

PUBLICATIONS

V. Marmara, A. Cook, A. Kleczkowski, Estimation of force of infection based on different epidemiological proxies: 2009/2010 Influenza epidemic in Malta, Epidemics, **9** (2014) 52-61.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction & literature review

1.1 Introduction

The aim of this study is to predict infectious disease outbreaks based on limited information. This thesis shall discuss early warning techniques that have the potential to provide signals to clinicians on the spread of diseases. This thesis will also focus on parameter estimation for various influenza datasets through the use of mathematical modelling. The study probes into the underlying factors related to influenza in order to improve the available information for the Maltese population. Ultimately, throughout this thesis, I aim to provide various techniques to predict the outbreak in real-time and as early as possible. The different methods are illustrated using real-life influenza outbreak data from Malta spanning five seasons from 2009 to 2015. This chapter serves as an introduction to the main themes of this thesis. I provide an overview of the history of epidemics both internationally and in Malta, followed by a literature review of epidemiological modelling, which is the paradigm I shall follow throughout this thesis. The last part of this chapter contains a brief overview of the chapters produced in this thesis.

1.2 Background

The history of epidemics goes back centuries and their associated human morbidity and mortality was a concern for a number of generations [1]. It is estimated that during the 14th Century, 25 million Europeans died from the Bubonic plague, representing between 30-60% of the whole population [2]. During the year 1520, about half of the population of Aztecs probably died due to smallpox and around 150 years later, 68,000 people died in London due to the plague epidemic [2]. Another 2.5 million are thought to have died from Typhus in Russia during World War 1 and during that same period, around 20 million people are estimated to have died from the world epidemics of influenza [2].

The value of scientific research in the field of epidemiology has long been recognised [3], in particular with the development of the 'germ theory of disease' [2]. This theory states that some diseases are caused by microorganisms (pathogens) and the diseases they cause are called infectious diseases [2]. Mathematical modelling also has a long history in the area of epidemiology [2]. Numerous developments in the area of mathematical epidemiology led to the availability of widespread information, improved understanding of the spread of disease, and advances in the area of medicine and computer programming

[3]. Subsequently, countries began to reap the benefits of understanding the spread of disease due to the setting up of surveillance systems across the globe [4]. Guidelines and incentivized vaccination programmes have been established across the years to prevent or control widespread transmission [5] and to increase vaccination rates in various populations [6].

Amidst such progress in the epidemiological field, there is still room for substantial improvement to better understand the dynamics of the spread of epidemics [2], owed to the continuous outbreaks of new influenza viruses affecting various populations [7]. Influenza epidemics bring with them serious health complications such as physical illness or death, or that pose a risk for people with weak immune systems [8, 9]. These implications result in an extensive burden on the health sector [10, 11] and welfare states [12]. This highlights the pivotal role and rising impact of mathematical modelling in epidemiology to map and predict the future state of populations [13] and most importantly, to quantify the uncertainty in these predictions [14]. In turn, this informs public health decision-making on the likelihood of an infectious disease outbreak, how the disease will spread and how it can be controlled [15].

1.3 History of Malta's Influenza Epidemics

Malta is a small island at the centre of the Mediterranean Sea and lies in between Libya (Africa) and Sicily (Italy). Malta is considered one of the most densely populated countries around the world with a population of around 414,000 in 2013 [16] and a total area of 216km^2. During the second quarter of 2015, the employment rate in Malta was estimated to be around 184,871 [17] with just under 5,400 unemployed individuals (≈3.9%) in 2016. This places Malta as the second best country in the European Union for achieving the lowest unemployment rate.

The first reference to influenza, in epidemic terms in the Maltese islands, was in a petition that was sent to the Grandmaster in 1682 by the Gozitan Apothecary [9]. During that time, the Apothecary requested funds for drugs which were required during an epidemic in Gozo. During 1730, it was estimated that all the Maltese population was infected with 'catarrhal influenza' and slight fever. Other major influenza outbreaks in Malta were recorded during 1733, 1746 and 1754 [9]. The term 'influenza' was then applied in the

Maltese context in 1803, in relation to a specific acute viral respiratory disease, during which time the infection caused an epidemic in Britain [18]. In 1836, almost all of the population in Malta at the time contracted the influenza [18]. Eleven years later, another outbreak hit the Island and most factions of the Maltese society were also affected. Approximately half a decade later (1890), compulsory notifications regarding influenza were introduced in Malta. During the same period, Asiatic Flu reached the Maltese islands and the case fatality rate was estimated to be 4% [18]. A revival of the same flu occurred in 1892 and 1894, recording slightly lower case fatality rates.

The Spanish Flu, a leading cause of death for more than 20 million individuals worldwide during 1918-1919, reached Malta in June 1918 and subsided a year later. The Spanish Flu reached Malta in three phases; the first wave occurred during June-August with a case fatality rate of 5.1%, followed by the second wave (September-November) with a case fatality rate of 3.9% and a third wave during March 1919 [18].

Another influenza outbreak occurred in 1920, though this was not severe. A year later saw the start of an epidemic in Malta, which consisted of two waves with a case fatality rate of 1.8%. Other outbreaks were recorded in 1929, 1936/1937, 1940-1943, 1948 and 1951/1952 [18]. Subsequently, the Asian Pandemic (H3N2) reached Malta in 1957 and had a case fatality rate of 0.13%. During 1968-1969, the Hong Kong Flu (H3N2) made its appearance in Malta, but with minimal number of reported deaths [18]. The subsequent Russian Flu (H1N) which occurred during 1977-1978 did not have any impact on the Maltese population. Following the last pandemic dating more than 30 years ago, a significant influenza pandemic (H1N1: Hemagglutinin Type 1 Neuraminidase Type 1, aka swine flu) reached Malta in 2009. This pandemic shall be discussed in detail in this dissertation.

1.4 Mathematical modelling in epidemiology

The applications of mathematical modelling in the area of infectious diseases appear to have emerged by Daniel Bernoulli during the 18th century to study the strength of mathematical methodologies against small pox in England [19]. It was only until the late 19th century that other researchers studied mathematical epidemiology yet again, for example William Far who fitted a normal curve to a smoothed quarterly small pox data

[2]. Later on, other mathematical epidemiologists made their important mark in this area of research. During the early years of the 20th century, John Brownlee published a research paper about the theory of epidemics [2]. During the same period, William Hamer and Ronald Ross applied the post germ theory to two specific quantitative issues and were the first epidemiologists to formulate specific theories related to the transmission of infectious diseases [2]. The work of the latter researchers, together with the research studies of Hudson, Soper, Kermack and McKendrick provided a solid base about the theoretical framework of observed diseases. Hamer and Ross used the important 'Mass Action Principle' to describe the epidemic behaviour, while Kermack and McKendrick developed the classical SIR model [2]. This opened the field of mathematical epidemiological modelling to further investigation on infectious disease dynamics and epidemiological phenomena.

Mathematical models exist to make more sense of the available data by enabling the estimation of disease parameters to understand the dynamics and control of infectious diseases [3]. Epidemiological mathematical models provide a framework for predicting epidemiological dynamics, though this field is still evolving due to the number of uncertainties found in various epidemiological data [14]. Various techniques exist to estimate the number of affected individuals at different time points for different model compartments. The above mentioned SIR model can be described by the following equations for a closed population:

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

This set of differential equations describe the number of Susceptible individuals S, i.e. individuals in the population who have not been infected but are at risk. The number of Infected individuals I refers to those individuals who are infectious and hence can transmit the disease; β is the infection rate. R is the number of Removed individuals that are no longer at risk of acquiring the disease because they are either immune or deceased, while γ is the recovery rate once the individual is infected and N is the population size. $S(t)$, $I(t)$ and $R(t)$ are functions of time t, and initial conditions are set appropriately. This is the simplest model which is designed for different stages (compartments) of the

disease. The model can be further enhanced by considering further compartments, such as the Exposed (*E*) individuals, and is known as the SEIR model.

1.4.1 Deterministic and stochastic compartmental disease models

Several studies have compared deterministic and stochastic models [20-23] in order to demonstrate the importance in their relationship. A deterministic model can be described by a set of ordinary differential equations in a single system (as shown above), while the stochastic model can be a Markov population process with continuous time and discrete space [21].

A deterministic compartmental model assumes that the population is homogenous; hence all people are the same and only differ in their disease state. In comparison to stochastic models, deterministic epidemic models are mathematically less complex and usually lead to powerful qualitative results [20]. Due to this reason, for a long period of time research work related to deterministic models dominated this research field [24]. Additionally, a deterministic model only deals with proportions rather than a finite population size [20] but are a good approximation of related stochastic models. Extant studies focus on stochastic modelling to improve epidemic models [24] by predicting the expected extinction time of the disease, as opposed to deterministic models [21]. This is one of the main differences between the two models. Another difference is that stochastic models provide a coherent picture of the uncertainty and variability that is related to the real-life epidemics due to factors such as the randomness of person-to-person contact [23]. The stochastic model can capture individual behaviour as well as the probability of the occurrence of an event.

During the past years, substantial progress was achieved in the applications of Bayesian inferential methods for epidemiological data through the use of stochastic compartmental models. In many cases, such models employ the Markov Chain Monte Carlo (MCMC) methods [25]. These methods are widely acknowledged nowadays, because they not only incorporate uncertainty in parameter values, but more importantly the population size of infected hosts is random. This includes the effect of a possible extinction and the re-emergence of an infection, the prediction of an individual realization of an epidemic and the understanding of the suitable period of application of a control treatment. Results

obtained from these models can be used to inform policy makers to plan health strategies and to understand the effectiveness of proposed control measures. We aim to infer biological processes from epidemiological patterns to control the epidemic.

1.4.2 The Bayesian Inference

Bayesian inference or likelihood inference is of fundamental importance in the field of mathematical statistics [25]. Nowadays, this technique is utilized in a wide range of statistical fields and will be used substantially throughout this dissertation. Bayesian inference requires sampling models that produce the likelihood function together with a conditional distribution of the data, given the parameters of the model. The Bayesian approach takes into account a prior distribution on the parameters of the model. Following this, the likelihood function and the prior distribution are combined through the use of the Bayes' theorem to compute the posterior distribution [26]. The posterior distribution is a conditional distribution of a set of unknown quantities, given that there is some observed data. This is the main distribution from which all Bayesian inference arises. The use of Bayesian techniques has grown rapidly in recent years [25]. Computers, together with powerful software, have contributed to the development of Bayesian techniques due to the power required to run such models. Such developments were well complemented with a class of iterative simulations methods known as the Markov chain Monte Carlo (MCMC) algorithms.

1.4.2.1 The Markov Chain Monte Carlo models

There is widespread activity and application of Markov Chain Monte Carlo (MCMC) models in various fields [27]. These models are not restricted to a limited number of applications and have thus been of substantial benefit in the finance and gaming industries. In the area of epidemiology such models are widely employed due to missing data. It is considered a standard approach to apply MCMC models when missing data occurs [28], thereby applying the right imputation techniques to 'fill-in' the missing gaps. MCMC models and the Bayesian framework offer an opportunity to address the arising challenge of missing data through the inclusion of extra parameters in the model [29]. In order to produce a likelihood function, estimation of missing data becomes a part of the model fitting mechanism [29].

The knowledge of epidemiological outbreaks has been improved through the use of these MCMC models as they provide further information to understand the mechanisms, diseases and main parameters of an outbreak [29]. In return, this information is extremely useful for control strategies, policy makers and health interventions. Parameter estimation by these MCMC models is of prime importance for epidemic predictions. Prior distributions in such mathematical models might influence considerably the accuracy of several Bayes factors and hence influence outbreak predictions [30]. MCMC models are mostly important in order to understand the transmission parameter estimates between different stages of an outbreak [31-33]. Therefore, acquiring accurate parameter values will improve the stochastic epidemiological compartmental models.

The epidemiological models can be complex [14]. Several attempts have been carried out to simplify the MCMC algorithms, so that they can be straightforwardly applied by non-experts [34]. This is being done at the detriment of assuming a lower number of parameters.

In this dissertation, I use the MCMC technique to obtain parameter values in conjunction with other statistical techniques/models to provide real-time forecasts. Although we are interested to understand and estimate parameters of an outbreak, simultaneously we need to establish the right control epidemic strategies as early as possible [35]. Thus, it is of great importance to find the right balance between these two objectives by optimizing historical information and current real-time data.

1.4.2.2 Particle filter algorithm

Particle filter algorithms are widely used to improve the prediction processes and the parameter estimates. Substantial studies use such algorithms [26, 36-38] which are considered the gold-standard tools in mathematical modelling [39]. Such algorithms provide an opportunity to estimate recursively a system of state variables and to apply inferential techniques on the model parameters [40]. Such filtering methods include basic particle filter (PF) [41], maximum likelihood estimation via iterated filtering (MIF) [42], particle Markov chain Monte Carlo (pMCMC) [39] and several ensemble filter variants [43-45]. Such methods can be used together with epidemiological models and reported influenza datasets to estimate parameters of the epidemiological model. For example, we

can obtain various estimates of the transmission rates between different epidemiological stages.

Particle filters are sequential Monte Carlo methods based on particles [41]. In order to apply a particle filter, there are various sequential simulation methods/algorithms [41, 46]. These methods use state-space model together with Kalman filters, particularly on time dynamic models that are usually non-linear and non-Gaussian. Arulampalam et al. (2002) [41] presented various algorithms, all focusing on particle filtering but with several variants. These include sampling importance resampling (SIR) filter, auxiliary sampling importance resampling (ASIR) filter and regularized particle filter (RPF). All filters are derived from the sequential importance sampling (SIS) algorithm. Doucet et al. (2000) [46] described in detail the various stages within these algorithms, as well as an analytical description of the efficiency of these models and their limitations.

During these last two decades, substantial research studies were carried out to analyze the implementation of such algorithms in epidemiological theory. Currently, these are the latest techniques to obtain reliable parameter estimates and accurate forecasting. Such algorithms have the flexibility to amend various steps in order to explore possibilities to improve the results. Extensive work is being carried out in this area to improve the understanding and application of such algorithms. In recent years, Ionides et al. [42] proposed new theoretical results in relation to the above particle filtering (PF) technique. Throughout the latter study, the researchers proposed a method on how to model state parameter estimates updated on multiple rounds of particle filtering, hence resulting in multiple iterations. On the other hand, the basic particle filter algorithm updates the state parameter values based on every single time point, producing one individual round of particle filtering.

Other variations exist, such as the ensemble filter variants, which differ only in how the observed variables are being updated [40]. Such mechanisms include the ensemble Kalman filter (EnKF) [43], the ensemble adjustment Kalman filter (EAKF) [44] and the rank histogram filter (RHF) [45]. Several filtering techniques might produce more accurate forecasts for different datasets, based on the characteristics of the filtering method [40].

Throughout this dissertation, the particle filtering algorithm is used as a main tool in this dissertation, where weighting, resampling and kernel smoothing are applied. However it is not our aim to analyse different particle filtering techniques, but to use these Bayesian techniques together with different data sources to obtain accurate estimates of epidemiological parameters.

1.4.2.3 Implementation to the S(E)IR models

Markov Chain models are widely used in epidemiological models, such as SIR, SEIR and SIRS models. Such epidemiological models are applied on various forms of data and are also useful when limited data is available. Most popular research papers are those where the implementation of such models is carried out on Influenza-Like Illness (ILI) and virology datasets [47]. There are other studies where the SEIR model was fitted on observations where only the removed (R) compartment, such as deaths, were recorded [48].

The above MCMC models, together with particle filtering algorithms, are the most important tools for the formulation of complex SEIR models [31]. Epidemiological models carry substantial uncertainty and hence through the use of Bayesian framework, one can estimate the unknown information and parameters of the epidemiological models. Such studies aim to account for control measures related to the spread of the infection period [35].

Researchers employed the SEIR model to analyze the immigration of infected individuals and the efficiency of the stochastic variation of the infection [49]. Others used the latter model to study the stability of the equilibrium points of the SEIR model [50-51] from a more theoretical perspective. Further theoretical analysis was carried out by Artalejo et al. (2015) [52], where efficient computation procedures and algorithms were studied to analyze the stochastic SEIR model.

Additionally, the design of control strategies is of particular importance in such studies. The SEIR model is commonly used as a model design control strategy to protect susceptible individuals from getting infected [53] and as an important tool to determine the best vaccination policies through the spread of the disease [54]. Others used the SEIR

model to apply real-time forecasting [26, 55], as defined throughout subsequent sections. In a substantial number of research studies, the epidemiological models were used primarily to estimate the reproduction number [56] (see below), since this gives a clear indication about the severity of the epidemic.

Throughout this dissertation, the SEIR model is used in different ways. Initially, it is employed to compare the parameter values between different data sources, more specifically the reproduction ratio. Then the SEIR model, together with other mathematical techniques, is used to predict the spread of the seasonal influenza outbreak, the severity of influenza, as well as the influenza peak.

1.4.3 The basic reproduction number (R_0)

The most important parameter in epidemiological modelling is the basic reproduction number (R_0), which value dictates whether or not a large epidemic outbreak can occur [25]. In simple terms, this value represents the number of secondary cases caused by an infectious individual in a completely susceptible population. When the reproduction number is greater than 1, the infectious disease will spread, resulting in a major epidemic. When this value is less than one, the disease will fail to spread. Therefore, when R_0 is greater than 1, there is a positive probability that a large number of individuals contract the disease, while when R_0 is smaller than 1, only a limited number of individuals will get infected. This value provides direction on whether the population is at risk from any emerging disease.

Mathematical models attempt to predict this crucial number in order to communicate it as a good indicator to health authorities. A substantial part of this dissertation focuses on estimating R_0 for influenza outbreaks in Malta.

Linear programming methods can be used to obtain acceptable bounds for the mean of R_0, given the time at which an active epidemic is observed [25]. In addition, it is interesting to analyse the dynamics of R_0, thus understanding how pathogens spread and transmit within their host populations [26, 57-65]. Most mathematical models assume that either the contact rate between hosts is linearly related to host density (density-dependent) or that the contact rate is independent of density, thereby considered as

frequency dependent [66]. Parameters, such as R_0, may prove to be difficult to estimate; in some datasets they cannot be estimated consistently from the final data due to multi-type[1] epidemic model [67]. Additionally, in some complex epidemics based on spatial-temporal evolution (different transmission models from one area to the other), Bayesian Markov Chain Monte Carlo methods are the best established algorithms to model the reproduction number [68]. Such reproduction ratio analysis helps to understand several characteristics of different epidemics and to use such information for future outbreaks [69]. Epidemiological studies and the reproduction ratio aim in understanding the evolution of infectious disease in real time [70], hence historic information might support such an objective.

Different researchers calculate the reproduction number from different sources of data. Some of the common data sources are the laboratory confirmed influenza cases [69-72], Influenza-Like Illness (ILI) [73-75] and serological data [76-78]. In my research paper [79], several different data sources were analysed with the intention to understand the reproduction ratio of four different related datasets. The data included the number of GPs consultations, the number of ILIs, swabbed H1N1 cases and confirmed H1N1 positive cases. The remaining part of the thesis introduces other datasets.

As discussed above, substantial research papers make reference to the basic reproduction ratio R_0, where it represents the average number of cases generated from another infectious individual during the course of the outbreak. Hence, R_0 does not vary over time. On the other hand, the effective reproduction ratio R_t varies over the time of the outbreak and for different seasons [80]. Thus, R_t represents the number of new infectious individuals at a given time *t* in the epidemic. Understanding and capturing all available information helps to explore the uncertainty that the reproduction ratio carries. This non-constant factor R_t may be influenced through several health control strategies [80-81] during the progression of the outbreak.

[1] Different parameter values for the same parameters due to different data demographics

1.5 Influenza

1.5.1 Defining the seasonal influenza

Seasonal influenza is one of the major epidemics that occurs on a yearly basis [82]. It has major implications towards healthcare services as its outbreaks occur frequently and are generally characterized by high levels of activity in the hospital setting, thus carrying a yearly cost which varies according to the severity. Such implications continue to emphasize to health authorities and policy-makers the importance of a comprehensive influenza prevention strategy and accompanying interventions that need to be designed well in advance and that can be applied across the entire spectrum of healthcare settings.

There are several different definitions related to seasonal influenza, but most converge to the same major symptoms. For instance, the UK National Health Service (NHS) states that the symptoms related to the influenza usually develop during the first three days upon becoming infected [83]. The NHS also highlights the major symptoms of seasonal influenza, which include a high temperature of 38°C or above, tiredness and weakness, headache, general aches/pains and a dry cough. All of the latter symptoms are similarly defined by the World Health Organization (WHO) [82]. However, WHO's definition also includes sore throat and runny nose. A definition is provided by the Center for Disease Control and Prevention (CDC) which is similar to that defined by the WHO [84]. In addition, CDC states that it is more common for children to experience vomiting and diarrhoea. The same definition as the one used by the NHS has been applied by the Health Authorities in Malta.

The terms 'influenza' and 'cold' are two different illnesses, although sometimes it is difficult to distinguish between the two [83]. Influenza symptoms tend to appear more quickly and usually include fever and aching muscles, making it more difficult to continue with the normal routine [83]. On the other hand, a cold is an illness which develops more gradually and mainly affects the nose and throat, thereby allowing an individual to continue with routine daily activities. Additionally, influenza might seriously affect several high risk people, especially young children (aged 2 years or younger), adults aged 65 year or older, pregnant women, people with several medical conditions including chronic diseases, and individuals with a weak immune system [82].

It is unlikely that an individual will get infected more than once with the exactly same strain of influenza during the same season [85]. However, there are several occasions where an individual can contract an infection more than once during the same season. Mainly this happens when an individual does not develop full immunity for seasonal influenza (or is not fully recovered), or when a person is affected by different strains of the influenza virus [84]. The Malta Health Promotion Department (MHPD) claims that a person usually contracts the influenza only once in a season, due to the circulation of one of the viruses which is more dominant (Appendix A, Meeting with the MHPD).

Seasonal influenza is widely described and tackled from different perspectives. Several research papers focus on the vaccine uptake, such as the intention to receive future influenza vaccine and uptake rates [86-87] or to investigate the knowledge, attitudes and practices of individuals regarding seasonal influenza vaccination [88-91]. Others focus on the economic aspects of the seasonal influenza, such as the estimation of the direct/indirect costs in relation to outpatient visits and hospitalisation [92], or the vaccine administration costs and vaccination costs [93].

Most of the common topics found in the aforementioned research papers mainly focus on the medical determinants of seasonal influenza and its health implications. In this dissertation, seasonal influenza datasets are being investigated mainly from the mathematical perspective in order to predict the outbreak. Several aspects of the seasonal influenza which were explored throughout this research provide new insights to the dynamics of the seasonal influenza.

1.5.2 The dynamics of influenza in relation to climate and temperature

Several important factors affect spread of influenza, including school holidays, seasonality, immunity and vaccination. However, one of the most important factors is the climate, predominantly the temperature. Some researchers focus on the temperature as a basis to account for the seasonal variation related to mortality and hospitalisations [94]. Others focused on the dynamics of the transmission of influenza in relation to the influence of climate conditions [95]. These studies found significant associations in relation to the influenza transmission and the minimum temperature. Other research studies compared the number of diagnosed individuals for bronchitis in relation to several

important variables such as influenza outbreak and low air temperature [96]. The findings suggest that there exists a correlation between these variables.

In a systematic review by Irwin et al. (2011), persistence of the influenza virus was studied for several different environmental conditions, including air temperature [97]. In this study, temperature, was categorized into three levels (2 to 12°C, 17 to 27°C, and > 27°C) to evaluate influenza virus persistence. It was found that the persistence of influenza was found to be longer at lower temperatures [97].

Several research papers used the temperature variable to model seasonality in avian influenza H5N1 [98]. Low temperature values across countries were associated with high intensity outbreaks, but were not associated for countries when the temperature remains constant throughout the year [98]. Other climate factors could play a role in understanding the spread of influenza. For example, in Indonesia and Egypt, the peak of the outbreaks corresponds to a wet season, while in Vietnam the peak corresponds to a dry season [98]. In addition, one needs to utilize the climate factors according to the characteristics of different countries. Although minimum temperature as an indicator for influenza may be consistent over different regions, this would nonetheless differ on a global scale [99]. Lower temperatures might encourage more crowding among populations, hence increasing the chances of influenza transmission [98]. Similar characteristics within the population, as well as related-influenza characteristics help in predicting other outbreaks whereby similar influenza transmission features might lead to similarities between different epidemics [100]. Similarly, there are also extensive studies related to poultry outbreaks which found an association between low temperatures and such outbreaks [98, 101].

It is clear from the extant literature that the transmission of the influenza outbreak is dependent on climate conditions, especially temperature [102]. It is also clear that there is an association between low air temperature and the spread of influenza. In this thesis, I address this relationship for the Malta datasets.

1.5.3 The role of surveys in studies related to Influenza

Cross-sectional surveys have a principal role in vast research fields such as marketing, media and political studies. These are considered important tools to explore the key determinants of the population under study. Surveys are also utilised in studies related to influenza [103-105], for example to predict the actual spread of an outbreak within a country.

Surveys related to influenza can incorporate different methodologies. For example, cross-sectional serological studies are used to explore the response to immunity before and after an influenza outbreak [103] or to estimate the proportion of symptomatic infected cases [104], or to estimate influenza infection rates [105]. Serological studies are very popular in epidemiology to understand various characteristics related to outbreaks and the main predictors related to an individual's risks in acquiring the influenza [106]. For example, in a research study by Soh et al. (2012), cross-sectional serological surveys were carried out to estimate the actual infection rates of school-aged children [107]. Other surveys quantitatively assess the knowledge and attitudes towards influenza vaccination amongst different populations [88-90, 108].

One of the known influenza surveys is the UK flu survey [109]. This online system of monitoring influenza is part of a European project with ten participating countries under the project name InfluenzaNet [109]. Participants are reminded to record and report their symptoms on a weekly basis. Such surveys aim to observe the spread of influenza through responses over the internet regarding participants' influenza-like illness (ILI) symptoms. Although such surveys help to monitor the spread of influenza, such data tends to have considerable bias towards those individuals that have internet access, and so many individuals do not have the same opportunity to participate in this survey. These surveys tend to be biased towards those with a higher level of education and who are younger in age. Additionally, such surveys tend to be biased towards those individuals who work in an office environment as they have continuous access to the internet. Amidst such bias, the acquired data can still be strongly indicative of the spread of influenza. UK flu survey data [109] is used by the research team at the London School of Hygiene and Tropical Medicine and Public Health England to monitor flu trends in the UK [109]. Since the latter survey data is available online, several researchers make use of such information.

For example, such a dataset was used by Camacho et al. [110] to analyse the duration for cases of ILI and acute respiratory infections (ARI). Their research findings [110] were analysed against several demographics. Others used the UK flu survey data to measure ILI and its related risk factors [111], suggesting that vaccination is linked to the reduced risk of becoming ill with ILI [111].

A similar model to the above UK flu survey is found in Portugal [112], where researchers use the online survey data to analyse the incidence rates of influenza for different locations within the country. Such work coincides with further development in the area of mathematical models and computational platforms. Similarly, in France [113], researchers analysed real-time data to study the spread of the influenza disease. In Spain, other researchers made use of their data to compare the incidence rates of countries that are participating in this project [114]. In addition, Spanish data is being used to understand the mechanisms of the spread of the influenza.

Participation and response rates in epidemiological surveys are very important [115]. Therefore, the right methodologies are needed to ensure that the response rate is a satisfactory one with limited research bias [115]. Such studies can already contain certain elements of bias, since several responses are based on the respondents' medical knowledge. For a good number of questions, respondents often base their judgement on self-medical diagnosis; this is considered an important element in epidemiological studies as it supports pandemic control strategies through self-management practices and the reduction of visits to healthcare facilities, thereby aiding to contain viral spread [116]. Self-reports have been compared to electronic medical records [117] in order to examine the accuracy of self-report vaccination status. Nonetheless, there is limited evidence about the accuracy of self-reports of influenza, particularly during pandemics [116], warranting further in-depth analysis, as found in this thesis.

Tan et al. (2013) found that surveys provided useful information about key epidemiological parameters in relation to the influenza [104]. Of particular importance is the use of surveys to identify several missing gaps from different perspectives. Although not thoroughly studied, surveys can be used to obtain improved and more informative prior distributions [118]. However, limited research exists about nationwide cross-sectional surveys to improve the understanding of the prior distributions as well as the

under-reporting rates (the percentage of influenza cases that are not reported by GPs or by any other health authority) of the influenza outbreaks. Most of the research papers related to this topic focus on serological surveys rather than nationwide cross-sectional surveys. Telephone surveys as used in this thesis, might offer a good solution to solve missing gaps about knowledge related to the influenza of individuals [119].

1.5.4 Influenza forecasting

The ultimate aim of the above research studies is to acquire enough information to forecast an emerging outbreak. However, time is a crucial factor in such studies. Thus, our objective is to create the real-time forecast as early as possible throughout the outbreak, based on the fact that one extra day could cost extra lives [120].

In one of the latest systematic reviews on the forecasting of the influenza outbreak dynamics, Nsoesie et al. (2014) focused on research studies designed in forecasting influenza outbreaks at local, regional, national or global level [120]. The systematic review discusses several models, namely the time series models, non-parametric forecasting (used in meteorology), SIR and SEIR models (including particle filtering), agent-based models and meta-population models. Some of these models use historical data and other current factors related to the influenza outbreak. In this systematic review it was found that several papers discuss the forecasts retrospectively, but the major challenge is evaluating and assessing the performance of such methodologies in real-time.

Researchers used several techniques to analyse the accuracy of the predictions. Some researchers employ correlation analysis to analyse the predicted values against the observed values [121-123]. Other methods used were percentage errors [124], root mean squared error [123], proportion of correct predictions [125] and confidence intervals [126]. For these research papers, the correlation varied between 58% and 93.5%, when comparing between the observed and predicted values.

Several studies focused on specific characteristics. For example, one particular study by Soebiyanto et. al (2010) used the temperature data as an input series, together with an ARIMA model to improve the accuracy of forecasted data [123]. Other research studies

attempted to estimate the percentage of infection rate within the population [127], whereby it was estimated that during the 2009 pandemic, between 57% and 63% individuals were infected. The latter study also focused in predicting the peak of the influenza. Other researchers focused their forecasting on web-based estimated of influenza activity [126, 128]. They found that the peak of the outbreak can be predicted 6 or 7 weeks in advance. However, web-based estimates carry certain level of uncertainty due to errors in capturing influenza trends [120]. The peak of the influenza can be predicted using the distribution of previous influenza seasons; however, it is not always easy to predict the height that corresponds to the peak [120].

In a research paper published by Shaman et al. (2013), the researchers stated that their research study was the first one to predict seasonal influenza which was carried out in real-time and which demonstrated accuracy of the forecasts [129]. Additionally, through some form of forecast, the researchers forecasted the seasonal influenza in a number of cities with an accuracy of 63%. Other estimates of accuracy in forecasting were established by Yang et al. (2015) who found that at 1 to 3 weeks lead time (how far in the future the peak is forecasted), the accuracy was 37%, and increased to 50% at 0 weeks lead time [38]. This paper used the SIR model together with the particle filter algorithm to predict future data points. Yang et al. (2014) compared filtering methods to forecast influenza epidemics retrospectively [40] and found that different filtering methods overestimated the outbreak's size when the forecasting was carried out close to the observed peak.

Through the use of different sources of information, historical data, models and methods, one can try to improve forecasting techniques [130]. Researchers attempt to use a combination of statistical, simulation and optimization techniques to forecast an epidemic curve [131] through the use of the previous parameter values of past epidemics. Combination of different methods can improve influenza forecasts and can prove to be the way forward in real-time forecasting, given the right assumptions together with good datasets [131]. If a proactive approach and model adequate strategies are to be adopted, the right practices are necessary to be implemented to forecast the influenza [47]. The right methodologies with direct comparisons of independent data, as well as sensitivity analysis, are of paramount importance to ensure that the proposed models are providing tangible results that can be used by mathematical experts and public health officials [47].

In the above research papers, epidemiologists attempted to predict the epidemic trend, duration, peak timing, peak height, and the size of the outbreak. Throughout this dissertation, I in turn, attempt to predict all these factors through the use of a combination of some of the above models and other statistical techniques. In comparison to the current research, the model and methodology developed in this dissertation are novel, and aim to produce real-time forecasting. Additionally, I examined thoroughly the assumptions about the initial values of $S(0)$, $E(0)$, $I(0)$ and $R(0)$ in the epidemiological models in this thesis, as these values can be the key factors in forecasting the characteristics of current outbreaks [132]. The ultimate question remains as to how early we can predict the progress of an epidemic based on limited information [33], which is one of the main research questions of this dissertation.

1.6 Thesis Overview

The thesis takes the form of nine chapters. The first chapter provided an introduction to this research study by introducing the main objectives of this thesis. This is followed by a thorough narrative literature review of up-to-date findings of various studies. Chapter 2 introduces the main methods used in this thesis. I will be explaining some basic and important information regarding the Malta context, and general information about Malta's health care system. All data used throughout this thesis, together with models and statistical tests will be described in detail in chapter 2.

Chapter 3 focuses on the H1N1 Influenza outbreak. I acquired a dataset on the Maltese population regarding the outbreak of H1N1 influenza during 2009 and 2010. All data collection was performed by the Maltese Health Authorities, led by the Malta Health Promotion Department (MHPD); my contribution is to provide statistical analysis and modelling. This research study describes four datasets (consultations, diagnosed, swabbed and positives), all of which will be used for epidemiological modelling. The novel part of this research is that the approach explicitly addresses multi-proxy signals and compares parameter estimates across different proxies. Additionally, several relationships between the different proxies is examined in detail, including their time-dependence. Chapter 3 was published in *Epidemics* in December 2014 [79].

The analysis in chapter 3 is extended in chapter 4 through the analysis of four seasonal influenza datasets. Several comparisons between different parameters for four different seasonal influenza datasets are carried out, including the effective reproduction ratio R_t. The analysis in chapter 4 is complemented by a detailed analysis to understand the relationship between diagnosed cases and consultation cases.

In chapter 4, I introduce and analyse a model that is able to combine multiple datasets together. The main aim is to incorporate different datasets together to refine the prediction of the outbreak and at the same time, predict multiple datasets in one single framework. I use this model to study the effectiveness of real-time forecasting, using a certain number of known time points (Chapter 5). I apply the above techniques to forecast the subsequent weeks of data. For the above analysis, I apply the SEIR model and Bayesian techniques for parameters estimation (particle filtering algorithm), which are implemented through the use of a statistical package 'R' [133]. The R particle filtering code is a modification of Professor Alex Cook's code and I used this code with the author's permission.

Throughout chapter 6, several model parameters are analysed to better understand the sensitivity of the results to changes in priors. The initial number of susceptible and infected individuals are not known. In this chapter, I explore sensitivity of the results (including R_t) to changes in the assumptions about $S(0)$, $E(0)$, $I(0)$ and $R(0)$. These are analysed in relation to the effective reproduction ratio. At the end of this chapter, a method is presented in relation to the sensitivity analysis.

The objective of chapter 7 is to understand several underlying factors related to the influenza, including the under-reporting rate of the seasonal influenza in Malta. Two cross-sectional surveys were performed to address several important factors related to the Maltese population, and to compare the survey results with the GPs reported data. In addition, throughout this research we aim to understand the most important symptoms related to the seasonal influenza in Malta. I examine the level of occurrences for such symptoms, the hospitalisation rates due to the seasonal influenza, consultations to GPs and other important medical information related to the seasonal influenza. This research can be considered innovative in the local context as it is a first study of its kind in Malta. On an international level, limited research also exists in the adoption of nation-wide

cross-sectional surveys to study factors related to seasonal influenza and to estimate the under-reporting rates.

Chapter 8 aims to combine all the above techniques in one new innovative prediction model to predict the outbreak at an early stage; hence, a new model and framework are developed. I analyse the temperature data in relation to the diagnosed data and the effective reproduction ratio, and compare the new model with the current up-to-date techniques used internationally. Finally, I use this method to predict the total number of infected individuals until the end of the season, the peak of the influenza season and the influenza spread throughout all weeks.

Chapter 9 contains discussion in which the methodology and results from chapters 3 to 8 are brought together. I also provide some directions for future research.

Chapter 2

Materials and Methods

2.1 Introduction

The main aim of this chapter is to define important materials and methods used in this dissertation. This chapter provides key information on Malta, its healthcare system and the role of several health departments in Malta. Subsequently, the chain of events surrounding an influenza infection is mapped. The methodology and datasets are defined with clear distinctions between different datasets. The SEIR model, the effective reproduction ratio, the particle filter algorithm, linear regression model and other statistical tests are all described in detail. This chapter is concluded with brief information of the software used throughout this thesis.

2.2 A brief description of Malta

Malta has generally a high humidity level, with an average of 74% during July'09 – June'10 period (Figure 2.1). The humidity level in Malta is relatively stationary; however the lowest levels are reached during the July-August period. This period is also associated with the highest average temperatures in Malta. During this timeframe, temperature exceeds the 30 degrees Celsius on average during the whole day, though reaching temperatures of 40 degrees Celsius during mid-day. Malta's average temperature is 19 degrees Celsius, with lowest average temperatures reached during the December–February period. Although there is no clear peak for the humidity level, the months March-April tend to show higher humidity levels in Malta. Figure 2.1 shows data for 2009-2010 which are typical of the Maltese weather.

The Maltese population enjoys 13 public holidays; during these days, most employees are off from work and all schools are closed. The school-holiday periods are represented in figure 2.1 through the shaded areas. Such data is important as it is believed that it is directly correlated with the spread of influenza (Maltese Department of Health Information and Research, 2015) (Appendix A). Malta registers high seasonal influenza spread following the Christmas period and as soon as schools commence. Maltese students enjoy a long holiday period during the Summer season, i.e. between the beginning of July and the third week of September. A mid-semester break follows, consisting of around 3 days in most schools during the beginning of November. Subsequently, between the end of the 3rd week of December and 1st week of January, there is the Christmas holiday break, followed by another semester break of 3 days during

the last week of February, and an Easter break of around 10 days during the first week of April.

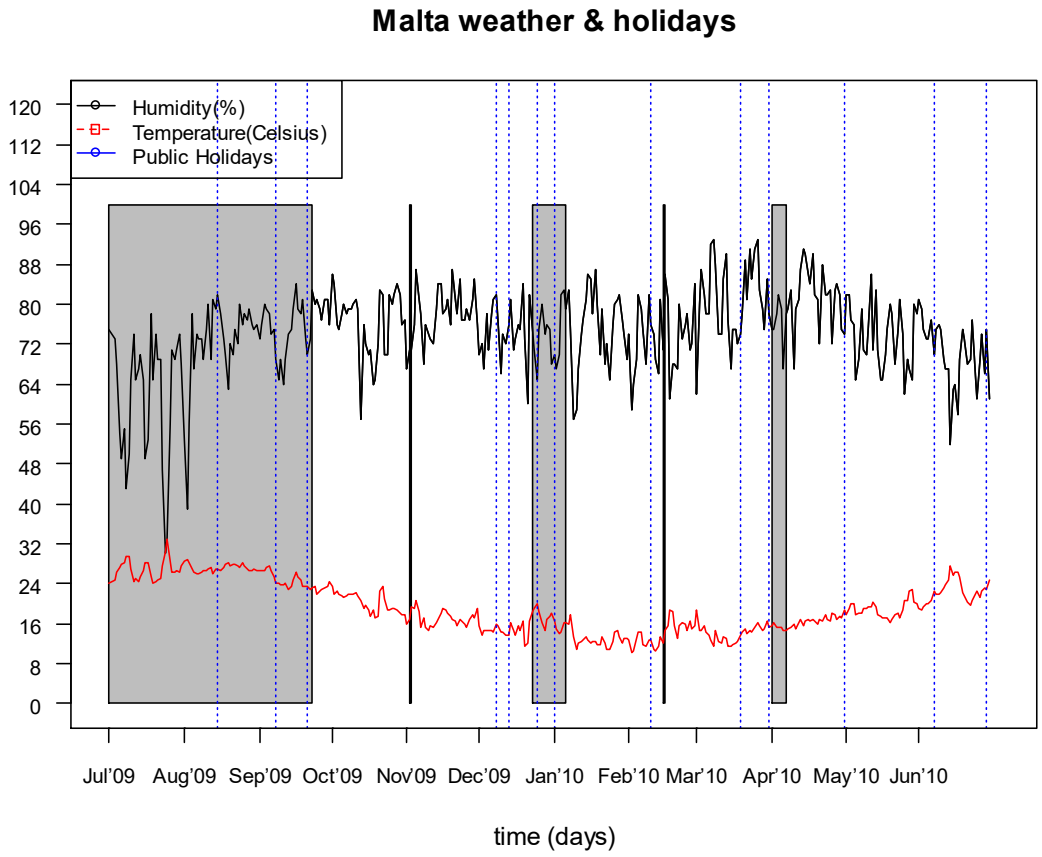


Figure 2.1 – The Maltese weather and holiday characteristics. The black line represents the humidity level for a whole year, the red line represents the Malta's typical temperature in Celsius throughout a whole year, the blue lines are the public holidays in Malta and the shaded grey areas are the school holidays.

2.3 Malta's healthcare system

Malta has a long-standing medical history of healthcare provision since 1372, when its first hospital began to function [134]. During World War I, the island earned worldwide reputation for the nursing care it offered to inpatients. In 2000, Malta ranked fifth in the World Health Organization's ranking of the world's health systems [135], superseding the United States (37th), Sweden (23rd), United Kingdom (18th) and Spain (7th). To date, the Maltese government provides comprehensive, publicly funded health care to all Maltese residents, similar to the British system [136]. It operates through public hospitals and health care centres, and is overseen by the Ministry for Health. Health care is funded through taxation and national insurance, covering a wide array of treatments, namely covering most medical services such as specialist treatment, hospitalisation,

prescriptions, pregnancy, childbirth and rehabilitation, amongst others. Individuals with lower income receive free pharmaceuticals following means-testing. Primary healthcare is provided through eight Health Centres: seven in Malta and one in Gozo, offering preventive, curative and rehabilitative services. Secondary and tertiary care is provided through public hospitals. The primary hospital in Malta is the Mater Dei Hospital which was inaugurated in 2007 as one of the largest medical buildings in Europe. It received a number of awards for medical excellence and research. For those who opt for private health care insurance or out-of-pocket payments, the island also offers a strong private health system [137]. Pharmacies across the island also offer services by General Practitioners, specialized doctors as well as allied health care professionals. Voluntary organisations, such as St. John Ambulance and Red Cross Malta, provide first aid/nursing services. Similarly, foreign residents are offered health care services through their private medical insurance [138]. The University of Malta has a medical school and a Faculty of Health Sciences which train students towards their undergraduate or postgraduate studies.

2.3.1 The role of the research department

The collection, analysis and delivery of health related information in Malta is led by the Directorate for Health Information and Research through the provision of high quality epidemiological indicators on the health of the Maltese population and local health services. The Directorate gathers, analyses and disseminates health information by conducting epidemiological studies and maintains disease registers. The Directorate is also responsible for the management of national health datasets on mortality, cancer, congenital anomalies, organ transplant, obstetrics, hospitals information system, accidents and injuries, as well as for a number of other databases on health service activity. This directorate is responsible for carrying out the Health Interview Surveys, such as the First National Health Interview Survey in 2002, and the European Health Interview Survey [139].

2.3.2 The role of the Malta Health Promotion department

Health promotion is the process of enabling individuals to increase control over the determinants of health, thereby improving their general health [140]. Health promotion not only embraces actions directed at strengthening individuals' skills and capabilities, but also increases actions directed towards changing social, environmental and economic

conditions. This, in turn, alleviates their impact on public health, enabling individuals to enjoy healthier lifestyles. The Health Promotion Unit was set up with the aim to support individuals in controlling their own health by investing in sustainable policies, actions and infrastructure to address the determinants of health. Apart from leading weight management classes, smoking cessation programs within Primary Health Centres, self-management programs and aerobics classes, particular attention is given to infectious disease prevention. The Infectious Disease Prevention and Control Unit, under the auspices of the Health Promotion Department, is the only centre in Malta that deals with surveillance of infectious diseases. Data is collected from various sources, namely medical doctors, laboratories and through local surveillance systems to provide information on prevailing issues in infectious diseases. The unit is also responsible to manage outbreaks of infectious diseases and to provide related data to the local and international scientific community.

2.3.3 Influenza vaccination in Malta

Routine annual influenza vaccination is offered free of charge to all healthcare professionals, other staff working with patients, employees working within the police force, soldiers, civil protection personnel, staff at detention centres and open centres, veterinary personnel, abattoir personnel, cleansing department staff, correctional facility staff and inmates, persons residing in institutions, students attending special schools, persons aged 55 years and over, children from the age of 6 months to 59 months and persons of any age suffering from chronic diseases (lungs, heart, liver, kidney, diabetes mellitus, and any immunodeficiency conditions, including HIV and AIDS). All other individuals need to call at their GP to receive the seasonal influenza vaccine. Health care providers offer vaccination in October and November at the healthcare centres, councils, family doctors, homes, institutes and hospitals. Most local councils and a number of family doctors participate regularly in yearly vaccination campaigns to promote influenza vaccination and raise awareness in the community [141].

2.4 Key definitions

2.4.1 Pathways through influenza illness

An individual faced with a disease may choose to follow different pathways throughout the course of the illness (Figure 2.2). Primarily it is best to define the target population.

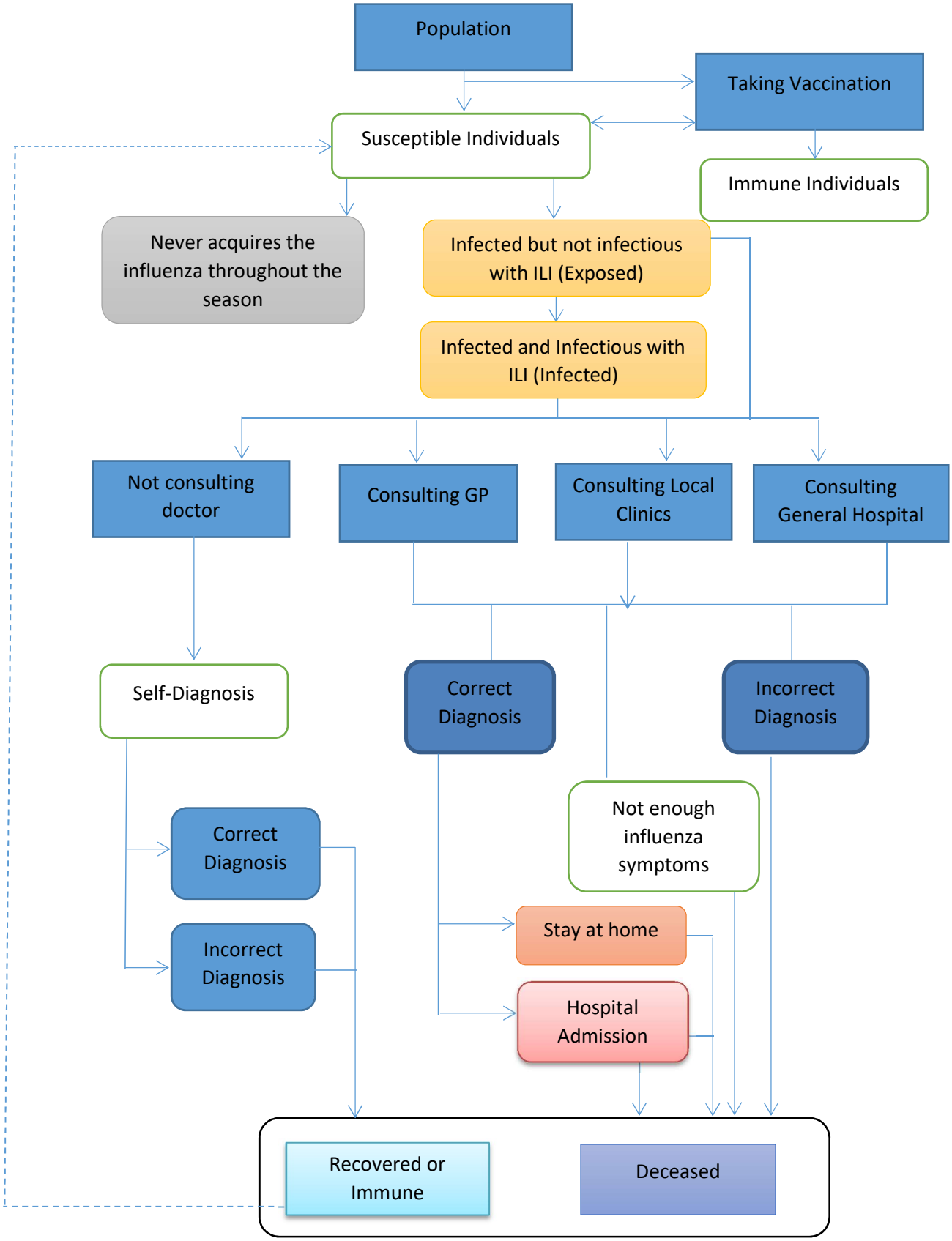
This includes all those individuals who are part of a country/area under study. During the influenza season, several individuals might decide to take the vaccine for influenza protection. Some vaccinated individuals might develop immunity but others might still develop the seasonal influenza later on during the year (maybe due to a lack of immunity or lack of response to the vaccine). Hence, after excluding immune individuals, the new sub-population becomes known as the group of susceptible individuals.

Figure 2.2 presents the different pathways taken by different individuals following their infection, as not everyone reacts to the same illness in the same way. Some might be very wary about their illness while others may feel that they can deal with it on their own.

An individual from the susceptible group and with symptoms related to an influenza-like illness may either consult a general practitioner (GP), the local clinic, the general hospital, or may decide not to consult anyone. If the latter option is selected, the individual might undertake a self-diagnosis with the risk of carrying out an incorrect diagnosis about the illness. If any of the first three options is selected, the doctors might correctly diagnose the patient for seasonal influenza or incorrectly not diagnose the patient for seasonal influenza. There is the possibility that the doctor concludes that there are not enough symptoms to diagnose the patient as positive to seasonal influenza. If one is diagnosed as positive to seasonal influenza, the patient may be admitted to hospital or sent home for the recovery period. This can either lead to a patient's full recovery or the patient is deceased. Hence, an individual has several options to consider when feeling unwell. Furthermore, an individual most likely will acquire immunity if recovered from influenza or if vaccinated. Most epidemiological studies aim to predict the total number of positively diagnosed individuals, irrespective of their preferred pathway.

In most cases in Malta, if a person is diagnosed with seasonal influenza by the GP, the diagnosis is not based on a blood test but on the GP's professional judgement. Hence, one may conclude that the patient is tested positive to an influenza-like illness or is positive to seasonal influenza but without the confirmation of a virological test based upon a nasal swab.

Figure 2.2 – This figure maps the whole process of the Influenza and all the potential pathways that several groups of individuals may experience during the seasonal influenza period. Individuals are faced with several possibilities and options throughout the whole period.



2.5 Data used in the thesis

This section presents several datasets related to the influenza in Malta. The first dataset concerns the H1N1 pandemic season (2009-2010). This includes the number of consultation cases resulting from the H1N1 pandemic season, data related to the number of people who were diagnosed positive to influenza-like illness, those who were swabbed during the H1N1 period and the number of individuals who tested positive to H1N1 through laboratory tests. Subsequently, data related to the seasonal influenza for four consecutive seasons (2011/2012, 2012/2013, 2013/2014 and 2014/2015) is presented. For each respective season, the data include three variables, the number of doctors reporting the cases, the number of consultations and the number of diagnosed individuals with influenza-like illness. Furthermore, a 2015/2016 seasonal influenza dataset will be only mentioned and used towards the end of this thesis. All data collection was performed by the Maltese Health Authorities and led by the Malta Health Promotion Department (MHPD). Additionally in this section, there is a description about the methodology and data obtained from a cross-sectional survey, and Malta's temperature data.

2.5.1 Influenza data

2.5.1.1 Doctors' consultations and diagnosed cases

When a patient feels ill, the first stage of the patient pathway is typically a consultation with a doctor. This is then followed by a diagnosis of the influenza or of any other illness. The MHPD collects the number of consultations and diagnosed Influenza-Like Illness (ILI) cases on a yearly basis (Figure 2.3) during every season related to influenza (October – May period). Both the consultations and the diagnosed data are collected on a weekly basis (Monday to Sunday). During the H1N1 2009/2010 pandemic season (Chapter 3), eight general practitioners (GPs) were selected (from around 300 GPs) to report the number of consultations and diagnosed ILI cases.

For the scope of chapter 4, four consecutive influenza seasons were analysed, spanning 2011 and 2015. Usually the data collection begins at the end of October till around mid-May. On average, the number of GPs submitting their weekly number of consultations and diagnosed cases varied between 6 and 7 GPs per week during the four seasons. All the seasonal influenza datasets include the number of GPs submitting their reports on a weekly basis.

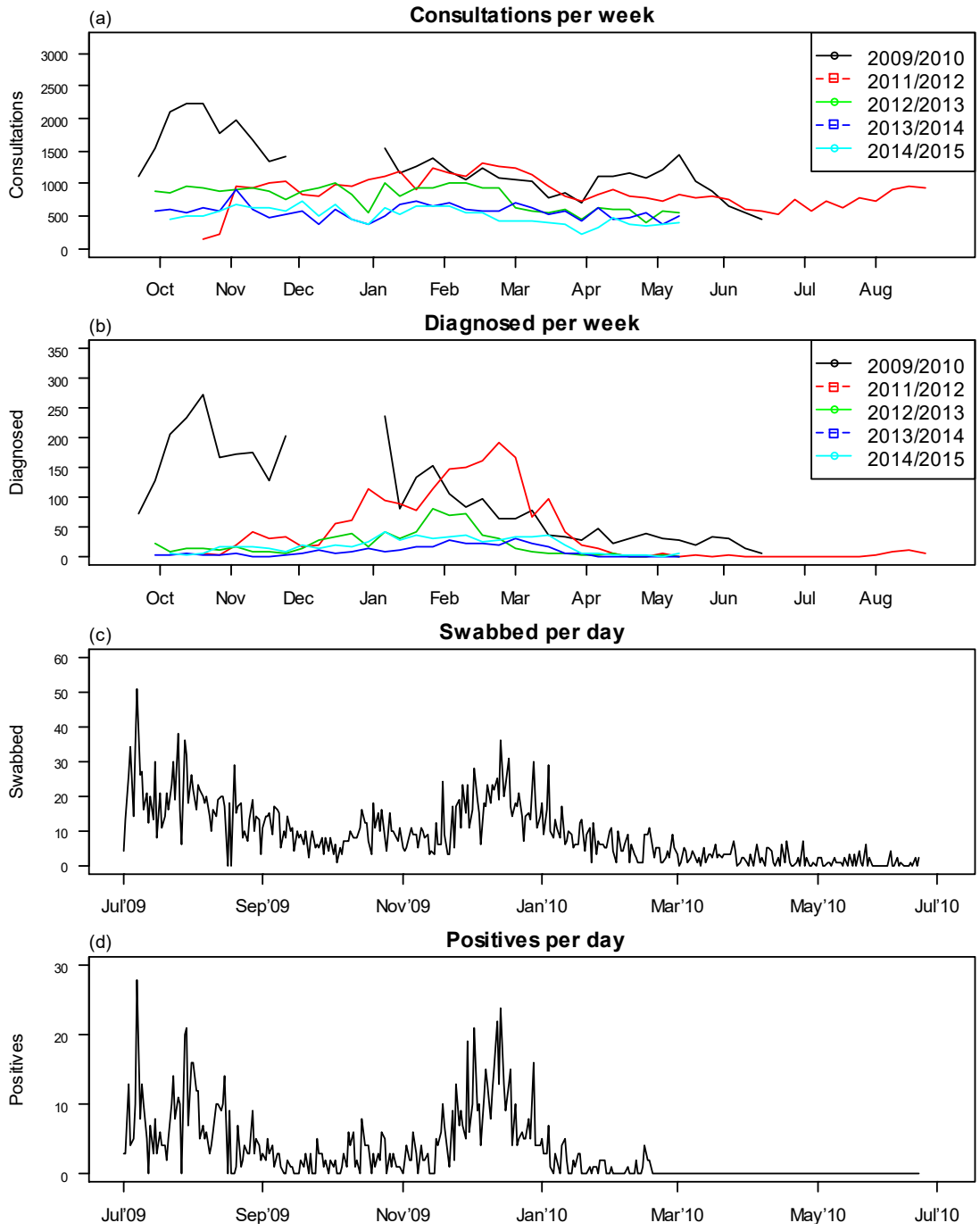


Figure 2.3 – All the original data as collected by the Malta health promotion department (MHPD). The first two charts ((a) and (b)) represent the weekly consultations and diagnosed ILIs by a selected number of GPs. No data were collected between week 49 (2099) and week 1 (2010) for the consultation and diagnosed datasets. The last two charts ((c) and (d)) represent the daily swabbed and positive cases during the pandemic season. Note that for the last two charts, all the GPs in Malta were invited to participate.

The consultation data include both influenza and non-influenza related data. Hence, the number of reported consultations include any consultation irrespective of the type of illness, medical condition, or any other request raised by the patient. Hence, the consultations data include a portion of patients that were tested for the influenza and

another portion of patients who were tested for other symptoms that were unrelated to the influenza. A number of patients were examined for ILI; these were either diagnosed positive or negative. Those who were tested positive (i.e. acute illness with onset during the last 7 days, with measured temperatures of >38ºC together with others symptoms as defined in Chapter 1) by the GPs are represented through all the diagnosed datasets. Some missing data exists in the consultations and diagnosed data (Figure 2.3) due to non-collected data during some periods. This missing data was imputed through the linear regression model which is later described in Chapter 3 [79].

The above datasets (Figure 2.3) are the main influenza datasets that I will be exploring; however a similar dataset was acquired for the 2015/2016 seasonal influenza. This dataset will only be used in chapter 7 to compare the 2015/2016 dataset with the survey data, and in chapter 8 to test the methodology developed in that same chapter.

2.5.1.2 ILI Swabbed and H1N1 Positive cases

During the H1N1 season (2009/2010), all GPs in Malta who had seen and diagnosed individuals with ILI, were encouraged to contact the MHPD to have their patients swabbed (Figure 2.3). Only these individuals, as well as those who were considered part of the high risk group, were eligible to be swabbed. On average, there were 8.5 GPs reporting cases on a daily basis. These GPs might differ from one day to another, as all GPs in Malta were invited to follow this process. As defined in chapter 3 [79], the high risk group includes: elderly, pregnant women, children under the age of 5, patients with a chronic disease and health care workers. Hence, swabbing patients includes further investigations in a laboratory, rather than the standard tests (such as checking patients' temperature) carried out by the physicians to examine patients for ILI. In total, 3204 people were swabbed by the MPHD between 1st July 2009 and 20th June 2010. Of these, 1100 tested positive to H1N1 (Figure 2.3). These were the only laboratory-confirmed H1N1 cases in Malta and include both hospitalised cases and cases in the community. However, one cannot assume that those who tested negative did not develop the H1N1 virus during the season, since the influenza during this period consisted of the H1N1 type virus. There are several reasons for this, which are discussed in further detail in chapter 3 [79].

For both swabbed and positive datasets, there were two main waves reaching their peaks in July 2009 and December 2009. The second peak resulted in a lower number of swabbed and infected cases when compared to the first peak. Swabbed and positive datasets commenced with a peak value without any build-up to reach the peak value of the influenza during the first wave. Hence, one may hypothesize that there is some missing data for the period prior to 1st July 2009. Although research has been carried out in other countries (as described in further detail in Chapter 3), there is a gap in our understanding of the epidemiological factors related to the Maltese population. In addition, determining innovative techniques may help explain any gaps in knowledge or misconceptions about seasonal influenza.

2.5.2 Malta's cross-sectional survey datasets

Two cross-sectional surveys were carried out as part of this thesis; however the data obtained from the first survey was the primary dataset that is presented and analysed in chapter 7. The first survey was carried out between week 35 (August 2015) and week 37 (September 2015), and its primary aim was to explore the under-reporting rate (defined in Chapter 1) of the seasonal influenza as compared with the above GP datasets during the 2014/2015 seasonal influenza. As defined above (Figure 2.2), this might be derived due to several reasons (self-diagnosis, not-enough symptoms to diagnose ILI, and incorrect diagnosis). In these surveys respondents were asked questions retrospectively for the previous year.

In this study, a questionnaire was designed to explore several characteristics related to the seasonal influenza, influenza-like illness, symptoms related to influenza and other medical topics (Appendix B). The research instrument consisted of 32 items, including socio-demographic factors, and other questions related to whether participants had experienced the seasonal influenza and whether they had any particular symptoms. Furthermore, respondents were given a list of symptoms to evaluate whether they actually had experienced these symptoms during the past year.

A similar second survey was carried out in April 2016. The data obtained in the second survey was used to compare and confirm some of the results obtained in the previous survey and hence, the results were not analysed to the same extent as in the previous

survey. This survey was carried out during the end stages of the 2015/2016 seasonal influenza. Hence, results were also compared (with the first survey) from the perspective that the second survey was carried out earlier (throughout the influenza season) when it might be easier for respondents to remember their ILI symptoms. It is important to emphasize that the second survey data can be considered as a secondary dataset in chapter 7, and mainly serves as a tool to analyse and compare the 2014/2015 survey dataset and the main objectives of chapter 7.

A pilot study was conducted with a small random sample of 20 individuals to ensure that all questions are understandable and to ascertain the practicalities of conducting the telephone survey. The results showed that the tool was feasible to conduct by telephone and that no changes were required. The individuals participating in the pilot study were not included in the larger study.

To ensure a good response rate, the study was carried out through the use of telephone interviews. The interviews were conducted in Maltese; however if participants preferred to answer in English, this option was offered. Each survey comprised a sample of 406 Maltese individuals from the eligible population of around 349,724 individuals [16]. In this study, the eligibility criteria to participate in this study was all Maltese residents of 18 years and older, and people residing in Malta. The study was carried out through a 95% confidence level and 4.86% confidence interval as shown below (Figure 2.4). The sample was stratified by age, district and gender. Telephone numbers were obtained from the two main telephone service providers in Malta (GO and Melita) and generated at random.

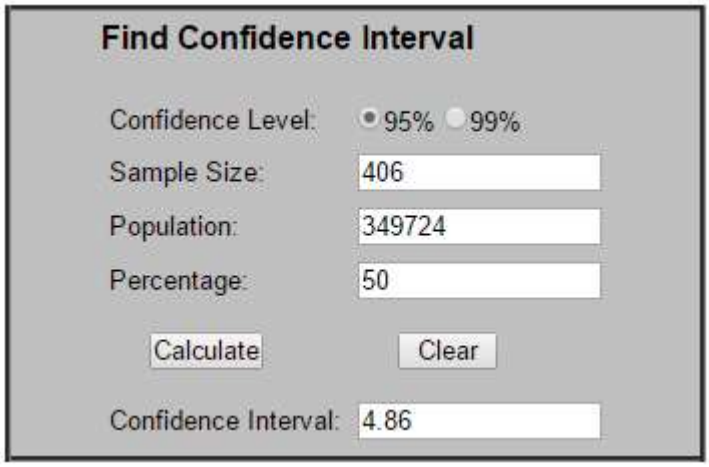


Figure 2.4 – Sample Size (Creative Research Systems, 2012) [142]

2.5.3 Temperature data

Malta's temperature data was obtained from the Maltese Meteorological Office [143]. This office is part of the private company, Malta International Airport. The Meteorological office offers an extensive range of products related to Malta's weather, including temperature data, humidity levels, wind speed and wind direction. These data can be provided at various locations around Malta. The Meteorological office collects this data every minute, and every day of the year [143]. For the scope of this thesis, the daily temperature data since 2009 was obtained. The data acquired was for Luqa, which is located centrally. Most on-line weather reports and weather forecasts also use this particular location. Since most of the above influenza datasets are in weekly format, weekly averages were calculated in order to compare the temperature data against the weekly diagnosed data and other variables.

2.6 Models

Throughout this section, we shall cover the most important mathematical modelling techniques, algorithms and statistical techniques used in this thesis. The SEIR model was the main modelling technique used. Parameters were estimated through the use of the particle filter algorithm. Throughout different parts of the thesis, the linear modelling technique and some other standard statistical tests were used to analyse different variables.

2.6.1 The SEIR model

The SEIR model [26] is the epidemiological model used throughout this thesis. The model includes four different compartments (Figure 2.5). The first stage of the SEIR model are Susceptible (S) individuals, i.e. those who can acquire the disease. Following this stage, individuals move to the Exposed (E) class (but cannot transmit the disease), followed by those who are Infected (I) and able to transmit the disease to other individuals. The last compartment is the Removed (R) individuals, which includes those individuals who become immune (recovered) or deceased due to the disease.

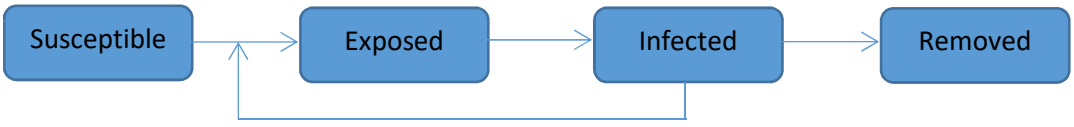


Figure 2.5 – The four different compartments of the SEIR model. An individual moves from one stage to another with the possibility of not being fully recovered and hence moves to the initial stage.

For the purpose of parameter estimation and prediction, we use the following set of equations [26, 79]:

$$\begin{aligned} S_t &= S_{t-1} - A_t \\ E_t &= E_{t-1} + A_t - B_t \\ I_t &= I_{t-1} + B_t - C_t \\ R_t &= R_{t-1} + C_t \end{aligned} \tag{1}$$

where S_t is the number of susceptible individuals at time *t* (day or week), E_t is the number of exposed (but not infectious) individuals at time *t* (day or week), I_t is the number of infected (and infectious) individuals at time *t* (day or week) and R_t is the number of removed individuals at time *t* (day or week). The values A_t, B_t and C_t are the numbers of newly infected individuals in the population (i.e. individuals from the susceptible compartment who are then moved to the exposed compartment), the number of infectious individuals (i.e. individuals from the exposed compartment who are moved to the infectious compartment) and the removed persons respectively (i.e. individuals from the infectious compartment who are moved to the removed compartment). These variables are assumed to have the binomial distribution and are defined by:

$$\begin{aligned} A_t &\sim Bin\left(S_{t-1}, 1 - e^{\left\{\frac{-[\varepsilon+\beta I_{t-1}]}{N}\right\}}\right) \\ B_t &\sim Bin\left(E_{t-1}, 1 - e^{\left\{\frac{-1}{\alpha}\right\}}\right) \\ C_t &\sim Bin\left(I_{t-1}, 1 - e^{\left\{\frac{-1}{\tau}\right\}}\right) \end{aligned} \tag{2}$$

Here, ε is the importation rate per week or per day (according to the dataset). This includes Maltese individuals who become infected due to travelling abroad, but does not include new travellers entering Malta. The parameter β is the infection rate of the Maltese population, α^{-1} is the transition rate between exposed to infectious, and τ^{-1} is the transition rate from infectious to the removed compartment. Hence, α is the latent period in days or in weeks (according to the dataset) that an individual takes to move from the exposed compartment (E) to the infectious compartment (I), while τ is the infectious period (in days or weeks) that a person takes to shift from the infections compartment (I) to the removed compartment (R). *N* is the population size of Malta which is assumed to be equal to 414,000.

These four different compartments (S_t, E_t, I_t, R_t) are not observable. Hence, through the collected data we aim to estimate these four compartments based on the above parameter values $(\beta, \varepsilon, \alpha, \tau)$. What is observable, is the dataset D_t. In fact, the above SEIR model is combined with an observation (reporting) model. The observations D_t are the actual number of cases (consultations, diagnosed, swabbed and positive) as reported by GPs. Through the observation model, we combine the number of infectious individuals (I), the background rate (ϕ) and the reporting rate (δ). D_t is assumed to be Poisson distributed with mean $N_t \delta_{d(t)} \left\{\phi + \frac{I_t}{300}\right\}$ where N_t is the total number of GPs submitting reports on day/week *t*. $\delta_{d(t)}$ is the probability of infected individual seeking medical help, where $d(t)$ is the day of the week for daily datasets, while for weekly data only one δ is used. The number of practicing GPs in Malta was estimated to be equal to around 300, as stated by the directorate for Health Information and Research during one of my one-to-one meetings. The SEIR model predicts the total number of infectious individuals from the whole population (414,000), while D_t predicts the total number of cases as reported by GPs. D_t can be directly compared with collected data.

The 'background' consulting rate (ϕ) for the consultations data is the number of non-influenza cases from the total number of consultations being reported by the doctors. For the diagnosed cases, ϕ is the number of non-ILI cases from the total number of ILIs being reported by GPs. For the swabbed/positive datasets, this is the number of non-H1N1 ILI/positive cases. Hence, for the consultation datasets, ϕ is expected to be higher, as substantial number of consultations cases are not related to the influenza. On the other hand, the diagnosed datasets are a more direct measure of the number of infectious (I), resulting in a lower number of 'background' consulting rate. Hence, the SEIR model tries to establish the actual number of individuals for different compartments based on the fact that there is a certain level of 'background' consulting rate and the reporting rate.

2.6.2 R_t for different datasets

As mentioned in chapter 1, the reproduction ratio is one of the most important parameters in epidemiological modelling. This is defined as the number of new infected individuals from one currently infected person at a given time. The effective reproduction ratio R_t is calculated from the above SEIR model. In fact, once the parameters $\beta, \varepsilon, \alpha$ and τ are

computed, the effective reproduction ratio (R_t) at any given time t is calculated through the following equation:

$$\frac{\beta(1 - e^{-\left(\frac{1}{\tau}\right)})S_t}{N}$$

For any dataset (consultation, diagnosed, swabbed and positive), the R_t value has the same meaning but is based on a different proxy. The S, E, I and R compartments are 'true' numbers and are not subject to interpretation. It is D_t that varies. Hence, the effective reproduction ratio has a consistent meaning for the SEIR model related to the number of new influenza infections. Thus, for all different datasets the R_t value is the number of newly infected cases produced by a single currently infected individual. The main difference is the level of uncertainty that each dataset (D_t) provides to the effective reproduction ratio. Hence, if the consultation dataset includes substantial amount of non-influenza cases, then this dataset includes a considerable level of uncertainty about the actual number of infected individuals. In contrast, the diagnosed dataset is specifically related to individuals with an influenza-like illness. Thus, this is a more direct proxy to the number of infectious individuals' compartment. The same applies for the swabbed and positive cases (both direct proxy of the infectious individuals). Therefore, the reliability of the estimated R_t value depends on the type of proxy being used.

2.6.3 Particle filtering algorithm

The particle filter algorithm (as defined in the previous chapter) is a sequential Monte Carlo algorithm [26]. It is a sampling method to approximate a distribution that makes use of its temporal structure [26]. The idea in this study is to represent the posterior density by a set of random particles with associated weights. The estimates are then computed based on these samples and weights. As defined in previous section, the SEIR model is based on a set of parameters $\theta = (\beta, \varepsilon, \alpha, \tau)$ and the unknown unobserved state $\Sigma_t = \{S_t, E_t, I_t, R_t\}$. Hence one can estimate these parameters and values through the use of the above SEIR model using the particle filter algorithm.

2.6.3.1 Initial stage

The algorithm starts at time $t=0$, and with a set of 10,000 generated particles (P) (or even more) from the prior distribution for the initial states Σ_0 and parameters θ.

2.6.3.2 Iteration of particles

For each particle, p at each time step $t+1$, Σ_{t+1} is drawn using Monte Carlo simulation from its conditional distribution, given x_t^p, where $x_t^p = (\Sigma_t, \theta)$ with an associated weight w_t^p [26]. At each time point, each prediction is calculated in light of what has already been discovered. Hence, the particles are being iterated by one time point at a time based on the new state space (Σ_t).

2.6.3.3 Weighting the particles

The likelihood function is estimated conditioned on the pathway of the particle and the associated parameter values. Hence, we use the likelihood function to weight each individual particle. Therefore, we set $x_{t+1}^p = (\Sigma_{t+1}, \theta)$ and the likelihood contribution $L_{t+1}^p = f(D_{t+1}|x_{t+1}^p)$ is calculated, conditioned on the path of the respective particle using the same parameter values and D_t, which is the number of reported cases on day t (for daily dataset) or week t (for weekly dataset). This likelihood is then used to find the weights by setting $w_{t+1}^{*(p)} = w_t^{(p)} L_{t+1}^{(p)}$ and then scaled to sum up to one: $w_{t+1}^{*(p)} = \frac{w_{t+1}^{*(p)}}{\sum_{q=1}^{P} w_{t+1}^{*(q)}}$.

2.6.3.4 Particle degeneracy and re-sampling

One of the problems when using the algorithm of particle filtering is that some of the particles will be assigned low values of weights; hence their relevance for the distribution is almost negligible. This problem is overcome by performing re-sampling [46], hence letting $x_{t+1}^{*(p)} = x_{t+1}^{*(q)}$ where q is selected from the set of integers $\{1,2,\dots,P\}$ with probability proportional to $w_{t+1}^{*(q)}$. Then for all p, $w_{t+1}^p = \frac{1}{P}$. Thus, whenever some of the particles fall below a certain threshold, the current set of particles are re-sampled.

2.6.3.5 Kernel smoothing

Particle diversity is retained by kernel smoothing [144]. As described in Ong et al. (2010) [26] and Trenkel et al. (2000) [14], let $x_{t+1}^p = \mu_{t+1}^p + h(x_{t+1}^{*(p)} - \mu_{t+1}^p) + Z\sqrt{1-h^2}$, and setting h=0.3, Z is generated from a multivariate Gaussian distribution with mean vector

0 and the variance derived from the variance-covariance matrix of $x_{t+1}^{*(p)}$ over all p, and μ_{t+1}^{p} is the vector of means $x_{t+1}^{*(p)}$ over all p if the estimated value x_{t+1}^{p} falls within the correct support or $x_{t+1}^{p} = x_{t+1}^{*(p)}$ otherwise. Hence, kernel smoothing is used to improve the precision and robustness of the parameter estimates [145]. The main strength of this step is that it solves the problem of particle failures (by retaining a good particle mixture) without the side- effects of increasing the variance.

2.6.3.6 Increment

This algorithm is repeated and Σ_{t+1} is observed again. This can be run in two different ways. For parameter estimation, we will run it up to the end of the observed data. For prediction, we will run it until the end of the prediction period.

2.6.4 Linear Regression Model

Linear regression modelling demonstrates the relationship between selected values of X and observed values of Y, from which the most probable value of Y can be predicted for any value of X. Hence, regression tries to find the line of best fit that predicts variable Y. A linear regression technique gives an understanding of the relationship between two variables. This technique establishes a linear regression equation:

$$Y_i = \kappa + \Delta X_i + \omega_i$$

where Y_i is the dependent variable or response variable for observation i, X_i is the independent variable or predictor variable for observation i and Δ is the regression coefficient. The latter is also the gradient/slope of the linear regression. This is one of the most important parameters of the linear regression model, as it defines the main relationship between the dependent and independent variable. In order to calculate the parameter Δ we need to use the least square estimation method to estimate:

$$\hat{\Delta} = \frac{S_{xy}}{S_{xx}}$$

where $S_{xy} = Cov(X,Y) = \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$, $S_{xx} = Var(X) = \sum_{i=1}^{n}(X_i - \bar{X})^2$, $\bar{Y}$ is the mean value of Y, $\bar{X}$ is the mean value of X and n is the sample size.

The parameter κ is the *y*-intercept of the linear regression. This variable captures the other fixed factor that influences the dependent variable. This parameter can be estimated through the least square estimation method:

$$\hat{\kappa} = \bar{Y} - \hat{\Delta}\bar{X}$$

For both parameters $\hat{\Delta}$ and $\hat{\kappa}$, the sum of squares of residuals are minimized. We assume that the random error ω is independent and identically normal distributed with mean '0' and variance σ^2, $\omega \sim N(0, \sigma^2)$.

The linear regression model produces the R^2 value, which is the degree of accuracy that the predictor variable *X* is predicting the response variable *Y*. The closer the values are to 100%, the better is the accuracy in predicting variable *Y*.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where Y_i are the original data values, $\hat{Y}_i$ are the modelled values, $\bar{Y}$ is the mean of the original data and *n* is the sample size.

A *t*-test can also be applied on the slope of the linear regression model to examine whether a linear relationship exists between the *X* and *Y* variables. The hypothesis for such a test can be defined as:

$$H_0: \Delta = 0$$
$$H_1: \Delta \neq 0 \; or \; H_1: \Delta > 0 \; or \; H_1: \Delta < 0$$

where H_0 is not rejected if no relationship exists between the *X* and *Y* variables and H_0 is rejected if a relationship exists between *X* and *Y*; hence a model does exist between these two variables. The *t*-statistic used in this case can be defined by the following equation:

$$t = \frac{\hat{\Delta} - 0}{se\,(\hat{\Delta})}$$

where the standard error can be defined by the following equation from the sample size n:

$$se\left(\hat{\Delta}\right) = \sqrt{\frac{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

2.6.5 Analysis for associations

2.6.5.1 Correlations analysis

Correlation analysis demonstrates the degree to which two quantitative and continuous variables are related. The Pearson correlation coefficient (r) is the measure of the level of accuracy between two variables X and Y. By drawing a scatter plot between the two latter variables, one can understand whether there is linearity between these two variables. If the scatter points between variables X and Y can be represented by a perfect line, then it means that the correlation value is 1 or -1, resulting in a perfect relationship between the two variables. The closer the Pearson correlation values are to 1 or -1, the higher the association between the two variables. If the values are close to 0, then it means that there is no association between the two variables. Positive correlation value represents a positive gradient; hence the higher the values of X, the higher the values of Y. A negative Pearson correlation value means that the higher the values of X, the lower are the values of Y. The Pearson correlation coefficient (r) is defined through the following equation:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{n(\sum_{i=1}^{n} {X_i}^2) - (\sum_{i=1}^{n} X_i)^2}\sqrt{n(\sum_{i=1}^{n} {Y_i}^2) - (\sum_{i=1}^{n} Y_i)^2}}$$

where n is the sample size. A t-test can also be applied to test whether the association between the X and Y variables is statistically significant. In order to test this association, we need to apply the following hypothesis:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0 \; or \; H_1: \rho > 0 \; or \; H_1: \rho < 0$$

where ρ is the population correlation coefficient (unknown). H_0 is not rejected if no relationship exists between the *X* and *Y* variables and we reject H_0 if a relationship exists between *X* and *Y*.

The *t*-statistic test used in this case can be defined by the following equation:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

with n-2 degrees of freedom for the above *t*-statistic.

2.6.5.2 Chi-Squared test

Chi-Square (χ^2) test is another test of association between two variables. However, unlike the Pearson correlation coefficient, Chi-Square test only compares categorical variables. This test compares the observed data against the expected data through a cross-tabulation. If there is a significant difference between the observed and expected data, then we reject the null hypothesis and hence we conclude that the responses within one variable are significantly different when compared to the second variable. In order to apply this test, the following chi-square distribution is used:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

with *n-1* degrees of freedom, O_i are the observed values, E_i are the expected values and *n* is the number of categories.

2.7 Software used

This section shall cover the main software used for the analysis performed throughout this dissertation.

2.7.1 R

In order to carry out the particle filtering algorithm, the statistical software 'R' was used [133]. Furthermore, R was used to visualize the above datasets, to apply several statistical

tests (Correlation analysis and Chi-Square test), to apply the linear regression analysis and to visualize the final outputs from the analysis. Since the commencement of my PhD study, several versions of the software 'R' were used; however, the latest utilized version was 3.2.0.

2.7.2 Microsoft Excel

Throughout the whole process of analysis, Microsoft Excel was used mainly to store all the above data, to carry out some quick analysis and to obtain some initial visualizations of the defined datasets. Ultimately, most of the charts were produced through the software 'R'. Several versions of Microsoft Excel were used; however the latest version was Microsoft Office 2013.

2.7.3 SPSS

SPSS software [146] was mainly used for the analysis of the under-reporting surveys. Hence, descriptive statistics, frequency tables, Chi-Square tests and Correlation analyses were all carried out through SPSS v21.

Chapter 3

Estimation of force of infection based on different epidemiological proxies: 2009/2010 Influenza epidemic in Malta

3.1 Introduction

The following is a research paper published in Epidemics **9** (2014) 52-61, and written by V. Marmara (main author), A. Cook and A. Kleczkowski. I was involved in the whole process of writing this research paper. I carried out all analysis and interpreted the results. The following content is exactly the same text as published in the journal. The reference numbers in this chapter are different than the reference numbers of the dissertation in general, as these are exactly the same as in the published paper.

Abstract

Information about infectious disease outbreaks is often gathered indirectly, from doctor's reports and health board records. It also typically underestimates the actual number of cases, but the relationship between the observed proxies and the numbers that drive the diseases is complicated, nonlinear and potentially time- and state-dependent. We use a combination of data collection from the 2009-2010 H1N1 outbreak in Malta, compartmental modelling and Bayesian inference to explore the effect of using various sources of information (consultations, doctor's diagnose, swabbing and molecular testing) on estimation of the effective basic reproduction ratio, R_t. Different proxies and different sampling rates (daily and weekly) lead to similar behaviour of R_t as the epidemic unfolds, although individual parameters (force of infection, length of latent and infectious period) vary. We also demonstrate that the relationship between different proxies varies as epidemic progresses, with the first period characterised by high ratio of consultations and influenza diagnoses to actual confirmed cases of H1N1. This has important consequences for modelling that is based on reconstructing influenza cases from doctor's reports.

Keywords: epidemiology, compartmental modelling, Bayesian inference, Markov chain methods.

Introduction

On the 1st of July 2009, the Health Authorities in Malta reported the first official case of the swine-origin influenza A (H1N1), but in the world, it was already during April 2009 that the first official cases were confirmed in United States (California) and Mexico [1]. Shortly afterwards the influenza started to spread in the European countries [2]. During

the initial stages of the epidemic the overall spread was similar in Europe but in autumn 2009 the second wave of infection primarily emerged in UK [2]. A lot of uncertainty about this influenza existed especially during the initial stages of the influenza, but the availability of datasets has now made this outbreak an excellent case for developing epidemiological models.

The main role of epidemiological modelling is to estimate the reproduction ratio, R_t of an unfolding epidemic of the infectious disease and to provide recommendations for its treatment. However, even the best models cannot perform their required function if the quality of data used to parameterise them is inadequate. Unfortunately, we are unlikely to ever have a complete data set of disease cases; instead we typically struggle with incomplete data sets using various proxies to estimate the numbers we need. One of the biggest problems in epidemiological parameter estimation is associated with low reporting rates. In fact the World Health Organization (WHO) in 2010 said that the total deaths from H1N1 is unquestionably higher [3, 4] due to a substantial amount of unreported cases. In the USA the reported number of H1N1 cases was "substantially underestimated" when compared with the estimated number of Reed et al. [5]. This happens due to several reasons, but the obvious ones are due to the fact that not all people go to visit their doctor when they fall ill, not all cases are sent to laboratories to be investigated and due to the timing of the specimen taken [5].

Additionally, the reporting efficiency often varies over the period of the epidemic. Thus, people might be reluctant to go and seek the doctor's attention early in the epidemic if they are not aware of the risks. Conversely, once the information about the unfolding outbreak is public, there is likely to be a rush to seek medical assistance. Thus, the relationship between what we observe (reported cases) and what is actually happening in the field is a non-trivial function of time, size of the epidemic and news coverage. As these relationships are complex, there are comparatively few studies that address the influence of choice of proxies and the time-and state-dependent reporting on the parameter estimation for epidemics and in particular on the estimation of the effective reproduction ratio, R_t [6-15]. In order to do so, for the case of the H1N1, several papers considered and compared different datasets coming from different states and countries [1, 2, 16-18].

Parameter estimation for epidemiological models has so far been mostly based upon positive cases of H1N1 (laboratory-tested-positive) [2, 12, 19-21] although some analyzed swabbed cases (Influenza-Like-Illness) [22, 23] and others compared swabbed and positive cases [1, 17]. Many datasets were analysed with resolution varying from weekly reporting [23, 24] to daily datasets [2, 8].

It is therefore very important to look for systems that would allow us to study in detail the relationship between different types of epidemiological data. The outbreak of H1N1 influenza in Malta gives us a unique opportunity to study such a relationship. The Malta Health Promotion Department (MHPD) was collecting various epidemiological data during the 2009/2010 outbreak. In this paper, we use a combination of these data and the Bayesian parameter estimation technique to explore how usage of different information about the epidemic influences our understanding of the disease progress. Our assumption is that health authorities would typically have access to only one of the data types that we include in our study and so would like to know how the estimation would be affected by which type of data is available. Our research will use data describing the number of people visiting their physician based on their symptoms (consultations), data about people that were diagnosed with any influenza (diagnosed), those that were swabbed for H1N1 (swabbed data) and those that were tested positive for H1N1 (positives data). The general idea is to give better understanding to the estimation of the force of infection based on different related sources of data. Furthermore, this analysis includes both daily and weekly data.

Material and methods

All data collection was performed by the Maltese Health Authorities and led by the Malta Health Promotion Department (MHPD). The H1N1 data began to be collected when the first cases emerged in Malta in 2009, but the MHPD also collects data informing about the seasonal influenza. The total population in Malta as end of December 2009 was ca. 414,000. This included the non-resident (tourists) population ranging from ca. 6,000 in December to as much as ca. 50,000 in August. Malta is a densely populated country with circa 1,311 inhabitants per square kilometer.

Doctors' consultations and Diagnosed

The first data set incorporates consultations to the Health Promotion Department between week 1 (1st January) in 2009 and week 21 (28th May) in 2011 (Figs. 1(a) and (b), based upon eight physicians selected by the MHPD to report on a weekly basis). Two types of information were collected, the number of patients who attended the practice with any medical problems (Consulted, see Fig. 1) and the number of those subsequently diagnosed with influenza (Diagnosed, Fig. 1(a)). The diagnosis was based on symptoms (a sudden onset of disease, cough, fever >38^{o}C, muscular pain and/or headache; MHPD, private communication). Unfortunately, no data were collected between week 49 – 2009 and week 53 – 2009. In our paper we concentrate on the period September 2009-June 2010, during which 52,016 patients sought the physician's help and 4,544 patients were diagnosed with influenza by the eight physicians.

Swabbed and H1N1 Positive

The physician's diagnosis typically is not based upon any microbial analysis and therefore is to some extend arbitrary. In order to study the process of reporting in more detail, we include in our analysis the data for individuals who were selected for further testing, based upon their increased risk of complications due to influenza. In the community, general practitioners were able to contact MHPD to have their patients swabbed if they developed flu-like symptoms (temperature of 38^{o}C or higher) and if they fell under one of the following high risk groups: elderly, pregnant women, children under 5 years of age, those with chronic disease and health care workers. These people were more at risk of developing complications and could be offered early treatment with antiviral drugs. On average there were 8.5 doctors sending reports each day. Moreover, all those admitted directly to hospital with influenza-like sickness and having a temperature of 38^{o}C or higher were swabbed during this period. Although testing was done centrally, not all people that should have been tested, were actually swabbed. MHPD estimates that for every swabbed person, there were another three people in the risk group who were not swabbed (private communication). A total of 1,847 people tested in this way between the 21st of September 2009 (week 39) and 20th of June 2010 (week 24), Fig. 2; of these, 622 tested positive to H1N1. Those who tested negative to H1N1 had flu-like symptoms, possibly due to various reasons such as having other respiratory illness. In addition, incorrect swabbing may have resulted in missed cases; late swabbing

or inaccuracy of the swabbing system may also have resulted in an inaccurate virus pick-up rate.

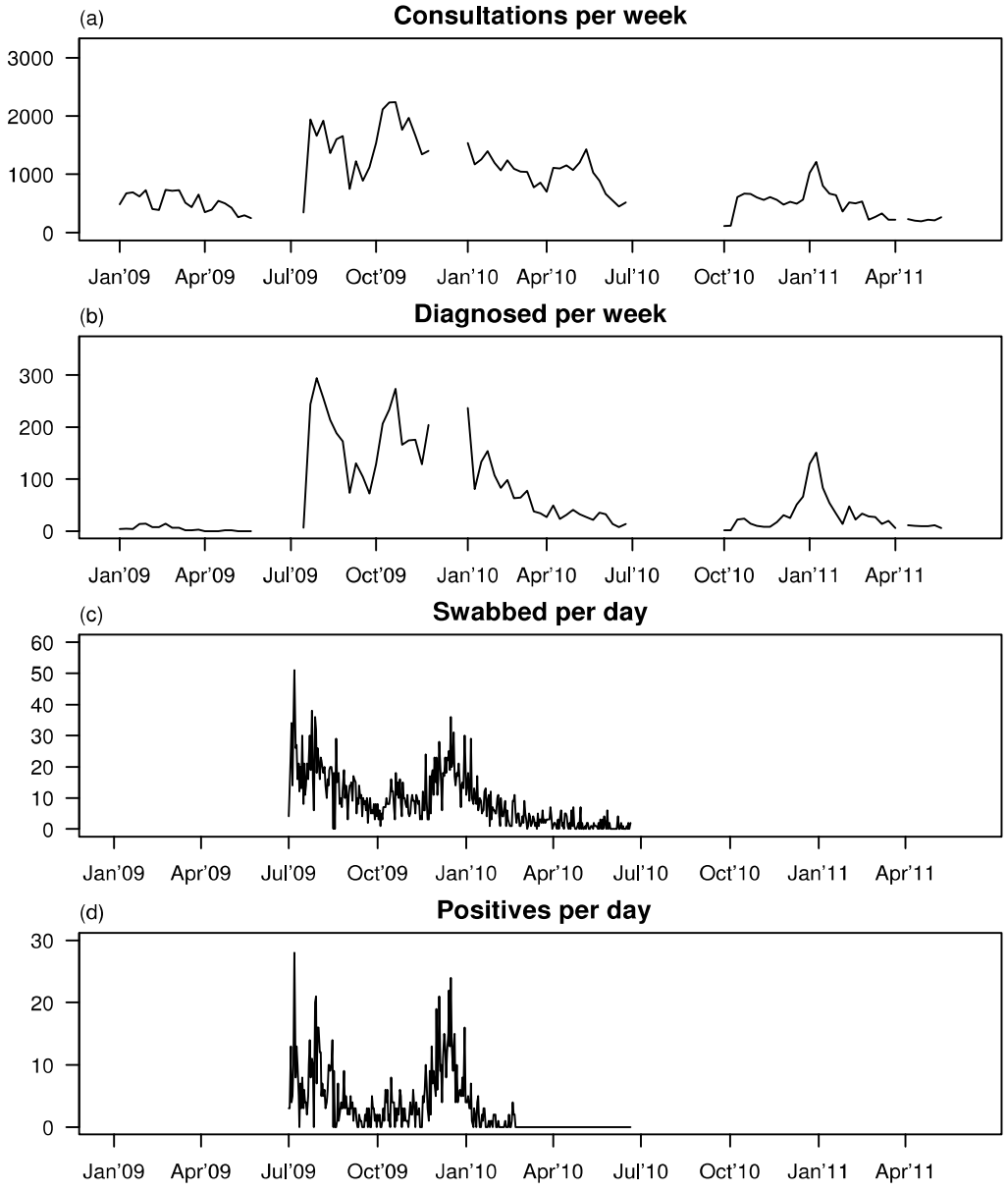


Fig.1 The epidemiological data from Malta covering the period from January 2009 to May 2011. Consultations and Diagnosed were reported weekly by 8 sentinel doctors selected by MHPD. During the H1N1 epidemic, data were collected daily for Swabbed and Positive patients from risk groups; data collected centrally for those doctors who selected to report the case (on average 8.5 doctors per day).

Most of the patients who were swabbed were followed-up, but doctors did not specifically record the date of recovery. Non-fatalities were considered to have recovered within seven days of their swab date, following the usual progression of influenza symptoms. During this period, there were five deaths due to the H1N1 in Malta. Epidemiological data included both residential people and tourists. In fact, one of the deaths recorded was that of a Spanish Tourist.

During January 2010 till the end of February 2010, the vaccine was available to everyone and so, March 2010 can be considered as the end of the epidemic. In total, Malta's Health Department dispensed 2700 courses of antiviral drugs through the government dispensary, but it is known that around 10% of the population had already bought a stock of antiviral drugs which had not yet expired, hence using their own medication. Following the end of February, there were no new positive cases.

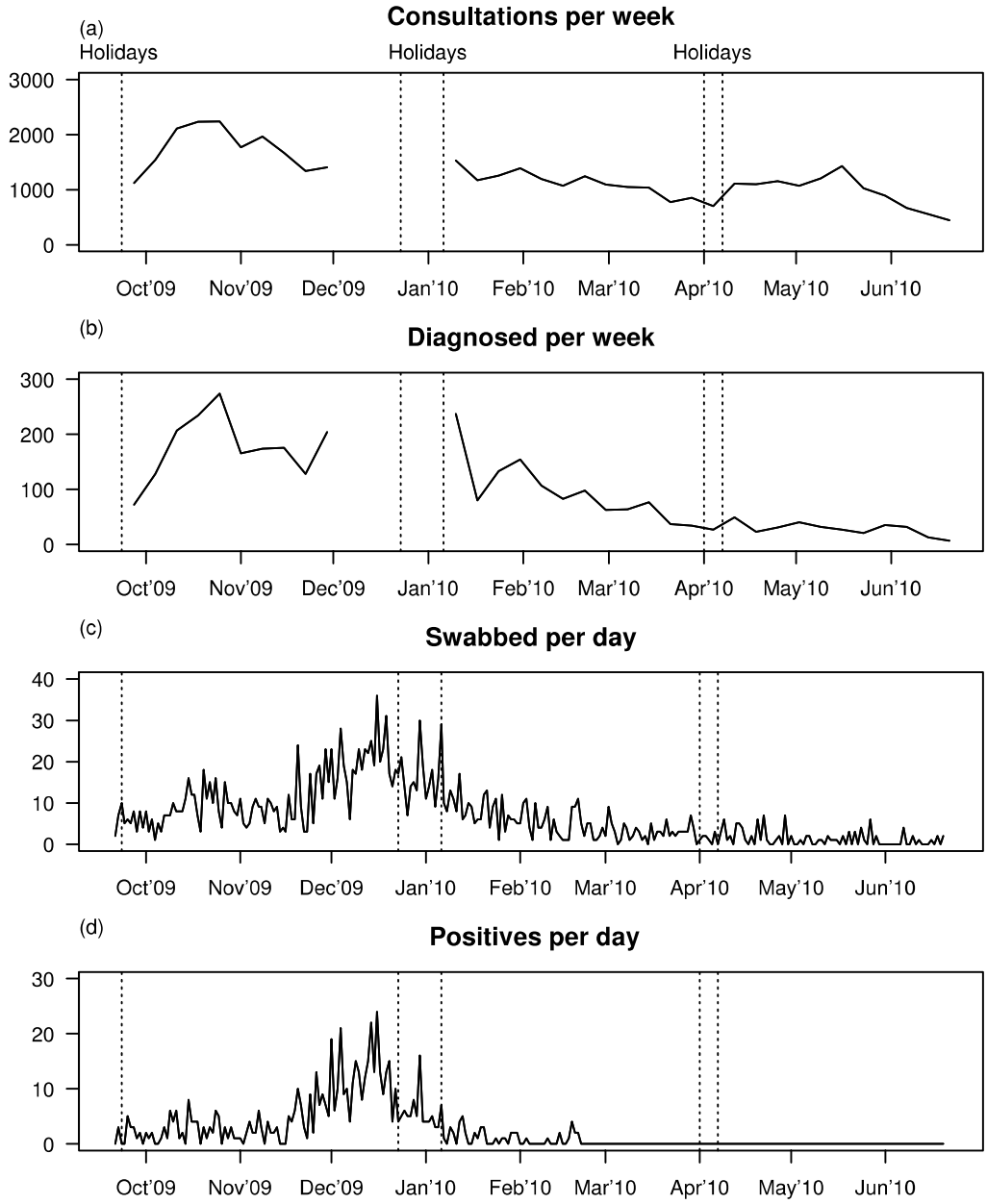


Fig. 2. Malta influenza data used in the analysis. The dotted lines denote Malta's holidays; no apparent correlation with holidays was found in the data.

Data aggregation

In order to compare data collected at different time steps (daily and weekly), we aggregated the daily data by summing the cases over the same intervals as covered by the weekly data. Thus, we analysed the data for swabbed and positive cases twice, once at the daily intervals (as collected) and once at the weekly intervals (corresponding to the consultations and diagnosed cases).

Model

A discrete time SEIR stochastic compartmental model [6, 25] was used to estimate the parameters. The model includes four compartments, Susceptible (*S*), Exposed (*E*) (infected but not infectious), Infectious (*I*) and Recovered (*R*). The SEIR model describes the flow of individuals between the compartments

$$\begin{aligned} S_t &= S_{t-1} - A_t \\ E_t &= E_{t-1} + A_t - B_t \\ I_t &= I_{t-1} + B_t - C_t \\ R_t &= R_{t-1} + C_t \end{aligned} \quad (1)$$

where A_t, B_t and C_t are the numbers of newly infected people in the population, the number of infectious and recovered, respectively. These variables are assumed to binomially distributed and are defined by:

$$\begin{aligned} A_t &\sim Bin\left(S_{t-1}, 1 - e^{\left\{\frac{-[\varepsilon+\beta I_{t-1}]}{N}\right\}}\right) \\ B_t &\sim Bin\left(E_{t-1}, 1 - e^{\left\{\frac{-1}{\alpha}\right\}}\right) \\ C_t &\sim Bin\left(I_{t-1}, 1 - e^{\left\{\frac{-1}{\tau}\right\}}\right) \end{aligned} \quad (2)$$

where ε, β, α^{-1} and τ^{-1} are the importation rate, infection rate of the local population, the rate of transition from exposed to infectious and the rate of transition from infectious to removed, respectively. Hence α represents the latent period, and τ the infectious period.

The population size is taken to be the total population of Malta, 414,000. The vector of parameters $\boldsymbol{\theta} = (\beta, \varepsilon, \alpha, \tau)$ and the current state $\boldsymbol{\Sigma}_t = \{S_t, E_t, I_t, R_t\}$ are unknown. Observations, D_t, are assumed to be Poisson distributed with mean $N_t \delta_{d(t)} \left\{\phi + \frac{I_t}{300}\right\}$ where N_t is the number of physicians submitting reports on day *t* and $\delta_{d(t)}$ is the weight

associated with a given day of the week $d(t)$ corresponding to the current day *t*; Monday being equal to 1, Tuesday being equal to 2 and so on. Then, δ_i is the proportion of individuals seeking medical help on the day of the week *i*. For weekly data, only one δ was used. ϕ represents the 'background' consulting rate (for consultations this term will represent all patients visiting a doctor for any non-flu illness; for other data this term corresponds to non-H1N1 ILIs). The number of physicians in Malta was estimated to be around 300 and so is used here to convert the actual total number of cases I_t to the number of observations by selected physicians.

Once the parameters are computed, the effective reproduction ratio at any given time t is calculated according to:

$$\frac{\beta\left(1-e^{-\frac{1}{\tau}}\right)S}{N} \qquad (3)$$

where β is the infection rate, τ is the recovery rate, S is the current number of susceptible individuals and N the population size.

Parameter estimation

The particle filter algorithm [26,27] is a sequential Monte Carlo algorithm designed to represent the posterior density by a set of random particles with associated weights. Details of the approach are given in [6] and we only provide a short summary here.

The algorithm starts at time *t=0*, and with a set *P* of initial states Σ_0 and parameters θ generated from the prior distribution. For each particle, *p*, at each time step *t+1*, Σ_{t+1} is drawn using Monte Carlo simulation from its conditional distribution given x_t^p, where $x_t^p = (\Sigma_t, \vartheta)$ with an associated weight w_t^p. Following this, we set $x_{t+1}^p = (\Sigma_{t+1}, \vartheta)$ and calculate the likelihood contribution $L_{t+1}^p = f(D_{t+1}|x_{t+1}^p)$ conditioned on the path of the respective particle using the same parameter values and on D_t , which is the number of reported cases on day *t*. This likelihood is then used to find the weights by setting $w_{t+1}^{*(p)} = w_t^{(p)} L_{t+1}^{(p)}$ which are then scaled to sum to one: $w_{t+1}^{*(p)} = \frac{w_{t+1}^{*(p)}}{\sum_{q=1}^{P} w_{t+1}^{*(q)}}$.

Re-sampling [27] is used to 'recover' particles that are assigned low weights by letting $x_{t+1}^{*(p)} = x_{t+1}^{*(q)}$ where q is selected from the set of integers $\{1,2,\dots,P\}$ with probability

proportional to $w_{t+1}^{*(q)}$. Thus, whenever some of the particles fell below a certain threshold, the current set of particles were re-sampled. Particle diversity is retained by kernel smoothing [6, 28]. The complete algorithm is then repeated and the state values at time *t+1* are calculated using parameters for time *t*.

Priors

The prior distributions were based on priors used in Ong. et al. [6] and were generally very broad. For the daily datasets, the infection rate, β was assumed to follow a normal distribution with mean and standard deviation equal to 1. The prior distribution for the daily importation rate, ε, follows a normal distribution with mean 30 and standard deviation equal to 15; for the latent period, α, the daily prior distribution was set to $N^{+}(1,1)$. For the infectious period, τ, the prior for the daily data was set to $N^{+}(2,0.5)$. For the daily background rate, ϕ, the prior was set to $N^{+}(1,0.25)$. For the four weekly datasets, β was assumed to follow a normal distribution with mean and standard deviation equal to 2; importation weekly rate, ε, a normal distribution with mean 80 and standard deviation 60. The prior distribution for the weekly latent period, α, was set $N^{+}(1,1)$ for all weekly datasets. For the infectious period, the prior followed a normal distribution with mean of 1 and standard deviation of 1. The prior distribution for the background rate, ϕ, for the consultations was set to $N^{+}(750,300)$, while for all the other weekly datasets to $N^{+}(1,0.25)$. The consultations dataset includes a substantial number of non-flu illness hence the high prior number for the background rate.

The prior distributions for *E(0)* and *I(0)* were derived using the number of confirmed cases at the start of the epidemic, normally distributed, with mean and variance related to the observed values of *I(0)* using similar approach to Ong et al [6]. As the epidemic analyzed here follows from the first summer wave, we used rough estimate of cases between July ’09 and September ’09 as a guide for choosing *R(0)*. For the consultation and diagnosed data, the *R(0)* value was assumed equal to 65,000, for swabbed equal to 50,000 and for positive equal to 20,000. For the consultation we assumed the same *R(0)* as diagnosed, but then for the consultation data we assumed a much higher prior for the background rate. The prior distribution for the proportion of infected seeking medical help, δ, for all data sets except consultation was assumed to follow beta distribution, $\beta(5,15)$, while for the consultation data $\beta(15,5)$. The mean for the prior beta distribution

for consultation is 0.75 while for the other data sets is 0.25, reflecting large number of consultations cases.

Simulation parameters

The performance of the simulations depends on the size of the datasets. The memory and time constraints limit the number of particles that can realistically be used for large datasets. Hence, for daily swabbed data, a series of 10,000 particles is used while for a smaller daily positive data set, a series of 15,000 particles is used. For the weekly data 50,000 particles were used. R statistical programming language [29] was used to run the particle filtering algorithm and the SEIR model.

Results

Three periods can be identified in the data that describe consultations and influenza diagnosed from January 2009 to May 2011, Fig. 1. The first (January 2009-June 2009) period is characterised by a very low level of influenza infections (Fig. 1b), whereas consultations for any illnesses (including influenza) are relatively stable at approximately 500-700 per week. The last (October 2010-May 2011) of these periods illustrates typical seasonal influenza outbreaks, characterized by a winter peak in flu cases (Fig. 1b), which is also visible in consultations above the background level of other illnesses (Fig. 1a). In contrast, the 2009/2010 outbreak shows a massive increase in consultations (Fig. 1a) that can be almost entirely associated with the H1N1 influenza (more detailed analysis below). Three waves can be identified in the period July 2009-June 2010, with the first (summer) wave essentially finished by the time children returned to school in September 2009 and the second (October-November) wave initiated shortly afterwards and the third (December-March) wave followed. Data recording is more complete for the second and the third waves and in particular we are able to capture the initial stages of this outbreak. Thus, in this paper we are concentrating our analysis on the period September 2009-June 2010, Fig. 2.

The data reflect the process of identification of H1N1 influenza among patients who sought help from the doctors. There is a broad agreement between the excess of consultations above the background and the number of diagnosed individuals, Figs. 2(a) and (b), and the relationship can be approximated by a linear function (R^2=0.71), Fig. 3a (we discuss this relationship in more detail later in the paper). The background level of

consultations (for any illnesses which are not related to the influenza) can be estimated from the linear relationship at about 770 consultations per week, in good agreement with the rest of the data shown in Fig. 1a. The approximately linear relationship seen in Fig. 3 can be used to reconstruct the missing portion of data for consultations and diagnosed for December 2009, see Fig. 4. Up to 64% of swabbed samples tested positively for H1N1 (cf. Fig. 2c with Fig. 2d), although no more positive cases were identified after 21 February 2010.

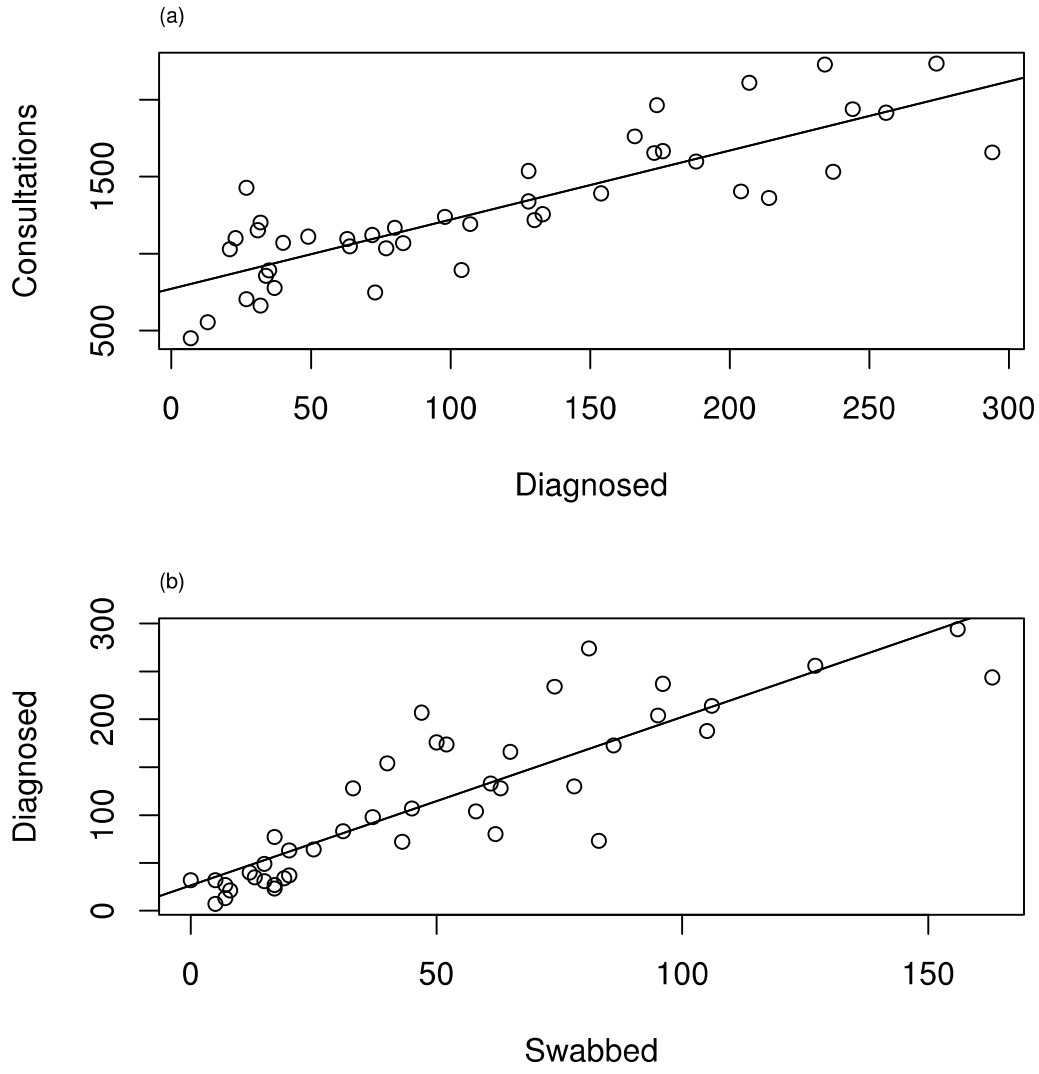


Fig. 3. Relationship between the number of consultations and diagnosed (a) and the number of diagnosed and swabbed (b) over the period shown in Fig. 2. Lines of best least-squares fit are used to 'reconstruct' the missing data. *Consultations = 772.32+4.49 (Diagnosed)*, R^2=0.76 and *Diagnosed=26.54+1.76 (Swabbed)*, R^2=0.71. The diagnosed was first 'reconstructed' from swabbed data and subsequently, the consultations from diagnosed.

All four data sets follow a typical epidemic curve, with an initial slow build-up up to mid-November 2009 followed by the main epidemic wave in December 2009 and a decline

to approximately constant level from March 2010 onwards, Figs. 2 and 3. This behaviour is broadly consistent with other data sets available in the literature [12, 18, 22, 30-33]. However, two main periods can be identified in the Malta data, Figs. 2 and 4. In the early phase (October-December 2009) the level of consultations and diagnosis was high but the number of individuals referred for further testing (swabbed) and the resulting number of confirmed cases of H1N1 remained relatively low. For instance, consultations peaked in October 2009 and again in December 2009, but swabbed and positives have only one peak in December, see Fig. 4. The data for swabbed and positive individuals aggregated at the weekly intervals unsurprisingly reveal more variation (Figs. 2c and 2d), some of which can be associated with the day of the week, see Fig. 5.

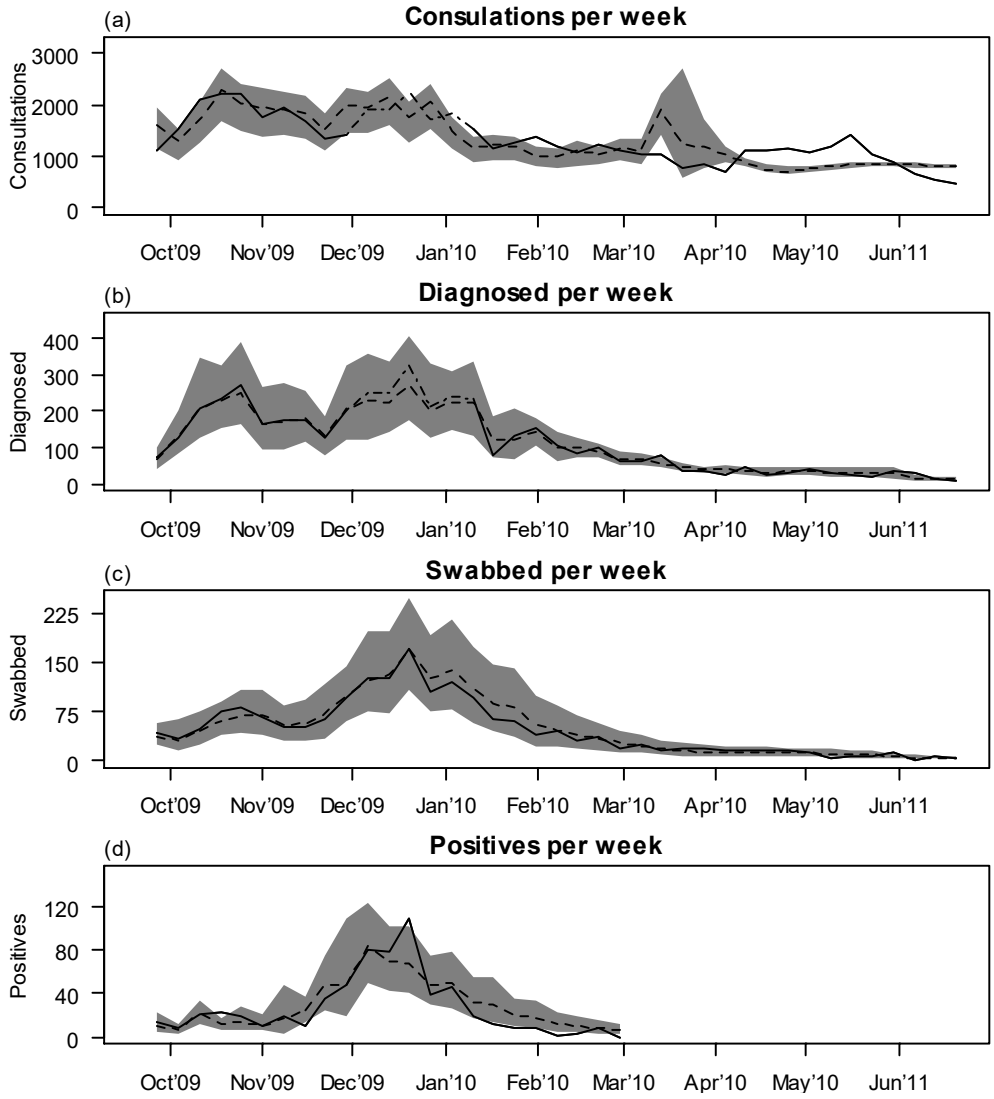


Fig. 4. Comparison of weekly (Consultations, (a), and Diagnosed, (b)) and weekly-aggregated (Swabbed, (c), and Positive, (d)) data, solid line, with the results of model fit, dashed line (mean) and shaded area (95% high predictive density regions). The 'reconstructed' data for consultations and diagnosed cases is marked by dashed-dotted line.

The model successfully represents the main features of all datasets, both for the weekly datasets (with the swabbed and positives aggregated over the weekly periods), Fig. 4, and for the daily sampling rate, Fig. 5. Note that we used the background consulting rate ϕ to represent the consultations that are not associated with the influenza outbreak. In particular, both waves (October and December 2009, respectively) are captured by the model and so are their relative strengths, revealed particularly in the weekly data. In addition, some fine scale oscillations are captured by the model at the higher resolution, Fig. 5.

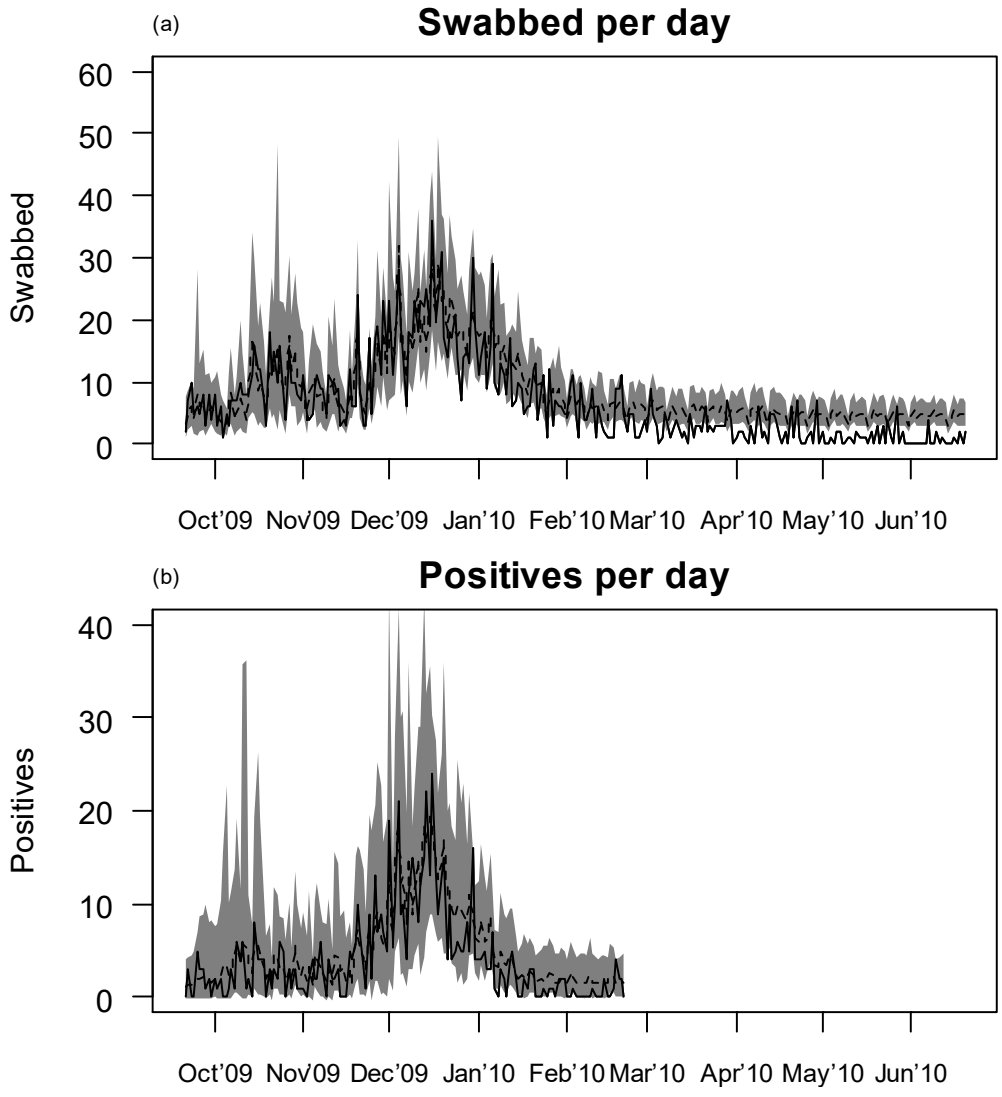


Fig. 5. Comparison of daily (Swabbed and Positive) data, solid line, with the results of model fit, dashed line (mean) and shaded area (95% high predictive density regions).

The estimates of individual parameters vary widely among different datasets and the sampling frequencies, Table 1, but the estimates of the effective reproduction ratios R_t

based on different epidemiological proxies are broadly consistent among the four datasets for the weekly sampling, Fig. 6. They are also consistent with other datasets available in the literature, for example see Ong. et al. [6]. The initial attack rate is high, with R_t values of order 3-6 and therefore well over the invasion threshold of R_t=1. The second wave in December has a lower rate of growth than the October one and is also initiated with a higher value of already infected individuals. It is therefore associated with relatively lower values of R_t. The epidemic peak is again reflected in the estimates of R_t for swabbed and positive data, with R_t consistently exceeding 1 until well into January 2010. Interestingly, the R_t estimates for consultations individuals drop below 1 already in November and stay below the threshold, Fig. 6.

The posterior variability in the estimates of parameters is initially high (Fig. 7), but quickly settles on the final values. These long-term estimates are largely independent of the prior choice, except for ε and ϕ.

Definitions	Parameter	Daily Swabbed (10,000 Particles)	Daily Positive (15,000 Particles)	Weekly Consultations (50,000 Particles)	Weekly Diagnosed (50,000 Particles)	Weekly Swabbed (50,000 Particles)	Weekly Positive (50,000 Particles)
Infection Rate (day^{-1} or $week^{-1}$)	β	0.29 (0.16-0.43)	0.38 (0.18-0.58)	0.86 (0.83-0.89)	1.29 (0.90-1.67)	1.13 (0.90-1.35)	0.74 (0.61-0.88)
Importation rate (day^{-1} or $week^{-1}$)	ε	58.32 (12.77-103.87)	67.88 (-15.16-150.91)	6.08 (4.10-8.07)	581.55 (45.03-1,118.06)	98.93 (-73.07-270.94)	246.19 (-17.95-474.44)
Latent period (day or week)	α	0.07 (-0.04-0.17)	0.31 (-1.28-1.89)	0.32 (0.17-0.48)	0.54 (0.14-0.95)	0.07 (-0.01-0.15)	0.05 (-0.01-0.11)
Infectious period (day or week)	τ	2.71 (1.91-3.51)	1.42 (0.77-2.07)	3.87 (3.42-4.32)	0.68 (0.42-0.94)	0.47 (0.27-0.68)	0.39 (0.26-0.53)
Background rate (day^{-1} or $week^{-1}$)	ϕ	1.06 (0.75-1.37)	0.70 (0.39-1.0)	294.95 (285.83-304.07)	0.79 (0.46-1.11)	1.08 (0.60-1.55)	1.50 (0.94-2.06)
Reporting rate	δ	n/a	n/a	0.33 (0.31-0.34)	0.60 (0.54-0.65)	0.31 (0.20-0.42)	0.13 (0.06-0.20)
Monday reporting rate	δ_1	0.15 (0.10-0.21)	0.21 (0.12-0.30)	n/a	n/a	n/a	n/a
Tuesday reporting rate	δ_2	0.24 (0.15-0.32)	0.27 (0.16-0.38)	n/a	n/a	n/a	n/a
Wednesday reporting rate	δ_3	0.29 (0.20-0.38)	0.26 (0.16-0.35)	n/a	n/a	n/a	n/a
Thursday reporting rate	δ_4	0.23 (0.16-0.31)	0.2 (0.12-0.29)	n/a	n/a	n/a	n/a
Friday reporting rate	δ_5	0.25 (0.16-0.35)	0.2 (0.11-0.29)	n/a	n/a	n/a	n/a
Saturday reporting rate	δ_6	0.26 (0.17-0.35)	0.22 (0.12-0.32)	n/a	n/a	n/a	n/a
Sunday reporting rate	δ_7	0.24 (0.16-0.31)	0.21 (0.09-0.33)	n/a	n/a	n/a	n/a

Table 1: Parameter values estimated for different data sets. Numbers in brackets represent highest density 95% symmetric credible intervals based on a normal approximation to posterior distributions.

Among the parameters for the weekly data, the infection rate, β, is decreasing as the proxy becomes more specific, except for the consultation data (diagnosed>swabbed>positive), Table 1. The estimate for the external infection pressure, ε, is characterised by huge variability (Fig. 7). In addition, the data resolution did not allow us to identify the imported cases to compare the estimate with the data. There is some uncertainty associated with the latent period (Table 1) suggesting that the data are not able to pinpoint its actual value. The infectious period based on weekly diagnosed, swabbed and positive data is on average about 3.5 days, slightly longer than Ong et al [6] estimates. The estimates for τ based on daily data are more consistent with Ong et al [6] (1-2 days). There does not seem to be much variation between days of the week for the weekly data, again consistent with Ong et al [6]. Finally, the background consultation rate is high for the consultations data reflecting the need for accounting for non-ILI patients, whereas for other datasets it is relatively low. Note that ϕ in Table 1 is calculated per doctor – with 8 doctors on average reporting per week.

Discussion

Epidemiological models can only be used in practical applications if we successfully and reliably can parameterise them. This, in turn, depends on the quality of available data. Unfortunately, this situation is rare in human epidemiology of influenza and similar diseases as we always struggle with incomplete data coming from different sources and at different sampling intervals. Moreover, we only rarely can infer the number of actual cases – more often we have access to various proxies which in different ways represent the progress of the epidemic. In this paper we use a multi-proxy dataset from the 2009-2010 H1N1 epidemic in Malta. The SEIR compartmental model is used to estimate the current value of the effective reproductive ratio, R_t. We show that the results from different proxies are basically consistent, although in some cases we observe $R_t<1$ from some proxies and $R_t>1$ for others. We also note a general linear relationship between different epidemic proxies.

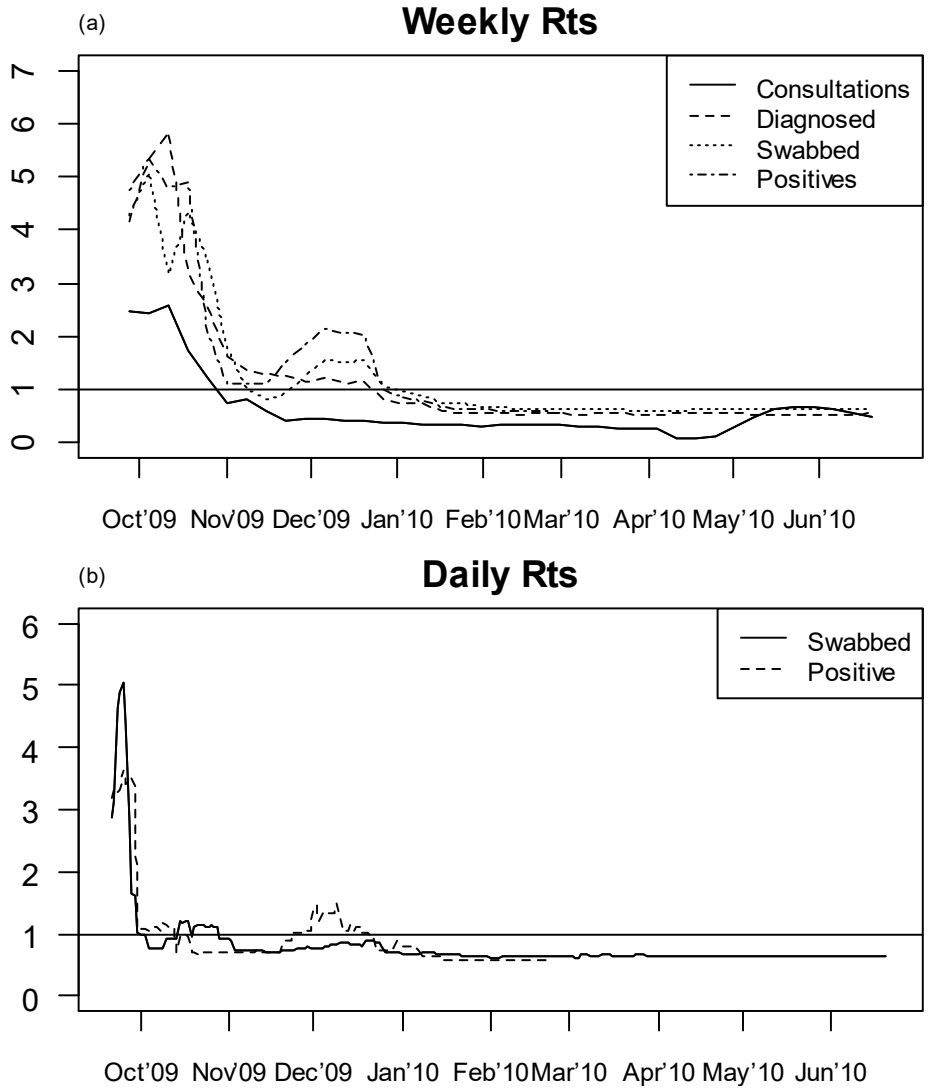


Fig. 6. Estimation of the effective reproduction ratio at any given point of the epidemic for different data sets, including weekly (Consultations and Diagnosed) and weekly-aggregated (Swabbed and Positives) data, (a), and daily (Swabbed and Positives) data, (b). Horizontal line corresponds to R_t=1, an invasion threshold.

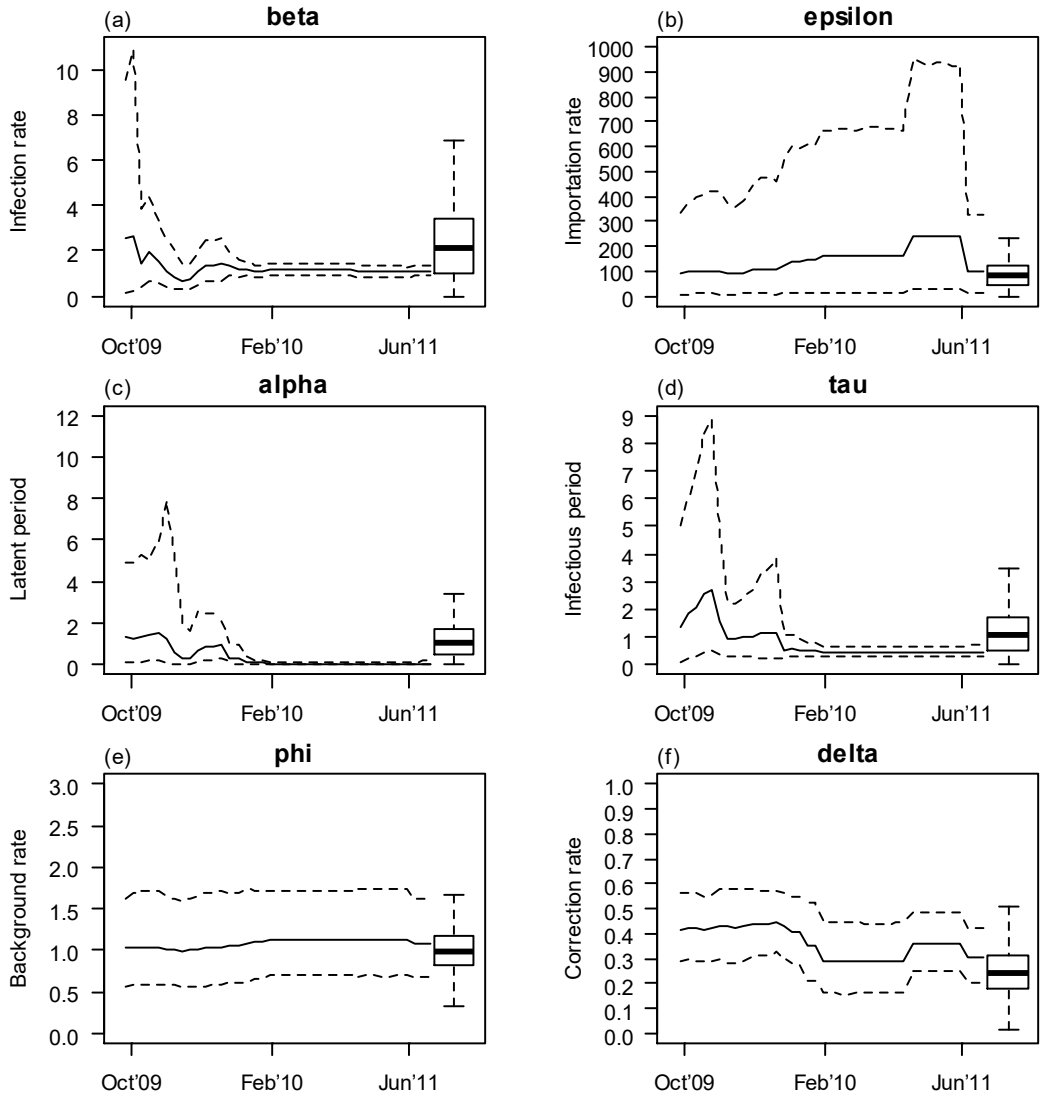


Fig. 7. Posterior and Priors parameter distributions for the swabbed weekly data (for illustration). The box-plot represents on right represents the prior distribution, whereas the graph shows the evolution of the posterior distribution over time (solid line represents the mean and the dotted lines show the marginal point-wise 95% credible intervals).

However, the datasets presented here allow us an even more detailed study of the relationship between different approximate datasets each describing the same epidemic. In particular, as the proxies become more specific, they introduce different biases and different processes underlying the reporting of data. The *consultations* reflect individual's need for seeing a doctor regardless of whether the person has or has not got influenza. In among consultations for other illnesses there will be patients with influenza, but who do not satisfy the 'official' criteria for influenza, as well as 'true' cases. The doctor will then assign the *diagnosed* status, again with some level of arbitrariness. The problem with these data is that they are only collected at the weekly period and reported by a small number of doctors. There is therefore a large uncertainty associated with the data. Only

individuals at risk are *swabbed* but the recording is much stricter and if we can assume that the disease affects both individuals at risk and not at risk equally, then the record of swabbed can be a good representation of doctor's diagnose of influenza. However, the swabbed person might not really have influenza or if he/she has one, it might not be H1N1. The *positive* result of testing confirms the H1N1 infection, but introduces further bias, as the test is not fully accurate. In this paper we have investigated the relationship between this different datasets and how the use of one proxy or another influences the parameter estimation. In particular, we found that broadly the different proxies are related to each other by an approximately linear relationship, Fig. 3 and Fig. 8.

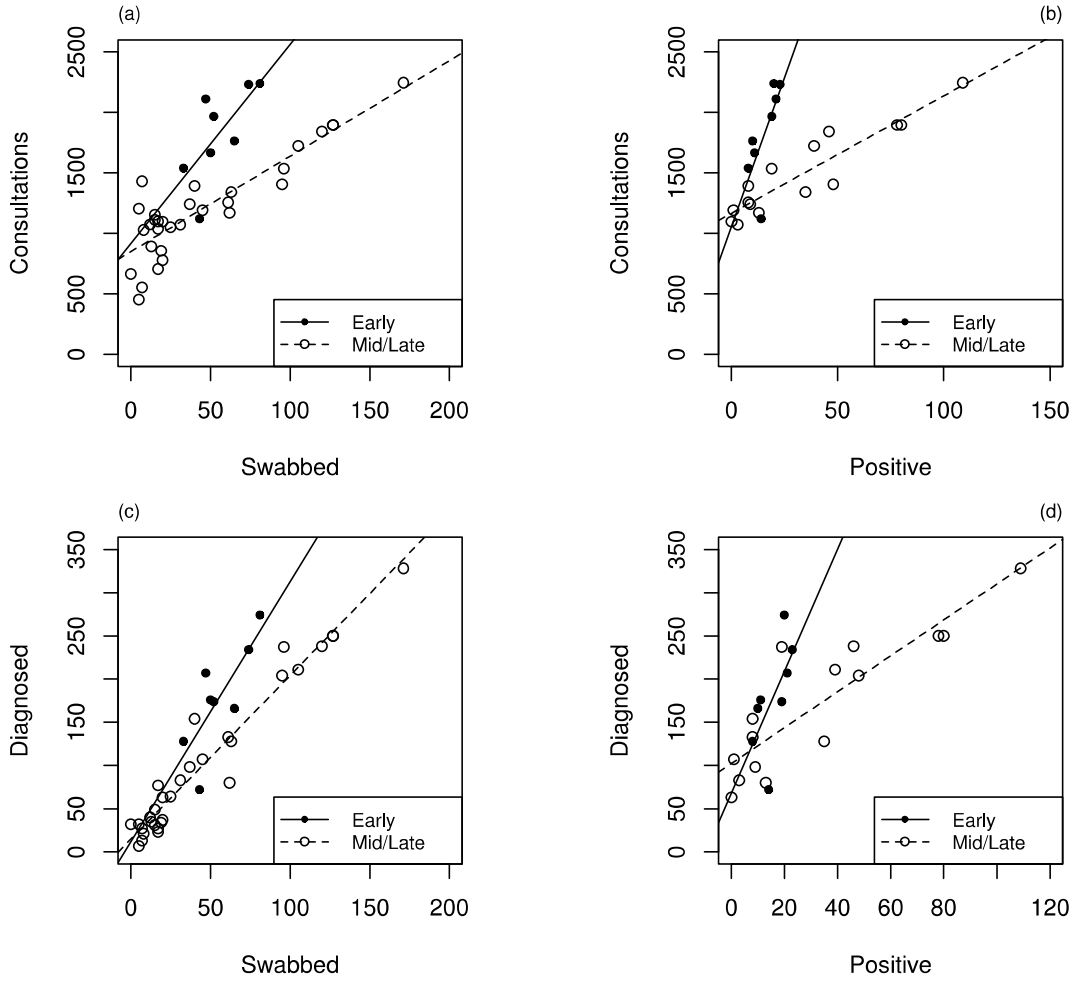


Fig. 8. Relationship between weekly and weekly-aggregated data for different periods in the epidemic timeline. Early period (weeks 39/2009 to 46/2009) is characterised by high overall levels and high variability of consultations and diagnosed cases as compared to swabbed and positive.

However, there is an additional time-dependent factor that becomes apparent when these relationships are considered for different parts of the epidemic (we limit ourselves here

to weekly data, with aggregation of the daily data for swabbed and positive). We split the period from October 2009 and June 2010 into two periods; see Fig. 2 and Fig. 8. In the early period (weeks 39-46 in 2009), the slope relationship between the level of consultations/diagnosed and swabbed/positive cases is much higher than in the second period (weeks 47 in 2009 to 13 in 2010). Thus, while the number of swabbed and positive cases is much smaller in the first (autumn) wave of the epidemic than in the second (winter) wave, the number of consultations/diagnosed cases is comparable between the two waves, Figs. 2 and 4. Thus it appears that many people actually sought consultations in the first period and were diagnosed by doctors as having influenza. However, most of these cases seem to be rather mild and so doctors were not performing swabbing in this period, Figs. 2 and 4. The number of positive cases was even smaller than the number of swabbed cases, further corroborating the interpretation of the first period as dominated by panic among the public.

In contrast, for the mid to late period (weeks 47-2009 to 24-2010), the number of consultations seems to largely follow the swabbed and positive cases (Fig. 8). As in the early period, it seems that the number of consultations rises again after April 2010, but this is not reflected in either diagnosed or swabbed cases (there are no positive cases after February 21 and so we do not show those data in Fig. 8).

This lack of stationarity in the relationship between the information that can be gathered from doctor's reports (consultations and diagnosed) and what the more detailed epidemiological analysis can reveal (swabbed and positives) is reflected in a small difference among the estimates of the effective reproduction ratios, R_t, Fig. 5. In particular, while the estimate based on diagnosed, swabbed and positive individuals remains above one in the winter period (November through January), the consultation data suggest that the influenza was not spreading during this time period (R_t close to, but below 1).

Further work needs to be done to understand the process by which different approximate data are produced and influenced, for example, by news. This might lead to an improved way of translating different proxies (and in particular ILIs) into infected individuals for the purpose of fitting dynamic, SIR-like models. The relationship between the observed and actual cases is usually assumed to be linear and independent of the stage of the

epidemic. Our results show that the relationship might be linear, but it is certainly not constant. The feedback between the number of cases and the reporting efficiency needs to be studied in more detail and might lead to modified SIR models leading to improved ability to predict a future course of any outbreak in real time. Similarly, prediction can be improved if different proxies can be combined into one framework. This can be achieved in the Bayesian framework, but probably would need an explicit model of various stages of data collection.

Acknowledgements

We are very much indebted to Malta Health Promotion Department for provision of datasets and for continuous help throughout the project.

References

[1] G. Chowell, S. Echevarria-Zuna, C. Vibound, L. Simonsen, J. Tamerius, M.A. Miller, V.H. Borja-Aburto, Characterizing the Epidemiology of the 2009 Influenza A/H1N1 Pandemic in Mexico, PLoS Med. **8**(5) (2011).

[2] S. Flasche, N. Hens, P.Y. Boelle, J. Mossong, W.M.V. Ballegooijen, B. Nunes, C. Rizzo, F. Popovici, P. Santa-Olalla, F. Hruba, K. Parmakova, M. Baguelin, A.J.V. Hoek, J.C. Desenclos, P. Bernillon, A.L. Camara, J. Wallinga, T. Asikainen, P.J. White, W.J. Edmunds, Different transmission patterns in the early stages of the influenza A(H1N1)v pandemic: A comparative analysis of 12 European countries, Epidemics **3** (2011) 125-133.

[3] WHO (2010) Pandemic (h1n1) 2009 - update 100. World Wide Web electronic publication. URL http://www.who.int/csr/don/2010_05_14/en/index.html

[4] A. Ishak , D. Tee, I. Nawmar, L.K. Pang, N. Ruslan, N. Che Mansor, L. Gam, H1N1 Influenza: A Viral Infection, WebmedCentral INFECTIOUS DISEASES, **2**(12) (2011) WMC002736.

[5] C. Reed, F.J. Angulo, D.L. Swerdlow, M. Lipsitch, M.I. Meltzer, D. Jarnigan, L. Finelli, Estimates of the Prevalence of Pandemic (H1N1) 2009, United States, April-July 2009, Emerging Infectious Diseases, *15*(12) (2009) 2004-2007.

[6] J.B.S. Ong, M.I.C. Chen, A.R. Cook, H.C. Lee, V.J. Lee, R.T.P. Lin, P.A. Tambyah, L.G. Goh, 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**(4) (2010) e10036.

[7] G. Chowell, H. Nishiura, L.M.A. Bettencourt, Comparative estimation of the reproduction number for pandemic influenza from daily case notification data, J. R. Soc. Interface, **4** (2006) 155-166.

[8] G. Chowell, C. Viboud, C.V. Munayco, J. Gomez, L. Simonsen, M.A. Miller, J. Tamerius, V. Fiestas, E.S. Halsey, C.A. Laguna-Torres, Spatial and Temporal Characteristics of the 2009 A/H1N1 Influenza Pandemic in Peru, PLoS ONE **6**(6) (2011) e21287.

[9] J.T. Griffin, T. Garske, A.C. Ghani, Joint estimation of the basic reproduction number and generation time parameters for infectious disease outbreaks, Biostatistics, **12**(2) (2011) 303-312.

[10] P.Y. Boëlle, P. Bernillon, J.C. Desenclos, A preliminary estimation of the reproduction ratio for new influenza A(H1N1) from the outbreak in Mexico, March–April 2009, Euro surveillance, **14**(19):pii=19205 (2009).

[11] C. Fraser, C.A. Donnelly, S. Cauchemez, W.P. Hanage, M.D. Van Kerkhove, T.D. Hollingsworth, J. Griffin, R.F. Baggaley, H.E. Jenkins, E. J. Lyons, T. Jombart, W.R. Hinsley, N.C. Grassly, F. Balloux, A.C. Ghani, N.M. Ferguson, A. Rambaut, O.G. Pybus, H. Lopez-Gatell, C.M. Alpuche-Aranda, I.B. Chapela, E.P. Zavala, D.M.E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings, Science, **324** (2009) 1557-1561.

[12] Y. Hsieh, K. Cheng, T. Wu, T. Liz, C. Cheng, J. Chen, M. Lin, Transmissibility and temporal changes of 2009 pH1N1 pandemic during summer and fall/winter waves, BMC Infectious Diseases, **11**:332 (2011).

[13] D. Clancy, P.D. O'Neill, Bayesian estimation of the basic reproduction number in stochastic epidemic models, International Society for Bayesian Analysis, **3** (2008) 737-758.

[14] L.F. White, J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, M. Pagano, Estimation of the Reproductive Number and the Serial Interval in Early Phase of the 2009 Influenza the Current Influenza A/H1N1 Pandemic in the USA, Influenza Other Respi Viruses, **3**(6) (2009) 267-276.

[15] G. Katriel, R. Yaari, A. Huppert, U. Roll, L. Stone, Modelling the initial phase of an epidemic using incidence and infection network data: 2009 H1N1 pandemic in Israel as a case study, J. R. Soc. Interface, **8** (2011) 856-867.

[16] A. Flahault, E. Vergu, P.Y. Boëlle, Potential for a global dynamic of Influenza A(H1N1), BMC Infectious Diseases, **9**:129 (2009).

[17] L. Opatowski, C. Fraser, J. Griffin, E. de Silva, M.D. Van Kerkhove, E.J. Lyons, S. Cauchemez, N.M. Ferguson, Transmission Characteristics of the 2009 H1N1 Influenza Pandemic: Comparison of 8 Southern Hemisphere Countries, PLoS Pathog, **7**(9) (2011) e1002225.

[18] E. Kenah, D.L. Chao, L. Matrajt, M.E. Halloran, I.M. Longini Jr., The Global Transmission and Control of Influenza, PLoS ONE, **6**(5) (2011) e19515.

[19] D. Buckley, D. Bulger, Estimation of the reproductive number for the 2009 pandemic H1N1 influenza in rural and metropolitan New South Wales, Aust. J. Rural Health, **19** (2011) 59–63.

[20] H. Nishiura, D. Klinkenberg, M. Roberts, J.A.P. Heesterbeek, Early Epidemiological Assessment of the Virulence of Emerging Infectious Diseases: A Case Study of an Influenza Pandemic, PLoS ONE, **4**(8) (2009) e6852.

[21] C.Y. Chang, C.X. Cao, Q. Wang, Y. Chen, Z. Cao, H. Zhang, L. Dong, J. Zhao, M. Xu, M. Gao, The novel H1N1 Influenza A global airline transmission and early warning without travel containments, Chinese Science Bulletin, 2010, **55** (2010) 3030−3036.

[22] A.M. Correia, L. Queiros, J. Dias, Pandemic influenza A (H1N1) in the North of Portugal: how did the Autumn-Winter wave behave?, Rev Port Pneumol, **16**(6) (2010) 880-886.

[23] C. Rizzo, M.C. Rota, A. Bella, S. Giannitelli, S. De Santis, G. Nacca, M.G. Pompa, L. Vellucci, S. Salmaso, S. Declich, Response to the 2009 influenza A(H1N1) pandemic in Italy, Euro Surveillance, **15**(49):pii=19744 (2010).

[24] H. Yu, S. Cauchemez, C.A. Donnelly, L. Zhou, L. Feng, N. Xiang, J. Zheng, M. Ye, Y. Huai, Q. Liao, Z. Peng, Y. Feng, H. Jiang, W. Yang, Y. Wang, N.M. Ferguson, Z. Feng, Transmission Dynamics, Border Entry Screening, and School Holidays during the 2009 Influenza A (H1N1) Pandemic, China, Emerging Infectious Diseases, **18**(5) (2012) 758-766.

[25] R. Anderson, R. May R, Infectious Diseases of Humans, first ed., Oxford University Press, Oxford, 1991.

[26] A. Doucet, S. Godsill, C. Andrieu, On sequential Monte Carlo sampling methods for Bayesian filtering, Statistics and Computing, **10** (2000) 197–208.

[27] A. Doucet, N. De Freitas, N. Gordon, Sequential Monte Carlo methods in practice, first ed., Springer Verlag, New York, 2001.

[28] V.M. Trenkel, D.A. Elston, S.T. Buckland, Fitting population dynamics models to count and cull data using sequential importance sampling, Journal of the American Statistical Association, **95** (2000) 363–74.

[29] R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2010) URL http://www.R-project.org, ISBN 3-900051-07-0.

[30] P. Poletti, M. Ajelli, S. Merler, The Effect of Risk Perception on the 2009 H1N1 Pandemic Influenza Dynamics, PLoS ONE, **6**(2) (2011) e16460.

[31] R. Omori, H. Nishiura, Theoretical basis to measure the impact of short-lasting control of an infectious disease on the epidemic peak, Theoretical Biology and Medical Modelling, **8**(2) (2011).

[32] H. Nishiura, Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009), BioMedical Engineering OnLine, **10**(15) (2011).

[33] A. Fierro, A simple stochastic lattice gas model for H1N1 pandemic. Application to the Italian epidemiological data, The European Physcial Journal E, **34** (2011) DOI 10.1140/epje/i2011-11011-2.

Chapter 4

Modelling seasonal influenza

4.1 Introduction

In Chapter 3, I developed the techniques for analyzing the way in which limited information about influenza outbreak affects the modelling and ultimately the prediction of the number of cases through the SEIR model. This analysis so far has been limited to the pandemic data from the 2009-2010 season. The main objective of this chapter is to extend the analysis of chapter 3 to the seasonal influenza over four different seasons. This will provide further information about the relationship between the number of diagnosed cases and number of consultation cases across four different seasons. Four seasonal influenza datasets were acquired from the Malta Health Authorities, as defined in chapter 2. Therefore, in addition to the analysis defined in chapter 3, the SEIR modelling techniques will be applied on the new acquired seasonal influenza datasets. Furthermore, we will try to establish the linear regression models (as defined in Chapter 2) between the consultations and diagnosed datasets for all the different influenza seasons (as in Chapter 3). However, the main challenge in this chapter is that a lower number of cases exist, which makes the analysis more difficult. Thus, the main question in this chapter is to understand the extent to which we can use the linear relationship (obtained through the linear regression model) between the diagnosed and consultation datasets to predict one dataset from the other. Furthermore, we aim to understand the variability of the posterior parameters (obtained through the SEIR model) of the diagnosed and consultation datasets between different influenza seasons. Finally, we will analyze whether there is an opportunity to combine the linear regression model together with the SEIR model (Joint model). In order to do this, first we need to look at the basic characteristics of all acquired datasets, followed by an analysis of the linear associations between the consultations and diagnosed datasets. Then, the SEIR model (as defined in chapters 2 and 3) will be applied to obtain the posterior parameters values of all datasets. The analysis is concluded by a joint model of the above two modelling techniques.

4.2 The influenza datasets

Throughout this section, I will analyse the data mentioned in chapter 2 in more detail. As explained previously, this data includes the number of doctors reporting the cases, the number of consultations and the number of diagnosed Influenza-Like Illness (ILI) cases seen by the same doctors. For the scope of this chapter, four consecutive years of seasonal influenza together with the pandemic influenza period, as defined by Marmara et al.

(2014) [79] will be analyzed. In fact, the 2009/2010 pandemic influenza, and four consecutive seasonal influenza datasets (2011/2012, 2012/2013, 2013/2014, 2014/2015) are all the datasets which will be the focus of our discussion throughout this chapter. For the 2011-2012 seasonal influenza, data is available between week 43 (2011) and week 35 (2012). For the next two seasons of influenza, data available is from week 40 of the starting year up to week 20 of the subsequent year, while for the latest seasonal influenza, data is available between week 41 (2014) and week 20 of the following year. The average number of GPs reporting the cases in 2011/2012 was 7.3, while for the other three consecutive years, the figures were 6.5, 5.9 and 5.7 respectively.

Details of the Maltese population as well as the 2009/2010 pandemic influenza are already described in chapters 2 and 3. Since the number of doctors reporting cases on a weekly basis is known, all data points were converted to an estimate of the total number of consultations and the total number of individuals with diagnosed ILI in Malta. Moreover, this takes into account that the total number of active GPs in Malta, which is around 300 (as described in chapter 2). Hence, the number of reported cases were multiplied by 300 and divided by the number of reporting doctors to get an estimate number of the total number of people consulting doctors and the number of people diagnosed with an ILI by all GPs in Malta. For the sake of consistency in this chapter, the pandemic data (2009/2010) was also converted to the total number of estimated consultation and diagnosed cases (Figure 4.1).

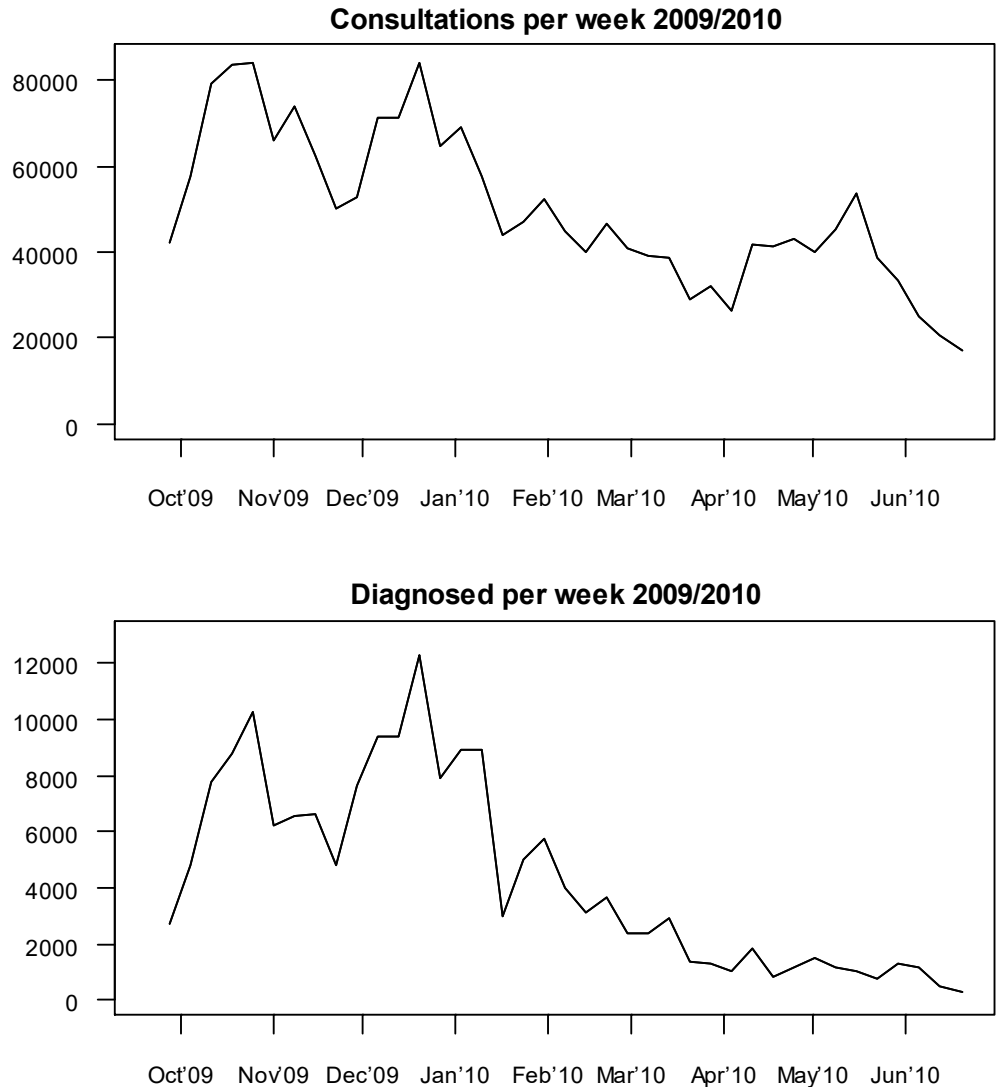


Figure 4.1 – Consultations and diagnosed charts during pandemic season (2009/2010) in Malta. The first chart represents the number of estimated weekly consultations in Malta and the second chart represents the number of estimated diagnosed ILI cases based on the GPs data.

Consultations data for the 2009/2010 period includes an estimated number of 1,950,600 consultations with a corresponding 170,400 ILI individuals (Figure 4.1). However, this was during the H1N1-influenza period; thus people were more wary about symptoms, resulting in a high number of consultations. As explained in the previous two chapters, for the pandemic period defined in figure 4.1, two high peaks were recorded (October'09 and December'09). These two peaks clearly show in the consultations and diagnosed datasets. Furthermore, unlike the diagnosed dataset, the consultations dataset recorded another lower peak during May 2010. However, this was recorded when the H1N1 virus was already considered inactive (Chapter 3). Following the H1N1 period, the number of consultations during the seasonal influenza period started to decrease rapidly. In relation

to this, it was estimated that during the 2011/2012 period, the number of consultations was equal to 1,640,991 consultations, followed by 1,182,374 during 2012/2013, 941,710 and 834,546 during the subsequent two years of seasonal influenza. Similarly, the number of ILI diagnosed individuals decreased quickly after the pandemic period. During the 2011/2012 period, 74,321 individuals were estimated to have been diagnosed of ILI, followed by 31,299 during 2012/2013, 15,450 and 31,514 ILI diagnosed individuals during the following two years.

The below plots (Figure 4.2) clearly show that for the consultations, data is relatively more stationary when compared with the diagnosed data, while the latter show a clear peak throughout the seasonal influenza period of each respective year. The consultation datasets vary between 20,000 and 50,000 individuals per week with a lot of fluctuations. There are clearly 2 groups of data. The first group of data consists of the 2011/2012 and 2012/2013 datasets, while the second group consists of the 2013/2014 and 2014/2015 datasets. For the first group of consultation datasets, minor peaks were recorded (early March 2012 and late January 2013), while for the second group of data no specific peaks can be observed. Some of this difference can be attributed to a higher number of influenza cases (which will be explored below) in 2011/2012 and 2012/2013 when compared with 2013/2014 and 2014/2015. The diagnosed cases vary between 0 and 7,000 cases. The peak is reached any time between February-April of each respective year. In general, the diagnosed datasets are rather stationary till around December. However, a sharp increase is registered during the beginning of January. A high number of diagnosed cases are recorded for around 3 months. By end of April, the number of diagnosed cases are at the same levels to the initial period (pre-January). The 2011/2012 dataset stands out, with a much higher number of diagnosed ILI cases. Hence, considering that the consultation datasets and the diagnosed cases have different characteristics, we aim to optimize the relationship between these two variables throughout this chapter.

One of our hypotheses is that the total consultations are a linear function of the diagnosed seasonal influenza cases. We will be testing this hypothesis throughout this chapter by comparing this linear association during different time periods of the four seasonal influenza outbreaks. Throughout this chapter, we will use the term ‘ratio’ to represent the

proportion of the number of diagnosed cases from the number of consultation cases. Hence, $^{diagnosed}/_{consultations}$ is a value between 0 and 1.

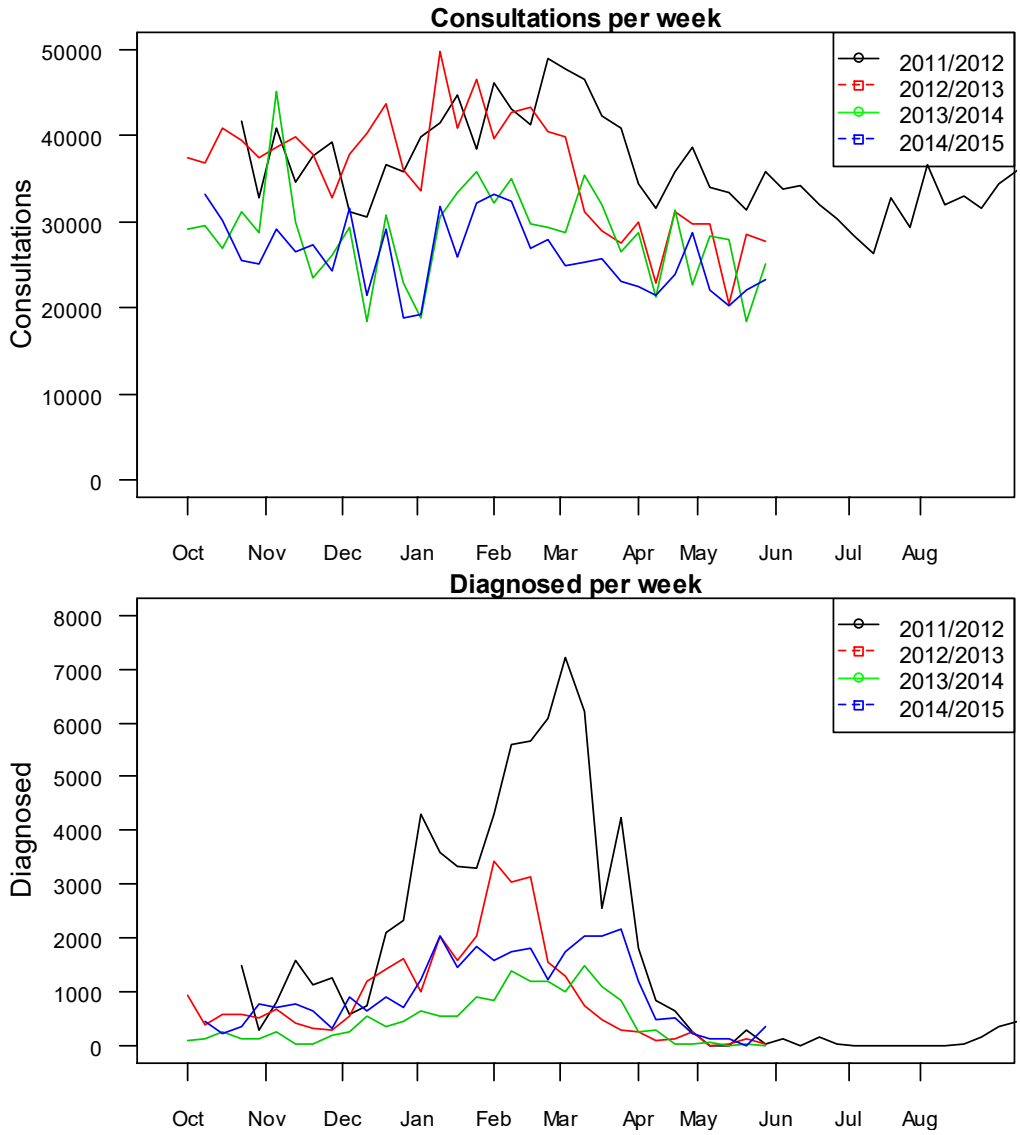


Figure 4.2 – Consultations and diagnosed charts during pandemic season (2009/2010) in Malta. The first chart represents the number of estimated weekly consultations in Malta and the second chart represents the number of estimated diagnosed ILI cases based on the GPs data.

4.3 Linear modelling of a relationship between diagnosed and consultations

Obtaining further understanding about the relationship between diagnosed and consultations will aid in the understanding of several epidemiological factors for the subsequent years. For the scope of the linear regression model, the consultation datasets are assumed as the dependent variables and the diagnosed datasets as the independent

variables. This format allows us to understand the 'background' consulting rate of the consultation datasets (non-influenza cases).

Similarly to what has been carried out previously (in Chapter 3), figure 4.3 shows the correlations between the consultations and diagnosed datasets (cf. Figure 3a in Chapter 3). For the 2013/2014 & 2014/2015 datasets, correlation gets weaker; in fact, the strongest correlation values are found in the oldest data (Table 4.1) which are being discussed in this chapter. Indeed, the highest Pearson-correlation value was found to be 0.897 for the pandemic data. The 2011/2012 data also shows a strong correlation value (r = 0.838) but weaker linear regression relationship than the previous dataset. As the number of influenza cases (and hence diagnosed individuals) decrease (2012/2013, 2013/2014), the correlation values drop. For the 2012/2013 datasets, the linear relationship is moderate (r = 0.685) and becomes weaker for the 2013/2014 influenza season (r = 0.308). A worse progression is noted for the 2014/2015 dataset (r = 0.235). These correlation values coincide with the R^2 – values when fitting a linear regression model, such that the R^2 – values decreased from one year to the other (0.806 (09/10), 0.702 (11/12), 0.469 (12/13), 0.095 (13/14) and 0.055 (14/15)).

The background level of consultations (non-influenza consultations) was estimated from the linear relationship for each individual dataset through the y-intercept. For all datasets, the background consultation rate varies between 24,000 cases up to 33,000 cases. The highest number of non-influenza consultations are found to be in the oldest three datasets due to a higher number of overall consultation cases.

The below results (Table 4.1) provide another important value Δ, within the linear regression equation $Consultations_t(c_t) = \kappa + \Delta * Diagnosed_t(d_t)$, as defined in chapter 2. For different datasets, this value varies between 1.8 and 5, which shows that the rate of increase of diagnosed seasonal influenza individuals varies according to dataset. The lowest Δ is for the latest dataset (2014/2015), where for every single diagnosed individual, on average the total consultations are increased by 1.8 cases. For the 2012/2013 dataset, for every single diagnosed individual, total consultations increase by around 5 cases (for detailed discussion, see below).

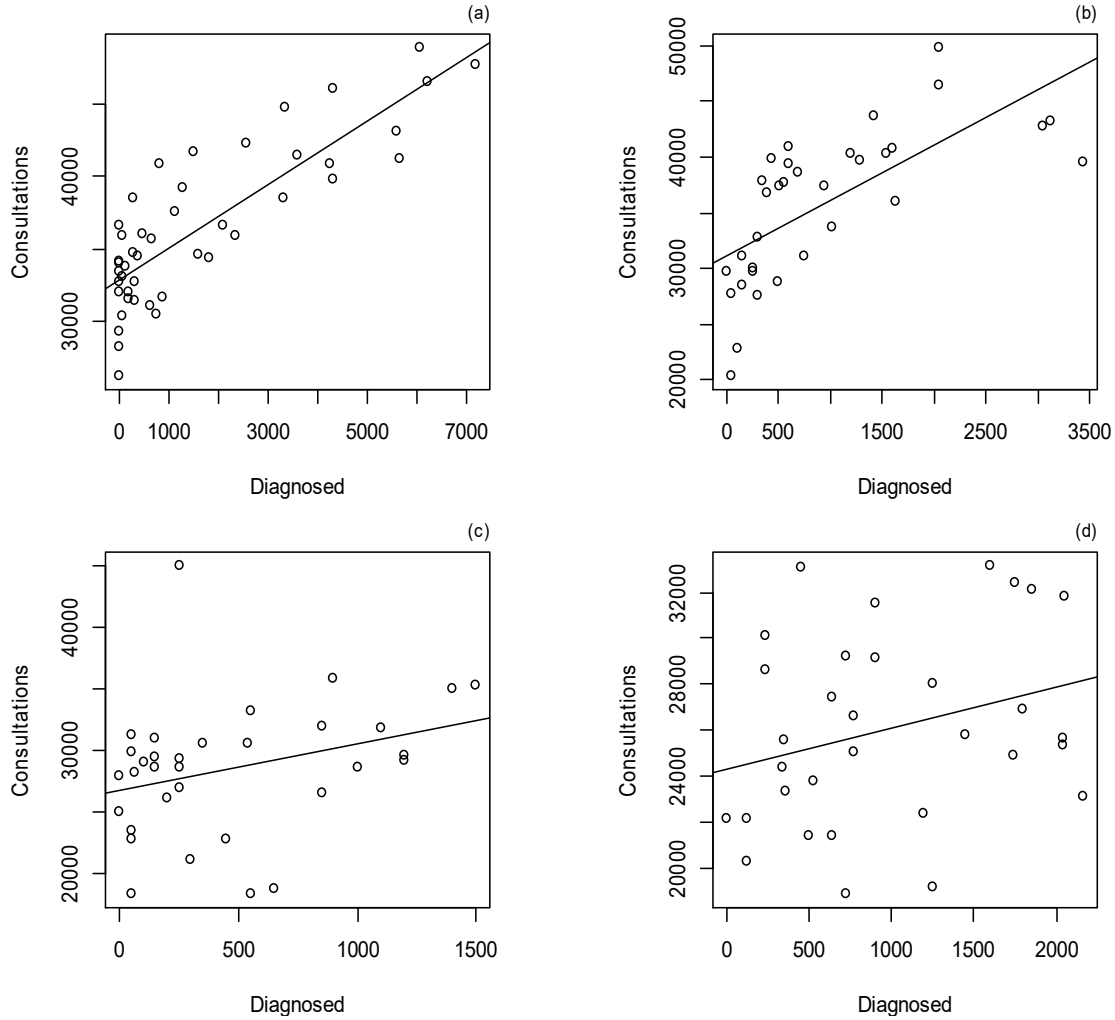


Figure 4.3 - Correlation plots between consultations and diagnosed where (a) is the 2011-2012 data, (b) 2012-2013 data, (c) 2013-2014 data and (d) 2014-2015 data. The straight lines in each plot corresponds to the regression line, which is the line of best fit between the two variables. All details for these plots are also found in table 4.1. Note the different horizontal and vertical scales in graphs due to different number of cases for individual seasons.

Furthermore, table 4.2 provides the confidence intervals for all individual parameters (κ and Δ). For the latest two datasets (2013/2014 and 2014/2015), the slope of the linear regression models provided wide confidence intervals (Table 4.2). In relation to this, these two predictors (diagnosed datasets) are not associated with significant changes in the response (consultations) variable (p-value > 0.05). Hence, H_0 is not rejected since no relationship exists between the consultations and diagnosed variables ($\Delta = 0$). All the other parameters for the other three influenza seasons were all proved to be good predictors (diagnosed ILIs) when compared to the response variable (Consultations).

Data	Pearson Correlation Value (r)	Linear Regression Equation	R^2 Value	Ratio (Diagnosed/ Consultations) Average
2009/2010	0.897	$c_t = 29210 + 4.762d_t$	0.806	0.077
2011/2012	0.838	$c_t = 32857 + 2.186d_t$	0.702	0.040
2012/2013	0.685	$c_t = 31103 + 4.983d_t$	0.469	0.024
2013/2014	0.308	$c_t = 26774 + 3.765d_t$	0.095	0.016
2014/2015	0.235	$c_t = 24332 + 1.774d_t$	0.055	0.038
All Data	**0.849**	$\boldsymbol{c_t = 27989 + 4.508d_t}$	**0.721**	**0.040**

Table 4.1 – Pearson Correlation Values and R^2 values for the relationship between consultations and diagnosed for five different years. The R^2 value was obtained through a linear regression model where c_t is the number of consultations at time t and d_t is the number of diagnosed individuals at time t. The 'Ratio' is the proportion of diagnosed cases from the consultation cases (defined above).

Data	Non-influenza Consultations	95% C.I.	Slope	95% C.I.
2009/2010	**29,210**	(25,074 - 33,346)	**4.762**	(4.009 - 5.515)
2011/2012	**32,857**	(31,738 - 33,976)	**2.186**	(1.761 - 2.611)
2012/2013	**31,103**	(28,647 - 33,559)	**4.983**	(3.118 - 6.848)
2013/2014	**26,774**	(24,141 - 29,407)	**3.765***	(-0.326 - 7.856)
2014/2015	**24,332**	(21,819 - 26,845)	**1.774***	(-0.343 - 3.891)

Table 4.2 – Error terms for the above parameter values for the linear regression models. (*) represents those parameter values which are not significantly different (p>0.05) (H_0 is not rejected when no relationship exists between the consultation and diagnosed ILI variables ($\Delta = 0$)); hence these predictors are not associated with changes in response. For all the other parameter values, the p-value is less than 0.05; hence this shows that we can reject the null hypothesis. Thus, these predictors are a meaningful addition to the above linear regression models (relationship exists between diagnosed and consultations). Changes in the predictor values are related to changes in the response variable.

All five datasets altogether provide a strong correlation (r=0.849) between consultations and diagnosed (Table 4.1). Furthermore, the regression model provides a satisfactory fit (R^2=0.721). The baseline non-influenza consultations (27,989) is in good agreement with figure 4.4. These results clearly show that there is a general strong relationship between the consultations and diagnosed ILI datasets. This relationship is universal across different seasons. Moreover, there is a significant number of points which relate to a lower number of consultations (18,000 – 28,000) and very low values of diagnosed ILI

cases (in some cases, 0 ILIs). Hence, this further confirms that low number of consultations correspond to non-influenza periods.

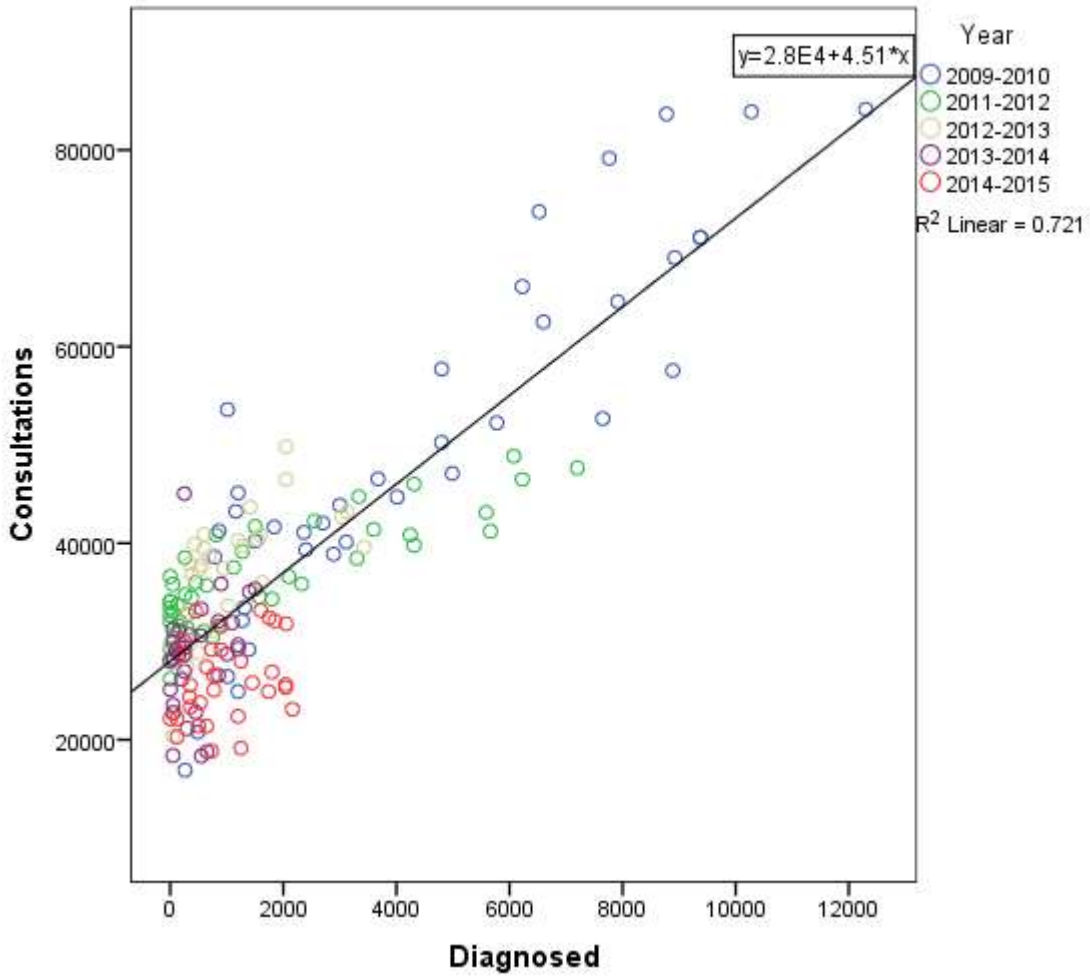


Figure 4.4 - Correlations of the 5 influenza periods combined. The straight line corresponds to the regression line, which is the line of best fit between the two variables. The accuracy of this model is 72.1%, hence the dependent variable can be predicted with this accuracy.

The regression models in table 4.1 provide a linear predictive technique between the consultation and diagnosed variables. In fact, these linear regression models were used to predict the consultation datasets (Figure 4.5). Hence the diagnosed dataset (independent variable) was used to predict the number of consultations at each individual time point. The linear regression models produced very accurate fits for the first two consultation datasets (2009/2010 and 2011/2012). However, by time the model fit started to get weaker. For the 2012/2013 dataset, the fit is rather reasonable, though for the latest two datasets (2013/2014 and 2014/2015) the linear regression models did not provide satisfactory predictions across the season.

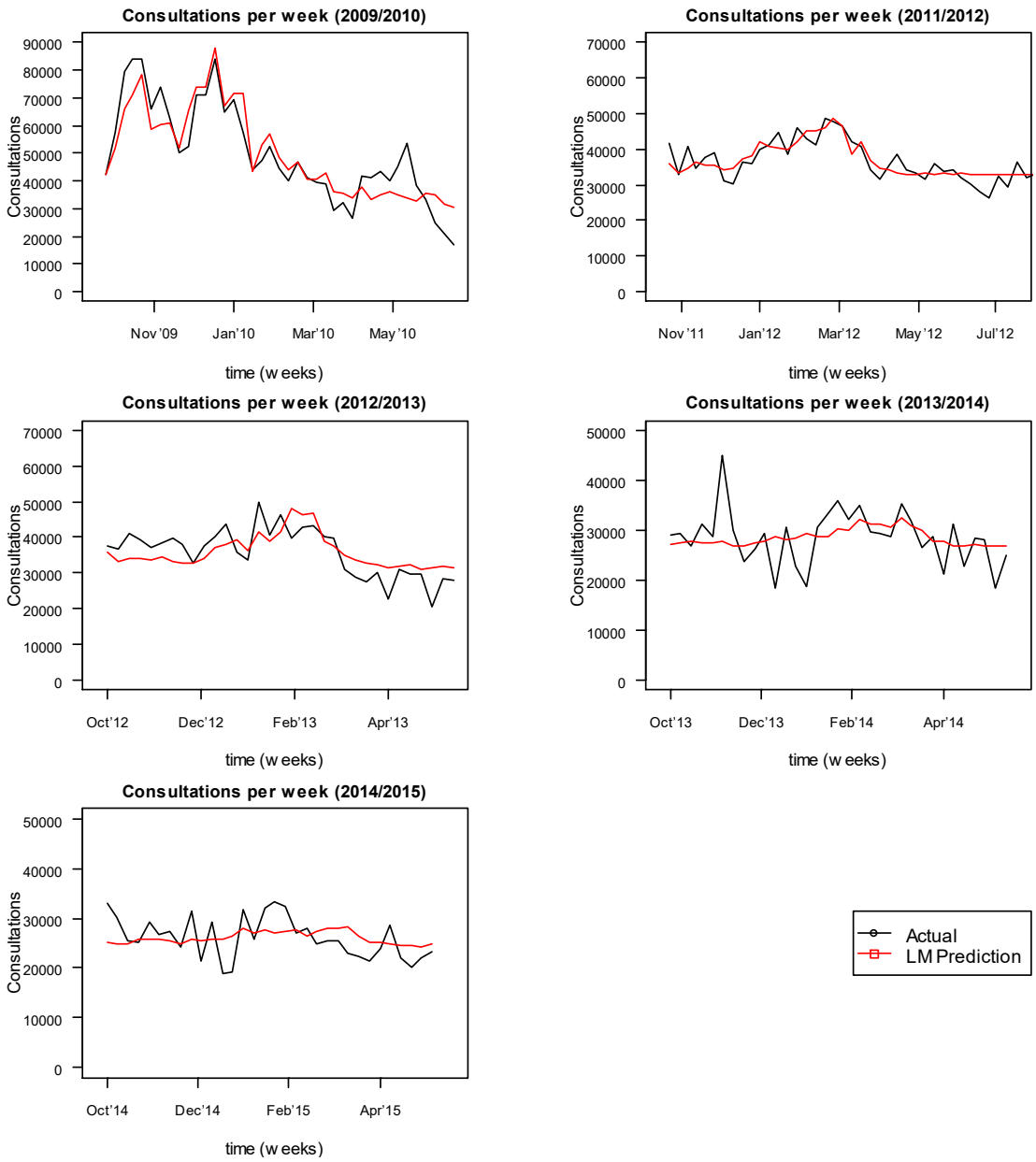


Figure 4.5 – This figure represents the linear model regression fit (defined in table 1) for all the five consultation datasets. The black lines represents the actual consultations data (GPs data) and the red lines corresponds to the fit produced through the linear regression model (obtained from table 4.1).

For the 2009/2010 consultations dataset, the linear regression model produced accurate peaks, except for the last and lowest peak (May'10). This can be attributed to the fact that for the diagnosed dataset, there were no further peaks from January 2010 onwards. To a certain extent, a smooth curve is produced for the 2011/2012 dataset with some minor oscillations. For the latter dataset, the major peak is predicted accurately with reasonable predictions for the other points. The 2012/2013 dataset produced a typical seasonal influenza wave, however missing the peak by few weeks. In fact, the peak of the produced prediction is around 3 weeks after the actual peak. For the remaining two consultation

datasets (2013/2014 and 2014/2015), the model fit is almost a straight line throughout the whole outbreak. Hence, no specific oscillations are detected when compared with the actual consultations data. This result coincides with the fact that for these last two datasets, the R^2 values are very low and that the independent variable is not a significant predictor (as defined above).

Throughout the next section, the SEIR model (as defined in chapter 2) was applied on all the four seasonal influenza datasets. Hence, through the use of the SEIR model, we aim to re-construct the above datasets. Following this, section 4.5 combines the linear regression model and the SEIR model into one joint model. Subsequently, results are analysed in light of the three different frameworks.

4.4 The SEIR model

Throughout this section, we used the particle filter algorithm and SEIR model to reproduce the seasonal influenza datasets (consultations and diagnosed). All the details of this model are provided in chapter 2. The prior distributions were mainly based on priors used in chapter 3 [79]. The prior distribution of the background rate (ϕ) for the consultations was set differently according to the year of the outbreak, due to a higher number of non-influenza illness. In fact, based on the linear relationship (as defined above) between the consultations and diagnosed, the baseline number of non-influenza cases (defined in chapter 2) was established for each individual year. For the 2011/2012 dataset the prior for (ϕ) was set to $N^{+}(750,300)$, for the 2012/2013 dataset $N^{+}(665,300)$, for the 2013/2014 dataset $N^{+}(530,300)$ and for the 2014/2015 dataset $N^{+}(420,250)$. For all the other diagnosed datasets the prior was set to $N^{+}(1,0.25)$ (same as Chapter 3). The prior distributions for the state $\Sigma_0 = \{S_0, E_0, I_0, R_0\}$ values were based on the priors defined in chapter 3 [79]. For all datasets, a series of 20,000 particles were used. The full algorithm and the R-language script is presented in Appendix C.

The particle filter algorithm [79] applied through the SEIR model and the observation model D_t provides a satisfactory fit for all the seasonal influenza datasets (Figure 4.6). In this case, the datasets are fitted individually with their own related parameters only. Hence, the relationship between the consultation and diagnosed datasets is not being used in any way. For the first two consultation datasets (2011/2012 and 2012/2013), the model

fits are very accurate, including good predictions of the oscillations. This same result also applies for the pandemic data (cf. Figure 4a in Chapter 3). For the latest two consultations datasets (2013/2014 and 2014/2015), the model fits are reasonable; however, some oscillations are not matched accurately between the actual data and predicted data. In fact, by late March 2014, the model predicted a peak, while actually this never happened. Similarly to the 2014/2015 dataset, the model predicted a peak in December 2014, though this never happened. For the diagnosed datasets, the features of the data are well represented, including all respective seasonal influenza peaks (same result applied to the pandemic data, cf. Figure 4b, Chapter 3).

Parameter posterior estimates were obtained through the same particle filter algorithm and SEIR model (Tables 4.3-4.4) with posteriors varying widely between consultation and diagnosed datasets. Therefore, one cannot use consultation estimates to directly measure the actual spread of seasonal influenza. However, there exists other relationships, as mentioned above and below that can provide further insights between the two.

The reporting rates, δ, (0.65-0.69) are relatively consistent for the consultations group when compared with all the different years. This latter result is similar to the diagnosed data (0.23-0.29), although not coherent with the 2009/2010 pandemic data (0.60). Parameter estimates for the diagnosed data are less spread when compared to the consultation datasets. In relation to this, the latent period, α, (0.01-0.06), the infectious period, τ, (0.33-0.47) and background rate per doctor, ϕ, (0.78-0.90) are all closely related. For the same parameter values for the consultations data, estimates vary broadly and hence one cannot draw any further results. The infectious period for the seasonal influenza data is estimated to be around 2.8 days (Table 4.4) which is slightly less when compared with the estimated values in chapter 2 (3.5 days). Moreover, table 4.4 shows that the higher the number of diagnosed ILI cases, the higher the infection rate. For example, the lowest number of diagnosed cases (15,450, as defined above) was registered in 2013/2014, and so was the infection rate (0.48) for 2013/2014 when compared with the other datasets.

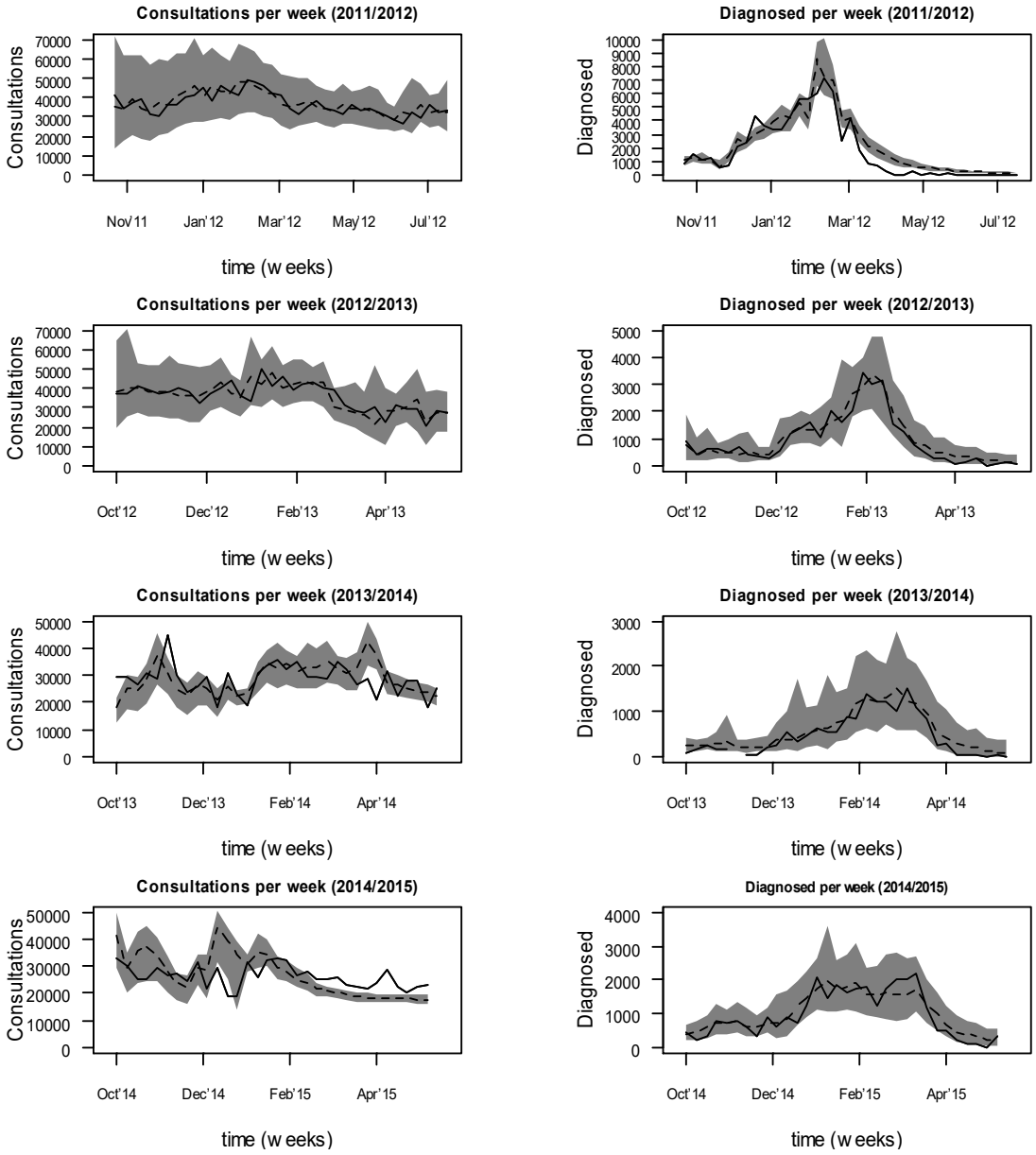


Figure 4.6–Comparison of weekly consultations (1st column) and weekly diagnosed (2nd column) for all the four seasonal influenza datasets. Data include the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region). The datasets were fitted individually, with their own related parameters only. Hence, the relationship between the consultations and diagnosed datasets is not being used in any way.

Definitions	Parameter	Consultations 2011/2012 (20,000 Particles)	Consultations 2012/2013 (20,000 Particles)	Consultations 2013/2014 (20,000 Particles)	Consultations 2014/2015 (20,000 Particles)
Infection Rate ($week^{-1}$)	β	0.55 (0.22-0.89)	0.84 (0.37-1.31)	0.47 (0.45-0.50)	2.92 (2.71-3.13)
Importation rate ($week^{-1}$)	ε	475.06 (-871.43-1821.55)	43.02 (-1.34-87.39)	235.18 (102.97-367.40)	22.85 (17.05-28.66)
Latent period (week)	α	2.48 (0.45-4.51)	1.83 (-0.35-4.01)	0.40 (0.29-0.50)	0.53 (0.51-0.55)
Infectious Period (week)	τ	2.95 (0.27-5.62)	1.56 (93.21-164.55)	8.53 (7.14-9.91)	0.08 (0.07-0.10)
Background rate ($week^{-1}$)	ϕ	138.53 (49.68-227.39)	128.88 (93.21-164.55)	21.83 (19.74-23.93)	86.06 (80.48-91.64)
Reporting rate	δ	0.65 (0.57-0.74)	0.66 (0.48-0.83)	0.66 (0.60-0.72)	0.69 (0.65-0.73)

Table 4.3– Posterior parameter values estimated for different weekly consultation datasets. Numbers in brackets represent highest density 95% symmetric credible intervals based on a normal approximation to posterior distributions.

Definitions	Parameter	Diagnosed 2011/2012 (20,000 Particles)	Diagnosed 2012/2013 (20,000 Particles)	Diagnosed 2013/2014 (20,000 Particles)	Diagnosed 2014/2015 (20,000 Particles)
Infection Rate ($week^{-1}$)	β	1.18 (1.11–1.26)	0.64 (0.43-0.86)	0.48 (0.28-0.69)	0.53 (0.29-0.77)
Importation rate ($week^{-1}$)	ε	35.16 (18.36-51.96)	299.02 (-46.21-644.25)	107.38 (-62.23-277.00)	102.76 (-55.66-261.19)
Latent period (week)	α	0.03 (0.02-0.04)	0.06 (-0.04-0.16)	0.04 (-0.04-0.13)	0.01 (-0.01-0.03)
Infectious Period (week)	τ	0.33 (0.29-0.38)	0.39 (0.12-0.65)	0.42 (0.07-0.77)	0.47 (0.16-0.77)
Background rate ($week^{-1}$)	ϕ	0.90 (0.79-1.01)	0.83 (0.45-1.21)	0.78 (0.37-1.20)	0.83 (0.41-1.25)
Reporting rate	δ	0.29 (0.19-0.38)	0.25 (0.03-0.46)	0.23 (-0.07-0.61)	0.25 (0.05-0.90)

Table 4.4– Posterior parameter values estimated for different weekly diagnosed datasets. Numbers in brackets represent highest density 95% symmetric credible intervals based on a normal approximation to posterior distributions.

The R_t values (Figure 4.7) for the consultations have a similar trend between each other, although they vary from the diagnosed R_t values. All R_t plots start with a high value. This feature was also observed in chapter 3 (see Figure 6, Chapter 3), and will be analyzed in further detail in chapter 6. For the 2011/2012 dataset, the first R_t value appears later than the other first R_t values for other seasons, since the data started to be collected at a later time when compared to the other years (Figure 4.2). The initial high value is followed by a sharp drop for both types (consultations and diagnosed) of datasets.

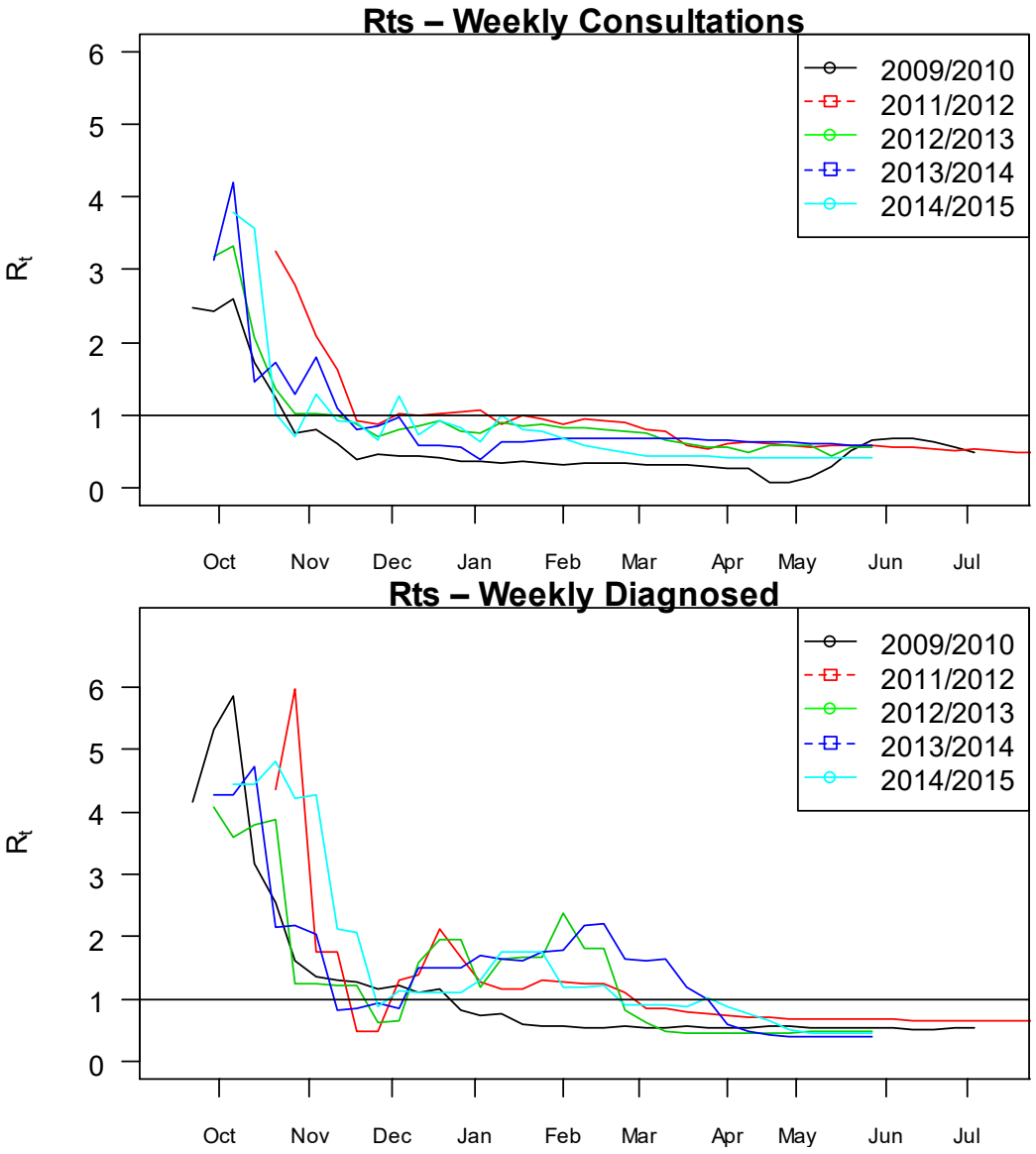


Figure 4.7–Estimation of the effective reproduction ratio at any given point of the epidemic for different datasets, including weekly consultations and diagnosed for all five datasets. The horizontal line corresponds to $R_t=1$, an invasion threshold.

For the consultations data, following December, the R_t values almost always remain under the value of 1 for all influenza seasons. On the other hand, for the diagnosed data, in general, during November the R_t value is under 1, but then it rise above 1 between December to March. The R_t values reach almost the value of 2 for all individual diagnosed datasets and then declines during the following weeks. For different datasets, the peak R_t value is reached during different periods. This difference can be attributed due to different influenza peaks for different influenza seasons. Although the R_t values for different datasets has the same theoretical meaning, the R_t value has a different level of accuracy

for different proxies (as described in Chapter 2). Note that the R_t values for the pandemic season are notably different when compared with the seasonal influenza datasets. The characteristics of the pandemic dataset are different when compared with the seasonal influenza datasets (as defined above).

4.5 Combining the SEIR and Linear regression model in one single framework (joint model)

The scope of this section is to extend the analysis of the previous two sections and chapter 3. Throughout this section, we aim to extend the model used in the previous analysis by incorporating different datasets together by attempting to refine and extend the prediction of the outbreak within a single framework. In fact, the ultimate aim is to use the relationship between the number of consultations and diagnosed ILIs to predict both outcomes during the same model run. In order to do this, I combined the SEIR model together with the linear regression model in one single joint model. The main question here is whether the joint model can improve the predictions when compared with the SEIR model and the linear regression model.

In order to combine two datasets, the same SEIR model (as defined in the previous section) was used. All the same details (as defined previously) related to the prior information, the particle filter algorithm and the SEIR model were adopted. However, several amendments in the R-code were carried out in order to calculate the number of reported consultations and the number of diagnosed ILI individuals during the same model run. The code was amended in a way to have one variable modelled through the SEIR model and the other variable through the linear regression model during the same run. Hence, the parameters of the linear regression model were updated during the particle filtering process, allowing the parameters of the linear regression model to be updated during every single time point (Appendix D – highlighted in red). All the details related to these amendments are defined below. Furthermore, the model script was adjusted to produce two outputs, i.e. the diagnosed and consultation predictions during the same model run.

In relation to the analysis carried out in the previous sections, it was established that for certain datasets there is a certain good level of linear relationship between the

consultation and diagnosed variables. For some datasets (2009/2010 and 2011/2012), this was also found to be a strong relationship. However, in this section our aim is to update the parameters of the linear regression model at each different time point, based on all the known information at the time point of analysis. The following time-dependent linear regression model was incorporated with the SEIR model:

$$c = \kappa_t + \Delta_t d_t$$

where κ_t is the parameter which refers to the y-intercept of the linear regression model and is dependent on time *t*, and Δ_t is equal to the slope of the linear regression model and is also dependent on time *t*. Detailed analysis about these two parameters were provided in chapter 2 and section 4.2. However, while these parameters were previously fixed during the whole process, in this case, the parameters are dependent on time. Hence, these will be updated at every single particle filtering iteration. For the above linear regression model, the parameters between the consultation and diagnosed variables will be updated at each different time point using all the previous known data points. Then, the number of consultations at $t_{n+1}$was estimated using the parameters obtained at t_n. Hence, based on the actual data points at t_n, the prediction of the number of ILI diagnosed individuals at time t_{n+1} is achieved using the SEIR model, while the number of consultations at time t_{n+1} is obtained using the above linear regression model at time t_n.

The first time point, where the weekly consultations were possible to be predicted, was from time t_2 (hence, two known weekly data points). At time t_1 (one known weekly data point) there is not enough data to estimate the parameters of the linear relationship between the two sets of data. Hence, this is a limitation for the above method, although in epidemiological studies, decisions and strategies are not based on the first data point of the outbreak.

Throughout the next paragraphs we will be looking at the main results related to the above model fit, including the time dependent parameter values of the linear relationship between the consultations and diagnosed, and the prediction plots. Furthermore, the pandemic 2009/2010 data and all the other seasonal influenza datasets will be used for the scope of this analysis.

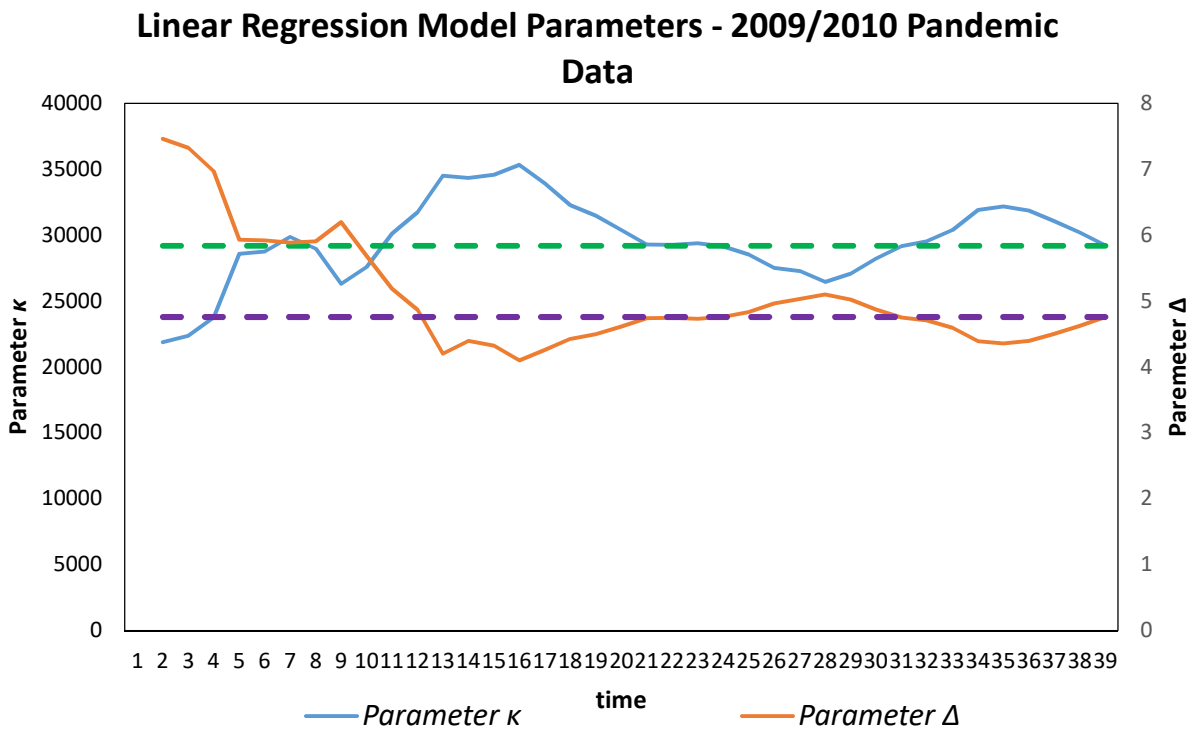


Figure 4.8 –Parameter values for the linear regression between the weekly consultations (dependent variable) and the weekly diagnosed (independent variable) of the 2009/2010 pandemic outbreak. These parameters were updated at each individual time point during the course of the outbreak. The green dashed line is the general parameter κ for the above 2009/2010 linear regression model (Table 4.1) and violet dashed line is the general parameter Δ for the above 2009/2010 linear regression model (Table 4.1).

Figure 4.8 shows the parameter values (κ_t and Δ_t) as defined above for the 2009/2010 pandemic data. For the first time point, there was insufficient data to calculate any parameter values, as defined above. Although the initial parameter values tend to be slightly inconsistent, after a period of time the parameters tend to stabilize. In fact, κ varies between 26,000 and 34,000, while Δ varies between 4 and 5. The variations are consistent with the confidence interval found in table 4.2. In general, even when plotting the parameter values for the other datasets, the same trends apply (Appendix E). Hence from early stages the parameter values for the relationship between diagnosed and consultations tend to stabilize. Due to few data points, the R^2 value starts with values which are close to 1 and then tends to vary between 0.6 and 0.9 (Figure 4.9).

As mentioned above, the diagnosed datasets were modelled through the joint model, however using the SEIR technique (Figure 4.6 – Diagnosed datasets and Figure 4 – Chapter 3). Hence, the results of the model fit for the diagnosed datasets are the same as

described in the previous section and chapter 3, while, the consultations model fit was obtained from the time dependent linear regression model as defined above. Hence, the new predicted diagnosed dataset (through the SEIR model) was used to predict the consultations data (through the above time dependent linear regression model), thus producing two outputs by using two different methods in one single framework (joint model).

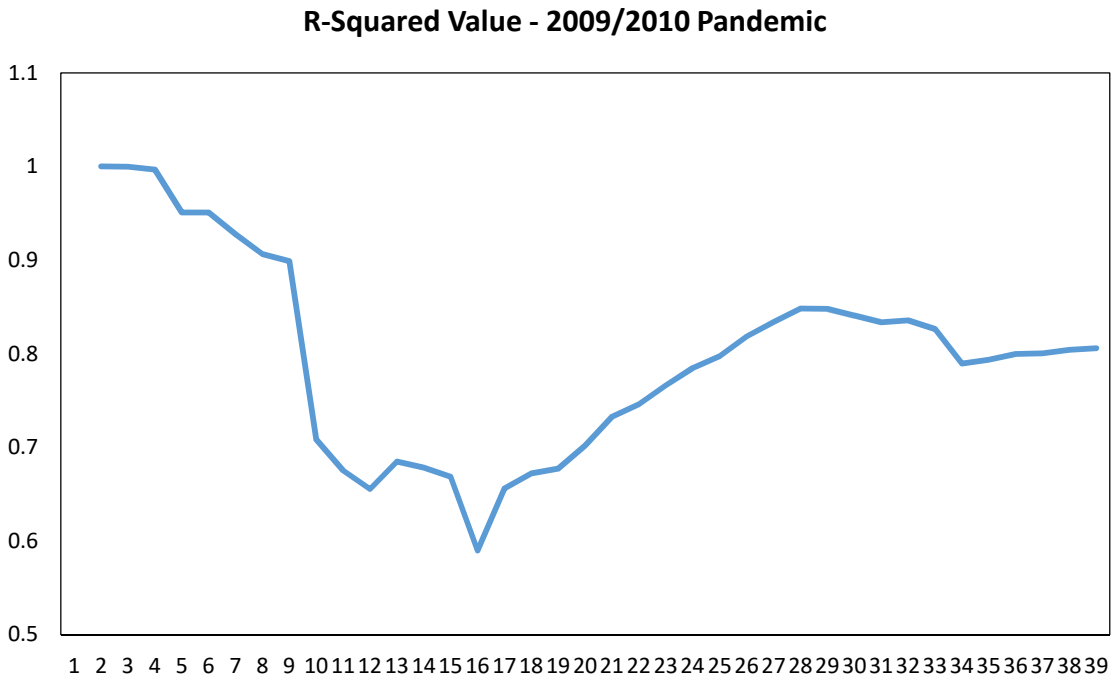


Figure 4.9 – R^2 values for the linear relationship between the weekly consultations and the weekly diagnosed datasets, of the 2009/2010 pandemic outbreak. The R^2 value was updated at each individual time point during the course of the outbreak.

Unlike the linear regression model (Section 4.3), the parameters of the linear regression model were allowed to be updated at each time point, based on the SEIR model fit of the diagnosed datasets. Hence, the parameters are time dependent. The joint model produced improved prediction charts for the consultations data, when compared with the constant (time independent) linear regression model technique (Figure 4.5). This improvement can be easily seen for the 2013/2014 and 2014/2015 datasets. For the constant linear regression model (Figure 4.5) the predictions are flat for the latter two datasets, while for the joint model the fit improved substantially. In fact, the time independent linear regression model fit produced a stationary line with few oscillations for the last two datasets (2013/2014 and 2014/2015). On the other hand, the joint model produced accurate predictions with corresponding oscillations to the actual dataset. Although the SEIR model fit for the consultations data (Figure 4.6) produced more accurate

predictions, for the joint model we are producing two outputs in one single framework, including two modelling techniques.

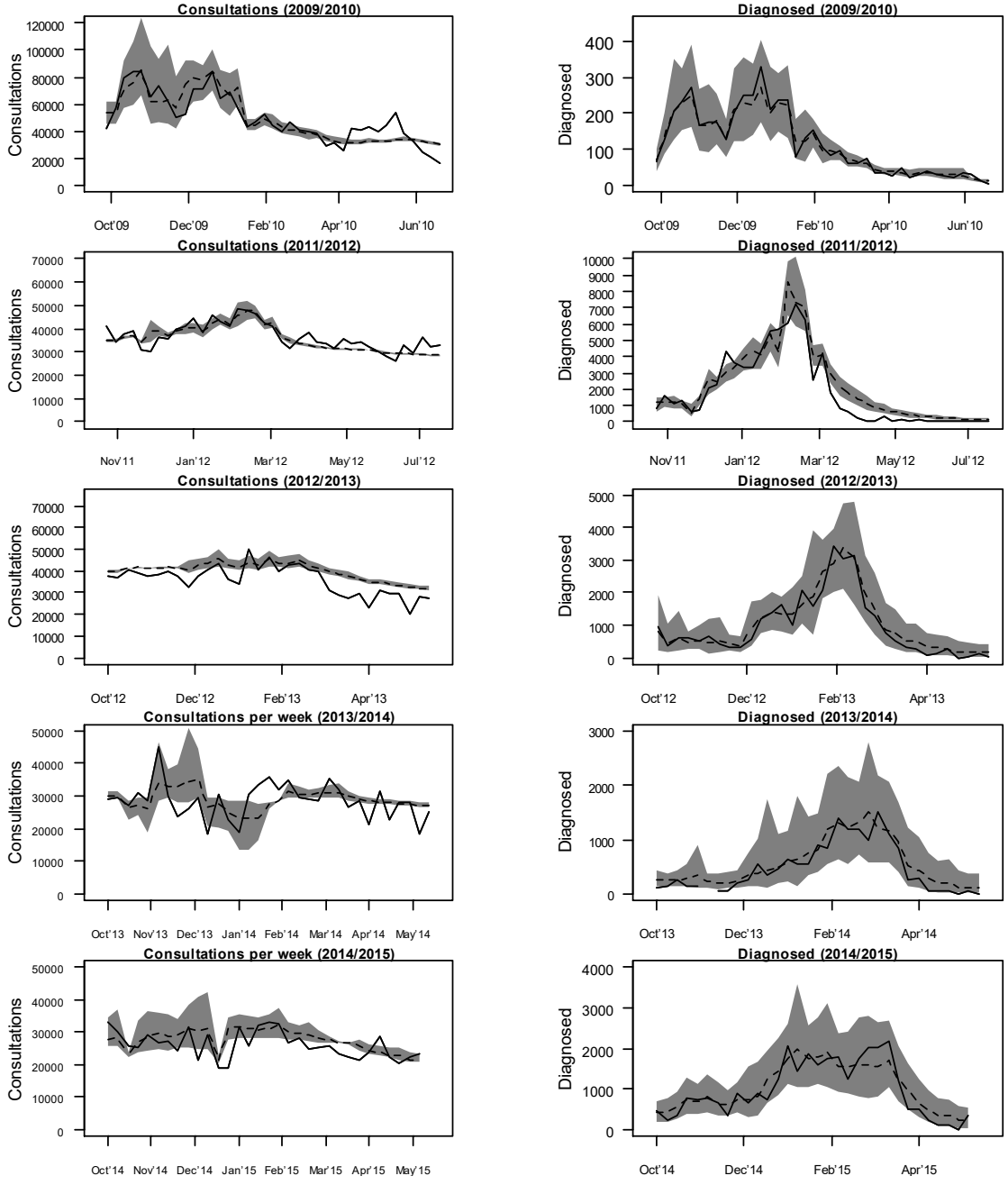


Figure 4.10 – Comparison of weekly consultations (1st column) and weekly diagnosed (2nd column) for all the five outbreak datasets. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region). All charts were plotted from the joint model. However, the diagnosed datasets were fitted through the normal SEIR model parameters, while the consultations datasets were fitted through the time-dependent linear regression model, as defined above.

Such technique (joint model) can possibly be useful when carrying out future (real-time) predictions, thereby using a more reliable dataset (diagnosed) to predict the number of future consultations cases (to be discussed in Chapter 5). For a limited number of

consultation predictions for each individual dataset, the joint model produced narrower (than expected) confidence intervals when compared with other predictions. This is attributed to a zero gradient of the time-dependent linear regression model. Since the consultations is being constructed from the diagnosed data, then this will end up with narrow confidence intervals for these particular points (with zero gradient).

4.5 Discussion

As seen in chapter 3, consultations are largely flat, possibly with some peak between January and May which can be associated with the influenza. The average consultation numbers are generally consistent, although 2011/2012 and 2012/2013 are higher than 2013/2014 and 2014/2015 with a general excess of about 30,000 individuals per week. Compared with the pandemic data, the excess in the number of consultations during the 2009/2010 season was higher, especially during the early stages of the influenza. People were quite wary during that period as H1N1 was an international concern, hence more people were inquisitive about this outbreak. By time the excess stabilized to the same levels of other influenza datasets.

In contrast to the pandemic data, the seasonal influenza datasets show a clear peak between February to April period. For the pandemic data, the first peak was reached during an unusual period of the year, July 2009, followed by October 2009 and December 2009 peaks. Following the pandemic data, the 2011/2012 season registered the highest number of diagnosed cases and then the numbers successively decreased.

Seasonal influenza undergoes a number of changes throughout the whole wave and hence, one needs to give the required attention to understand clearly the underlying results. The initial ‘early’ stage of the influenza represents the period where the number of diagnosed cases are flat and stationary. The ‘mid’ part of the influenza represents the period where the influenza starts to increase, reaches its peak and declines again. The ‘late’ part of the influenza corresponds to the end stages of the influenza where it has the same characteristics of the ‘early’ stage of the influenza (flat and stationary). Unfortunately, as discussed in the literature review, under-reporting in epidemiological studies exists, and thus maximizing the understanding and the information available of the seasonal influenza is of utmost importance.

For most epidemiological and seasonal influenza models, the early part of the influenza incorporates a low number of infected cases followed by the peak of the influenza and then a decline where the spread of the influenza dies out. This proves to be an important point when analyzing the relationship between consultations and diagnosed. In tables 4.5 – 4.10, we show the relevance of different periods within the whole influenza season. Hence, all datasets were defined in three different periods, based on the above definitions (early, mid and late). For the purpose of analysing the 2009/2010 data, 'early' influenza means between weeks 39/2009 to 46/2009, 'mid' part of the influenza between weeks 47/2009 to 13/2010 and 'late' influenza refers to the weeks 14/2010 to 25/2010. For all the other datasets, 'early' influenza refers to week 40 to week 50 (if week 40 is not available, the first available point is considered), 'mid' part of the influenza is between week 51 and week 13 of the following year, while 'late' influenza refers to week 14 up till any known weekly data point. The pandemic period was defined differently as the characteristics of this particular influenza vary from the other influenza seasons (as defined above).

Relationship between consultations and diagnosed proves to be stronger during the mid-part of the influenza period. In fact, the Pearson-correlation (r) and R^2 values are much higher during the mid-part of the influenza when compared with the early and late stages of the influenza seasons (Tables 4.5 – 4.10). This result applies for all the five datasets. R^2 values for three (2009/2010, 2011/2012 and 2012/2013) of the five datasets are higher than 0.5 during the mid-part of the influenza season, while only in one other period for one dataset is this value exceeded (2009/2010 – Early Period). In fact, the R^2 value for the early stage of the pandemic season is 0.906. This is substantially higher when compared with all the early R^2 values of the other seasonal influenza datasets (R^2 – 0.318 (2011/2012), R^2 – 0.221 (2012/2013), R^2 – 0.075 (2013/2014) and R^2 – 0.0002 (2014/2015)). The early high R^2 value for the pandemic season can be attributed to the early peak (as defined above) of the diagnosed ILI cases. As discussed above, this is associated with a high number of consultation cases during the early period for the same season.

For the mid-part of the latest two datasets (2013/2014 and 2014/2015), the R^2 values are below 0.2, hence resulting in a weak linear association. Such a weak association is mainly attributed to flat and stationary number of consultations for these two datasets (further

information below). For the first three datasets (2009/2010, 2011/2012 and 2012/2013), the early influenza period carries a stronger relationship between consultations and diagnosed datasets when compared with the late influenza period. However, for the latest two datasets (2013/2014 and 2014/2015), the early influenza period has a weaker relationship when compared with the late influenza period.

Early Influenza Data	Pearson Correlation Value (r)	Linear Regression Equation	R^2 Value	Ratio (Diagnosed/ Consultations) Average
2009/2010	0.952	$c_t = 28990 + 5.906d_t$	0.906	0.094
2011/2012	0.564	$c_t = 30508 + 5.530d_t$	0.318	0.027
2012/2013	0.470	$c_t = 35815 + 3.833d_t$	0.221	0.016
2013/2014	-0.274	$c_t = 31379 - 12.800d_t$	0.075	0.008
2014/2015	0.013	$c_t = 27300 + 0.214d_t$	0.0002	0.022

Table 4.5 – Pearson Correlation values (r) and R^2 values for the relationship between consultations and diagnosed for five different years for the early period. The R^2 value was obtained from a linear regression model. c_t is the number of consultations at time t and d_t is the number of diagnosed individuals at time t. The 'Ratio' is the proportion of diagnosed cases from the consultation cases (defined above).

Early Influenza Data	Non-influenza Consultations	95% C.I.	Slope	95% C.I.
2009/2010	**28,990**	(18,281 – 39,699)	**5.906**	(4.388 – 7.424)
2011/2012	**30,508**	(23,517 – 37,499)	**5.530** *	(-0.956 – 12.016)
2012/2013	**35,815**	(32,760 – 38,870)	**3.833** *	(-0.873 – 8.539)
2013/2014	**31,379**	(24,463 – 38,295)	**12.800** *	(-42.102 – 16.502)
2014/2015	**27,300**	(20,397 – 34,203)	**0.214** *	(-10.901 – 11.329)

Table 4.6 – Error terms for the above parameter values for the linear regression models. (*) represents those parameter values which are not significantly different (p>0.05) (H_0 is not rejected when no relationship exists between the consultation and diagnosed ILI variables ($\Delta = 0$)); hence these predictors are not associated with changes in response. For all the other parameter values, the p-value is less than 0.05; hence this shows that we can reject the null hypothesis. Thus, these predictors are a meaningful addition to the above linear regression models (relationship exists between diagnosed and consultations). Changes in the predictor values are related to changes in the response variable.

Mid Influenza Data	Pearson Correlation Value(r)	Linear Regression Equation	R^2 Value	Ratio (Diagnosed/ Consultations) Average
2009/2010	0.969	$c_t = 27200 + 4.407d_t$	0.938	0.099
2011/2012	0.795	$c_t = 33120 + 2.089d_t$	0.632	0.098
2012/2013	0.715	$c_t = 30518 + 4.818d_t$	0.511	0.042
2013/2014	0.437	$c_t = 25428 + 5.295d_t$	0.191	0.030
2014/2015	0.365	$c_t = 20546 + 3.798d_t$	0.134	0.061

Table 4.7 – Pearson Correlation values (r) and R^2 values for the relationship between consultations and diagnosed for five different years for the mid period. The R^2 value was obtained from a linear regression model. c_t is the number of consultations at time t and d_t is the number of diagnosed individuals at time t. The 'Ratio' is the proportion of diagnosed cases from the consultation cases (defined above).

Mid Influenza Data	Non-influenza Consultations	95% C.I.	Slope	95% C.I.
2009/2010	**27,200**	(23,797 – 30,603)	**4.407**	(3.869 – 4.945)
2011/2012	**33,120**	(29,237 – 37,003)	**2.089**	(1.222 – 2.956)
2012/2013	**30,518**	(25,729 – 35,307)	**4.818**	(2.258 – 7.378)
2013/2014	**25,428**	(19,925 – 30,931)	**5.295** *	(-0.626 – 11.216)
2014/2015	**20,546**	(11,903 – 29,189)	**3.798** *	(-1.461 – 9.057)

Table 4.8 – Error terms for the above parameter values for the linear regression models. (*) represents those parameter values which are not significantly different (p>0.05) (H_0 is not rejected when no relationship exists between the consultation and diagnosed ILI variables ($\Delta = 0$)); hence these predictors are not associated with changes in response. For all the other parameter values, the p-value is less than 0.05; hence this shows that we can reject the null hypothesis. Thus, these predictors are a meaningful addition to the above linear regression models (relationship exists between diagnosed and consultations). Changes in the predictor values are related to changes in the response variable.

The 2013/2014 and 2014/2015 data registered lower values of consultations, indicating that such low values are more likely to provide a weak signal for the seasonal influenza cases. In the above tables, we show that a ratio (defined above) higher than 4% between diagnosed and consultations provided an R^2 value higher than 0.5 together with a strong correlation (r>0.71) value, with the exception of one particular period (2014/2015 mid-part of the influenza, R^2=0.134 and r=0.365).

Late Influenza Data	Pearson Correlation Value (r)	Linear Regression Equation	R^2 Value	Ratio (Consultations/ Diagnosed) - Average
2009/2010	0.538	$c_t = 20783 + 13.980d_t$	0.290	0.030
2011/2012	0.255	$c_t = 32438 + 3.130d_t$	0.065	0.005
2012/2013	0.346	$c_t = 25359 + 16.520d_t$	0.120	0.004
2013/2014	-0.387	$c_t = 26221 - 16.930d_t$	0.150	0.004
2014/2015	0.142	$c_t = 22595 + 1.893d_t$	0.020	0.017

Table 4.9 – Pearson Correlation values (r) and R^2 values for the relationship between consultations and diagnosed for five different years for the late period. The R^2 value was obtained from a linear regression model. c_t is the number of consultations at time t and d_t is the number of diagnosed individuals at time t. The 'Ratio' is the proportion of diagnosed cases from the consultation cases (defined above).

Late Influenza Data	Non-influenza Consultations	95% C.I.	Slope	95% C.I.
2009/2010	**20,783**	(5,462 – 36,104)	**13.980 ***	(0.417 – 27.543)
2011/2012	**32,438**	(30,949 – 33,927)	**3.130 ***	(-2.078 – 8.338)
2012/2013	**35,815**	(30,628 – 41,002)	**16.520 ***	(-22.680 – 55.720)
2013/2014	**31,379**	(27,135 – 35,623)	**16.930 ***	(-18.389 – 52.249)
2014/2015	**27,300**	(23,527 – 31,073)	**1.893 ***	(-9.698 – 13.484)

Table 4.10 – Error terms for the above parameter values for the linear regression models. (*) represents those parameter values which are not significantly different (p>0.05) (H_0 is not rejected when no relationship exists between the consultation and diagnosed ILI variables ($\Delta = 0$)); hence these predictors are not associated with changes in response. For all the other parameter values, the p-value is less than 0.05; hence this shows that we can reject the null hypothesis. Thus, these predictors are a meaningful addition to the above linear regression models (relationship exists between diagnosed and consultations). Changes in the predictor values are related to changes in the response variable.

Otherwise, any other ratio less than 4% provided a weak R^2 value (< 0.32). Furthermore, in almost all cases, such low ratio values provided weak/moderate correlation values with the exception of the 2012/2013 data which provided a Pearson correlation value of 0.685 (Table 4.1). Elsewhere, all correlation values are less than 0.564, with a substantial number of correlation values showing very weak relationship. For ratios higher than 4% and strong correlation values between diagnosed and consultation datasets, this provided a baseline of non-influenza consultations between 27,000 and 33,120 cases. Otherwise, for other combinations of ratios and correlation values, baseline non-influenza

consultations do not follow any particular trend. In relation to this, we show that values range widely between 20,000 up to 36,000 non-influenza consultation cases. The lowest ratios between diagnosed and consultations is during late influenza period for all the 5 different datasets. During the mid-part of the influenza period, the highest Pearson correlation values and R^2 values were registered. This also coincides with a higher ratio between diagnosed and consultation datasets. In general, for early and late influenza periods, most of the latter values (r and R^2) are lower when compared with the mid-part of the influenza season. The only exception is for the 2009/2010 data as discussed above.

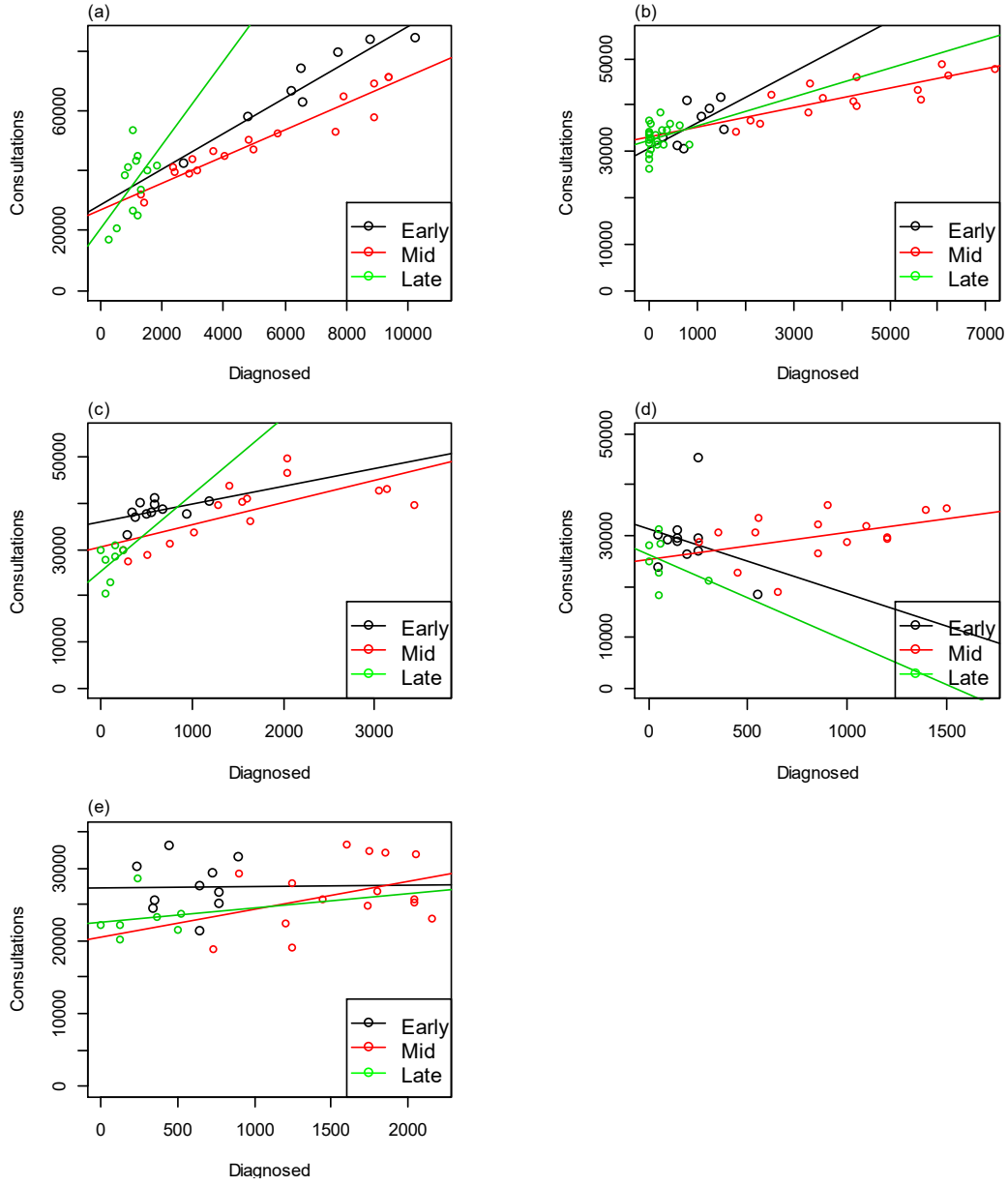


Figure 4.11–Relationship between weekly data for different periods (early, mid, late) in the influenza timeline for 5 different years: (a) represents the 2009/2010 influenza pandemic, (b) 2011/2012 seasonal influenza, (c) 2012/2013 seasonal influenza, (d) 2013/2014 seasonal influenza and (e) 2014/2015 seasonal influenza. The straight lines (black/early, red/mid and green/late) in each plot corresponds to the regression line, which is the line of best fit between the two variables.

The above results show clearly that a higher ratio between consultation and diagnosed datasets are more likely to provide better linear relationship (R^2) between both datasets. Furthermore, correlation values between consultations and diagnosed are stronger during the mid-part of the influenza period. Moreover, since for the last two years the consultations were lower than the other three years, this provided a weaker relationship between the two variables. Hence, this leads to an interesting result, where lower number of consultations weakens the potential to predict the number of consultation cases when applying the linear regression model. Furthermore, at low baseline level of non-influenza consultations, this provides a weak signal for the diagnosed ILI cases. After reaching a certain number of consultations, this is likely to give a stronger signal about the severity of the outbreak.

Early and late periods of influenza are characterized by high variability between consultation and diagnosed datasets as compared with the mid-part of the seasonal influenza (Figure 4.11). This relates well with the fact that the ratio between diagnosed and consultations is higher during the mid-part of the influenza season. Hence, during the mid-part of the influenza season, there is a higher number of individuals who sought a consultation from their GP and a higher proportion that were diagnosed with seasonal influenza. As showed in the joint model section, there is an additional time-dependent factor when discussing such datasets. For the latest two datasets (2013/2014 and 2014/2015), it is very clear that the relationship between the consultation and diagnosed variables for different periods is weak (Figure 4.11(d) and 4.11(e)). Figure 4.11(e) (2014/2015) shows that there are a substantial number of data points which are scattered around the three regression lines, hence the weak association. In contrast, for figure 4.11(d) (2013/2014), the 'early' and 'late' stages provide contrary results to all the other linear associations. In fact, these two stages (early and late) show that the association between diagnosed and consultations results in a negative slope. As explained in chapter 2, positive slopes signify that higher diagnosed cases imply higher consultations cases, while negative slopes show that higher diagnosed cases imply lower consultation cases (Figure 4.11).

Furthermore, the confidence intervals for the slope of the linear regression equations are rather wide (Tables 4.6, 4.8 and 4.10). However, for the mid-part of the seasonal

influenza, confidence intervals are more likely to be narrower. In relation to this, for the first three datasets, diagnosed ILIs are a good predictor for the number of consultations (p-value < 0.05). Only for the 2009/2010 dataset, the early part of the diagnosed data provides a significant contribution to the above linear regression model in order to predict the consultation dataset (R^2 =0.906). For the other predictors of the other models (early and late models), these are not proving to be beneficial to predict the number of consultations on a weekly basis.

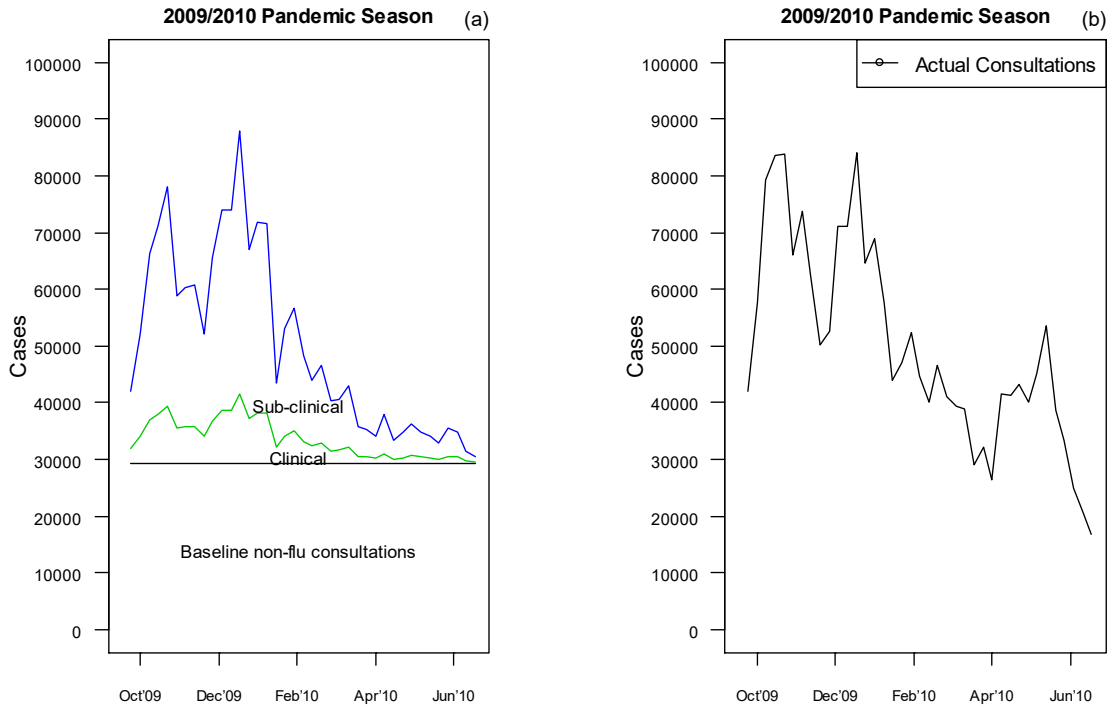


Figure 4.12 – Chart (a) represents three important stages for the 2009/2010 pandemic data. The black horizontal line represents the baseline of the non-influenza consultations (obtained from table 4.1), the difference between the black line and the green line represents the actual clinical diagnosed ILI cases and the difference between the green line and the blue line corresponds to the sub-clinical cases. Chart (b) represent the actual number of consultation cases.

The dataset presented in figure 4.12 allows us to observe an important point about the 'sub-clinical' cases. Figure 4.12 illustrate the split of each dataset into three categories. The baseline of non-influenza consultations was established through the linear regression model (Table 4.1). Furthermore, the split includes the number of weekly diagnosed cases (clinical cases) during the same season. Hence, the difference between the black line and the green line in figure 4.12 represents the actual diagnosed 'clinical' ILI cases. Then the remainder from the estimated consultations data (Figure 4.12(b)) carries a level of uncertainty, since the characteristics of the 'sub-clinical' part contains some ambiguity.

The difference between the green line and the blue line in figure 4.12 corresponds to these sub-clinical cases. Thus, the sub-clinical cases give further insight regarding those individuals who acquired influenza but did not have enough symptoms to be diagnosed as an ILI case, individuals who did not acquire influenza but were suspicious of having influenza, misdiagnosed individuals, cases that were not reported as ILI by the GPs to health authorities, or non-influenza related consultations (Figure 2.2, Chapter 2).

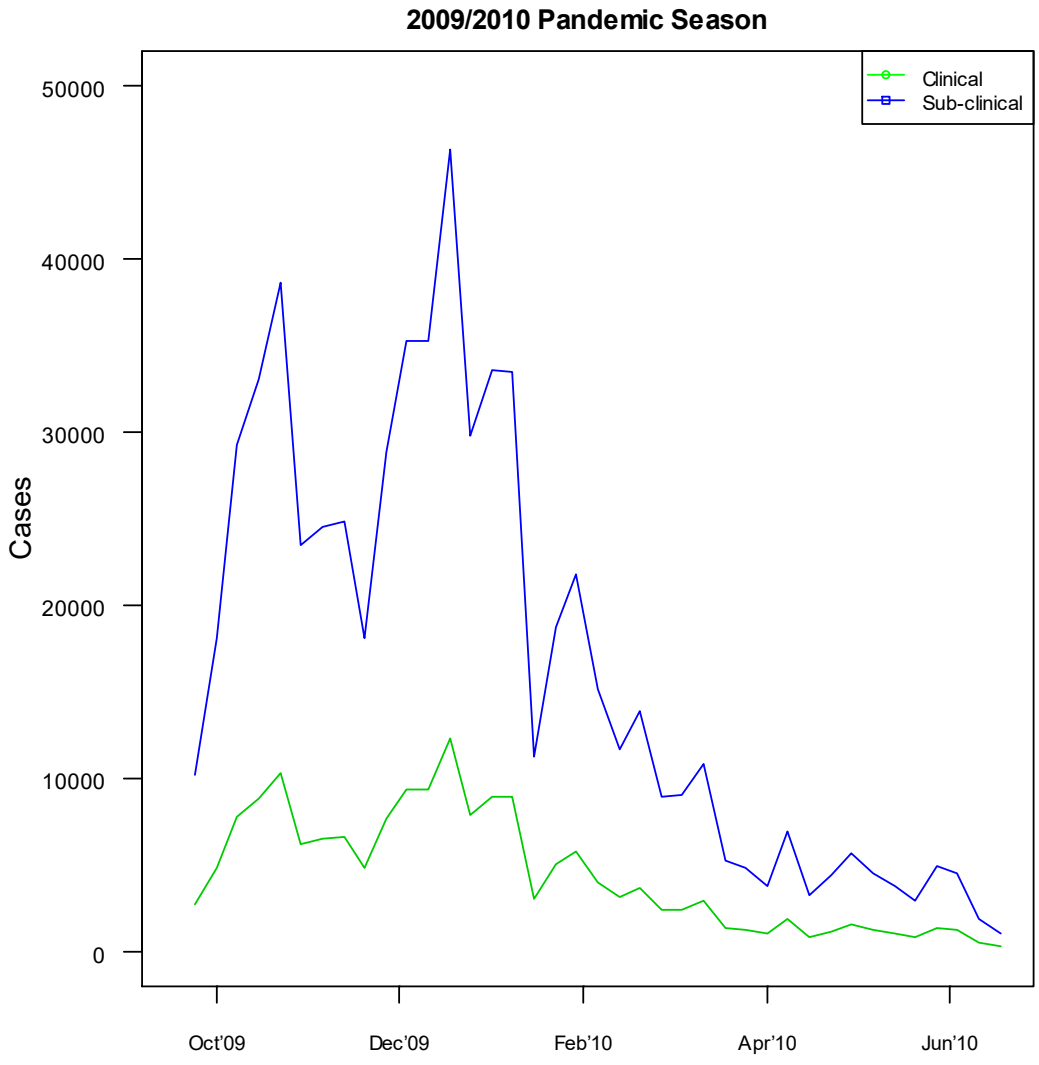


Figure 4.13 – This chart gives the actual clinical diagnosed ILI cases (green line) and the 'sub-clinical' cases (blue line). Note that in figure 4.12 the values are cumulative but in this case they are not.

The above figure (Figure 4.12) highlights the level of baseline non-influenza consultation cases which is substantially high. Furthermore, the level of 'sub-clinical' cases are substantially higher than the actual diagnosed ILI (clinical) cases (Figure 4.13). The issue

of the sub-clinical cases is related to the issue of under-reporting of the seasonal influenza cases. Although a major portion of under-reporting is due to people not consulting their GP due to their seasonal influenza (Chapter 2, Fig. 2.2), there exists a portion of under-reporting due to the sub-clinical cases as defined above. In the following chapters, we will be exploring the issue of under-reporting in further detail.

Parameters obtained through SEIR models help us to understand further the relationship between seasonal influenza datasets across different years. The average individual posterior SEIR parameter values (Tables 4.3-4.4) for the above datasets can be used as an approximation to estimate the prior parameter values for the succeeding seasonal influenza. However, consultation parameter estimates vary more widely when compared with the diagnosed datasets; hence, there is higher variability when predicting the state parameter values through the use of the consultation datasets. This corresponds well with the arguments raised in chapter 2, when defining the 'R_t for different datasets'. Nevertheless, for the scope of the seasonal influenza, diagnosed parameters are the strongest signal to understand the actual spread of influenza. As discussed in chapter 2, the diagnosed datasets are a more direct proxy of the measure of influenza, when compared with the consultations datasets. This is due to the higher number of 'background' consulting rate found in the consultation datasets. Moreover, infection posterior parameter rates (Table 4.4) for the diagnosed data show that, for a higher number of diagnosed cases the infection rate is more likely to be higher.

The above SEIR model was incorporated with the linear regression model to extend the latter model into a time-dependent one. For the scope of the time-dependent linear regression model, the diagnosed data (independent variable) was incorporated into the consultations data (dependent variable). We showed that when the linear regression model is adjusted as a time-dependent model, the predictions of the linear regression technique improved substantially (when compared with the time-independent model). Hence, this shows that although for the latest seasonal influenza datasets (2013/2014 and 2014/2015) the relationship between consultations and diagnosed datasets is very weak, when the relationship is analysed at each different time point, strong associations between the two variables can be established. Through the linear regression models (time-independent), we concluded that for lower number of consultations, the potential to

establish a strong relationship between consultations and diagnosed ILIs is weak. However, this issue was resolved by assuming a time-dependent linear regression model.

The above analysis produced very important results in understanding the relationship between the consultations and diagnosed ILI datasets. When all the datasets for different seasons were combined together, a strong linear relationship between consultations and diagnosed was recorded. This shows that the relationship between these two variables is collective for different seasons. Such findings suggest that for a new epidemic this result might also hold. This would be an interesting future research to compare such results for other influenza seasons and other different types of epidemics.

During every meeting that I held with health authorities the key health officials stated that any early signal that the seasonal influenza datasets can provide, this would be very useful for planning health strategies (Appendix A). Hence, all the above information helps to enhance our understanding of the seasonal influenza and to gain further insight that supports health authorities to better plan their health policies based on early warning techniques.

This chapter covered several signals related to the seasonal influenza that can aid to the further understanding of any other future outbreaks. The forthcoming chapters aim to use the above information to expand on this material, namely real-time forecasting, sensitivity analysis, under-reporting studies and further joint models.

Chapter 5

Real-time forecasting: The SEIR model and the joint model

5.1 Introduction

As discussed in chapter 1, one of the ultimate aims in epidemiological studies is to improve the prediction of the disease spread as early as possible in the epidemic. The main challenge is to develop a set of robust techniques that provide an early warning signal. As discussed in chapter 4, this proves difficult when limited information exists. Hence, our ultimate aim is to acquire as much information as possible to enhance our understanding of any outbreak under study. For example, we already showed in chapter 4 that when applying a time-dependent linear regression model, this improved the model fit when compared to a general linear regression model. Throughout this chapter, I will be looking at the extent to which the SEIR and joint models (as defined previously) can be used to accurately predict future forecasts based on real-time data. Real-time forecasting aims to carry out a 'stock-take' of the collected data, and then through the use of the SEIR model this further predicts the number of consultations and diagnosed cases for the following weeks. Furthermore, we carry out real-time forecasting to understand whether the forecasts for the consultation datasets can be improved through the use of the joint model (SEIR model and time-dependent linear regression model). Therefore, we also examine whether the predictions of multiple datasets can be facilitated through this integrated framework.

5.2 Method

Throughout this chapter, we use the consultations and diagnosed datasets for the five influenza seasons (2009/2010, 2011/2012, 2012/2013, 2013/2014, 2014/2015). For every season, I will produce three outputs as follows:

1. Real-time forecasting for the consultations dataset through the SEIR model;
2. Real-time forecasting for the diagnosed dataset through the joint model but using the SEIR model;
3. Real-time forecasting for the consultations dataset through the joint model but using the time-dependent linear regression model.

Since I aim to explore the forecasting of consultation and diagnosed variables at different stages of the outbreak, six cases are considered for every dataset (as defined above). The following are the number of known data points that are considered in the application of real-time forecasts:

1. Case 1: 9 data points (chart (a))
2. Case 2: 12 data points (chart (b))
3. Case 3: 15 data points (chart (c))
4. Case 4: 18 data points (chart (d))
5. Case 5: 21 data points (chart (e))
6. Case 6: 24 data points (chart (f))

and predicting the next 20 data points for every single case (or fewer if these extended beyond the end of the season). Hence, actual data points are considered unknown for the forecasted 20 data points (or less). For the time-dependent linear regression model, only the parameters up till the 'known' data points are considered to forecast the consultations through the joint model. The 95% high predictive density regions are portrayed to understand the level of accuracy for each individual forecast.

5.3 Results

5.3.1 2009/2010 pandemic data

As discussed in previous chapters, the pandemic data has different features when compared to the seasonal influenza datasets. The peaks of the pandemic data were reached during different periods (Oct 2009 and Dec 2009) when compared to the seasonal influenza datasets (Feb-April period). Figure 5.1a shows that when using 9 weeks of data to predict the next 20 weeks of data, the predictions have narrow confidence interval when compared with the next two cases (Figures 5.1b and 5.1c), but missing the prediction of the second wave. In fact, the forecasts clearly underestimated the actual data. Hence, up to the 9^{th} week of data, the information is very limited to predict future weekly data points. Figure 5.1b shows that when using 12 known data points, the real-time forecast (for the next 20 points) is fairly accurate, though missing the December 2010 peak. For the remaining plots (hence using 15/18/21/24 weeks of data to predict the next 20 weeks (or fewer)), the real-time forecasts are also accurate. For figures 5.1b and 5.1c, the 95% confidence intervals are rather wide; hence this shows the level of uncertainty during the first part of the outbreak. However, as time progresses, the predictions stabilize to lower confidence intervals (Figures 5.1d, 5.1e, 5.1f).

Figures 5.2b and 5.2c for the 2009/2010 diagnosed dataset are better real-time forecasts when compared with the consultations 2009/2010 dataset (for the same cases). In fact,

when using 12 weeks of data (Figure 5.2b), another peak is being predicted, although with 2 weeks of time lag. In general, for the diagnosed data, the confidence intervals are narrower (compared with consultations), resulting in a lower uncertainty in the provided predictions. For the last four real-time forecast cases (Figures 5.2c – 5.2f), the 20 weeks of forecasted data points are very accurate, with a very smooth curve passing through most of the actual data points.

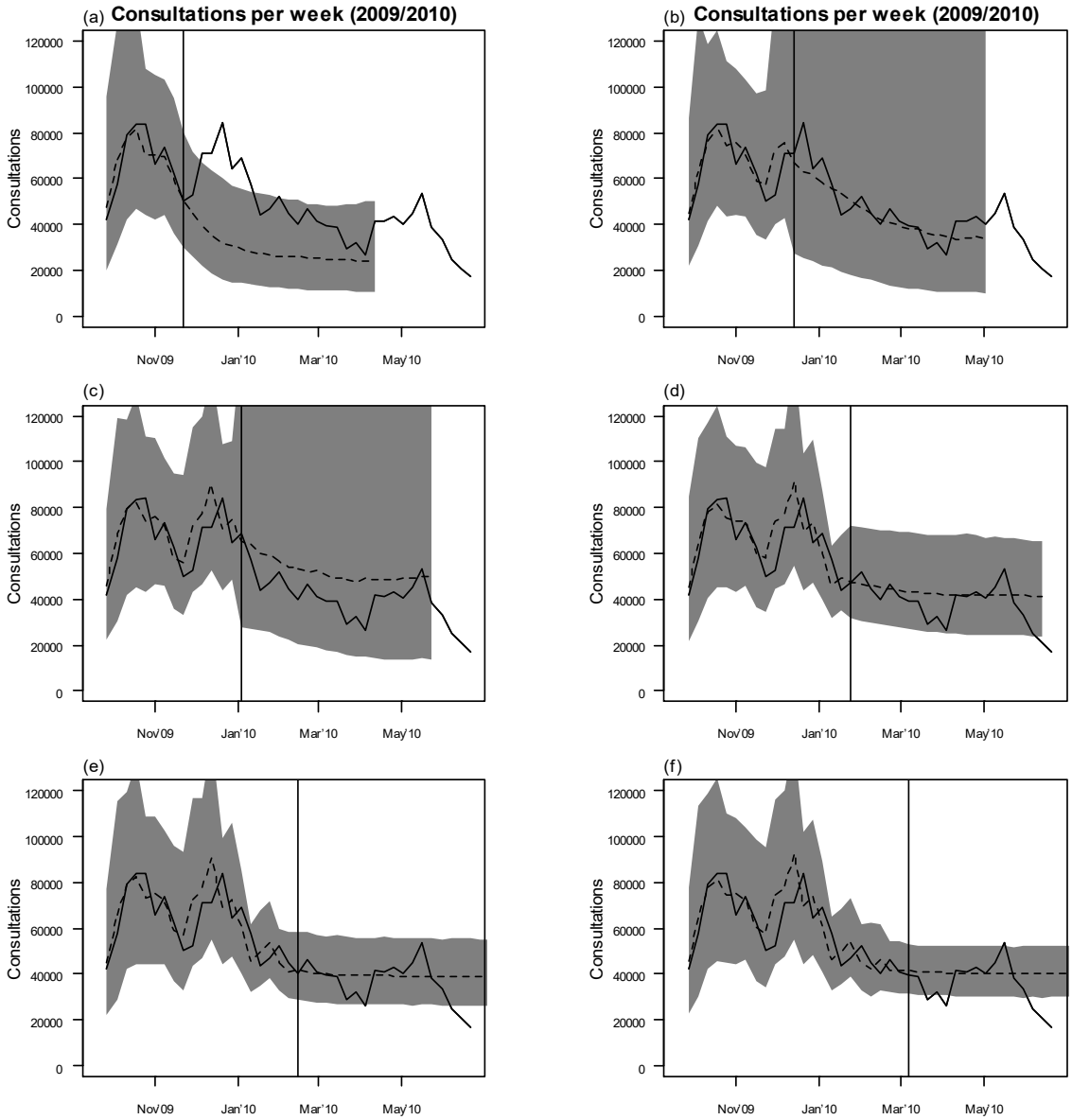


Figure 5.1 – Prediction plots at different time points when the model was fitted for the consultations dataset (2009/2010) using the SEIR model as defined in Chapter 3, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

The first dataset, which tested the joint model for the real-time forecast, is the 'Consultations per week' data for the pandemic season 2009/2010. Hence, the diagnosed dataset was set as the independent variable, and through the time-dependent linear regression model (described in Chapter 4), the consultation data points were estimated.

When the joint model was run through the methodology as defined in chapter 4, the forecasts for different consultation time points (Figure 5.3) are more accurate than in figure 5.1. When using 9 known data points (Figure 5.3a), the model produced an improved fit when compared with figure 5.1a (consultations 2009/2010, SEIR model). Furthermore, the joint model predicted the 2nd peak of the consultations data at week 12 (Figure 5.3b), in contrast to figure 5.1b. The joint model fit also improved for Case 3 (Figure 5.3c) when compared with figure 5.1c, producing better forecasted values.

The confidence intervals of the consultations data for the joint model (Figure 5.3) are narrower when compared to the consultations data for the SEIR model (Figure 5.1). In addition, this shows that the joint model for the 2009/2010 dataset improved the certainty and accuracy in the consultation predictions. This can be attributed due to lower confidence intervals for the diagnosed data, since the consultation predictions (joint model) are dependent on the diagnosed data (linear regression model). Hence, in this case, a strong relationship between consultations and diagnosed data (as described in Chapter 4) improved the forecasts of the consultations dataset.

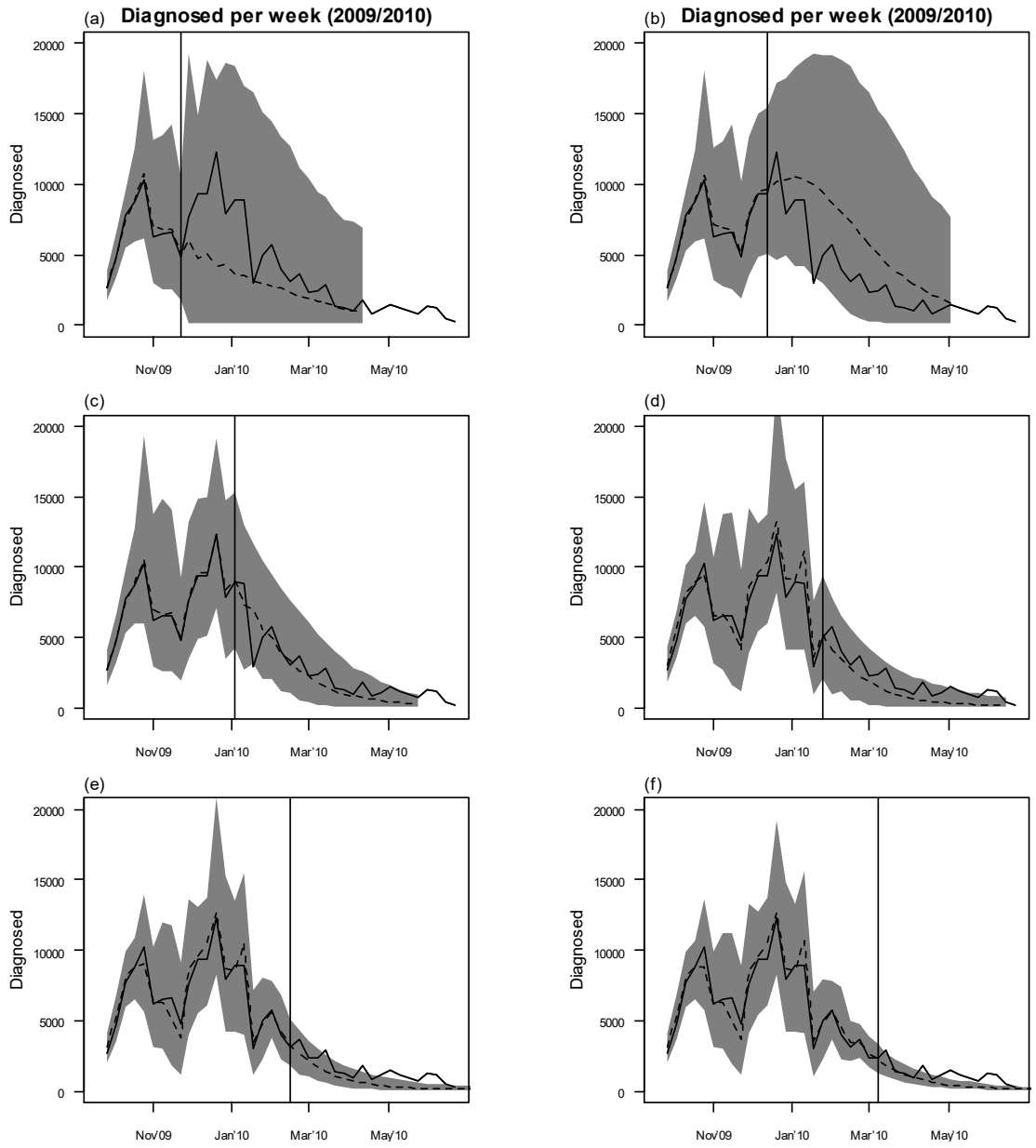


Figure 5.2 – Prediction plots at different time points when the model was fitted for the diagnosed dataset (2009/2010) using the SEIR model as defined in Chapter 3, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

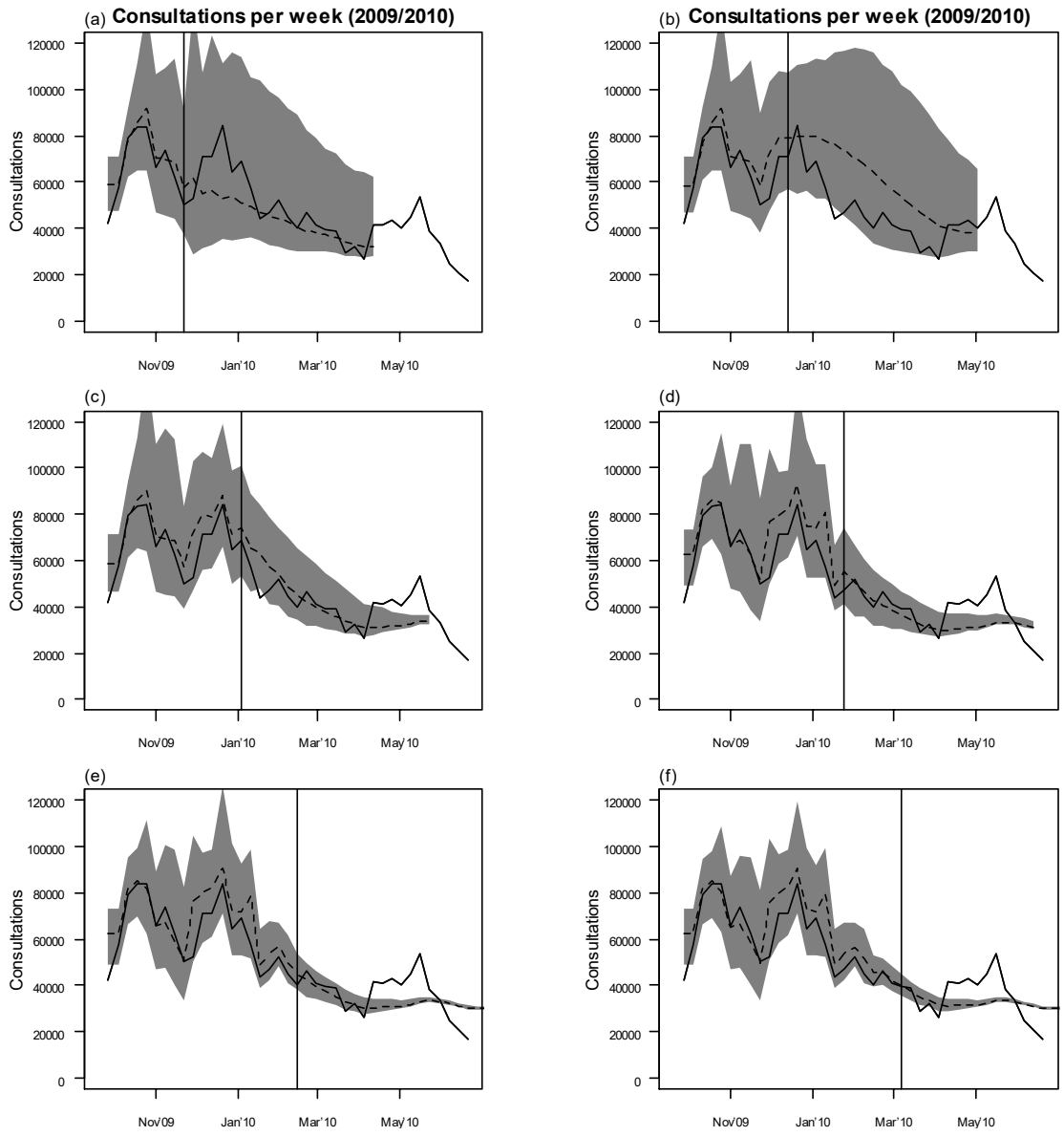


Figure 5.3 – Prediction consultation plots (2009/2010) at different time points when the model was fitted through the time-dependent linear regression model (joint model) between consultations (predicting variable) and diagnosed ILIs (independent variable) as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

5.3.2 2011/2012 seasonal influenza data

Similarly to the above, the analysis was run for the 2011/2012 seasonal influenza datasets. Hence, the SEIR model was run on its own for the consultations dataset. The joint model

was subsequently run to forecast the diagnosed data through the SEIR model, and the consultations data through the time-dependent linear regression model.

Similar to the outcome of figure 5.1a, the consultation (20 weeks) predictions are underestimated for figures 5.4a and 5.4b. As more data points were observed, the forecasts improved (Figures 5.4c, 5.4d and 5.4f). The results produced in figure 5.4e (assuming 21 known data points) again underestimate the number of actual consultation cases.

For the first prediction plot of the diagnosed data (Case 1, as defined above), the model produced overestimation of the peak of the diagnosed data (Figure 5.5a). Although the forecasts improved when using 12 data points to predict the next 20 data points (Figure 5.5b), the prediction is still not accurate. The peak is being predicted later during the outbreak with a wide confidence interval. Hence, this shows that, up to this point, the SEIR model is not predicting future data points accurately. For higher number of known data points (Figures 5.5c, 5.5d), the model is still predicting a larger outbreak, with a lot of uncertainty (wide confidence intervals). Only for the latest two cases (Case 5 and 6, the end stages of the influenza) do the predictions become notably better.

The real-time forecast for the consultations data, using the joint model, overestimated the number of consultation cases, when assuming 9 known data points (Figure 5.6a). This result is attributed to the overestimation of the forecasted diagnosed cases (consultations dependent on diagnosed cases). From case 2 onwards, the forecasts improved. For figures 5.6c and 5.6d, the joint model technique is forecasting higher consultations when compared with the SEIR model (Figure 5.4c and 5.4d), while figures 5.6e and 5.6f produced more accurate forecasts, with narrower confidence intervals when compared to figures 5.4e and 5.4f. For the 2011/2012 consultations dataset, the joint model also produced forecasts with a lower level of uncertainty and more reasonable forecasts (Figures 5.6e and 5.6f) when compared with the same cases of figure 5.4. Since the diagnosed forecasts (Figure 5.5) are providing a signal of a potential outbreak, the consultation forecasts also follow the same outcome, although with some overestimated predictions.

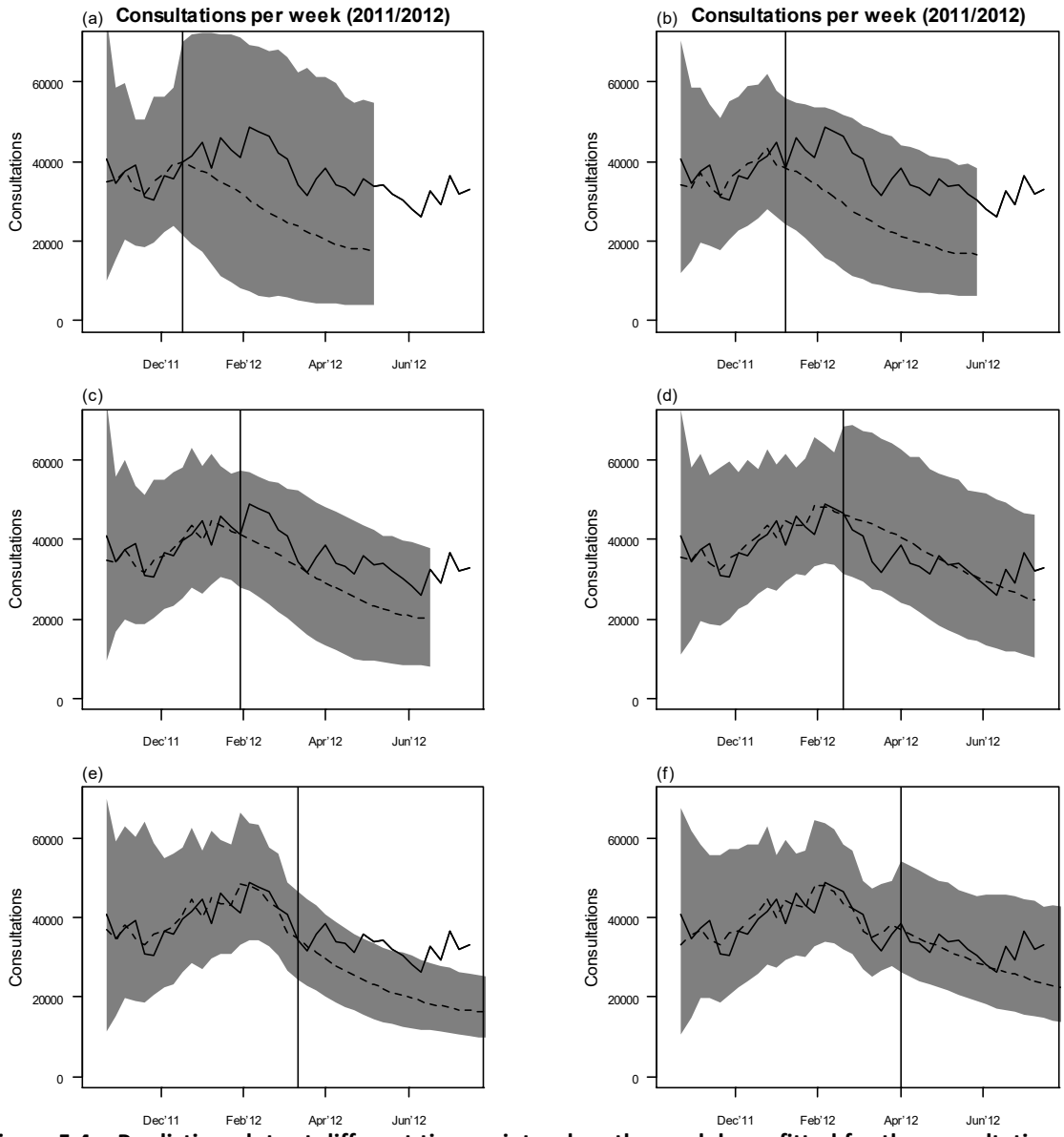


Figure 5.4 – Prediction plots at different time points when the model was fitted for the consultations dataset (2011/2012) using the SEIR model as defined in Chapter 2, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

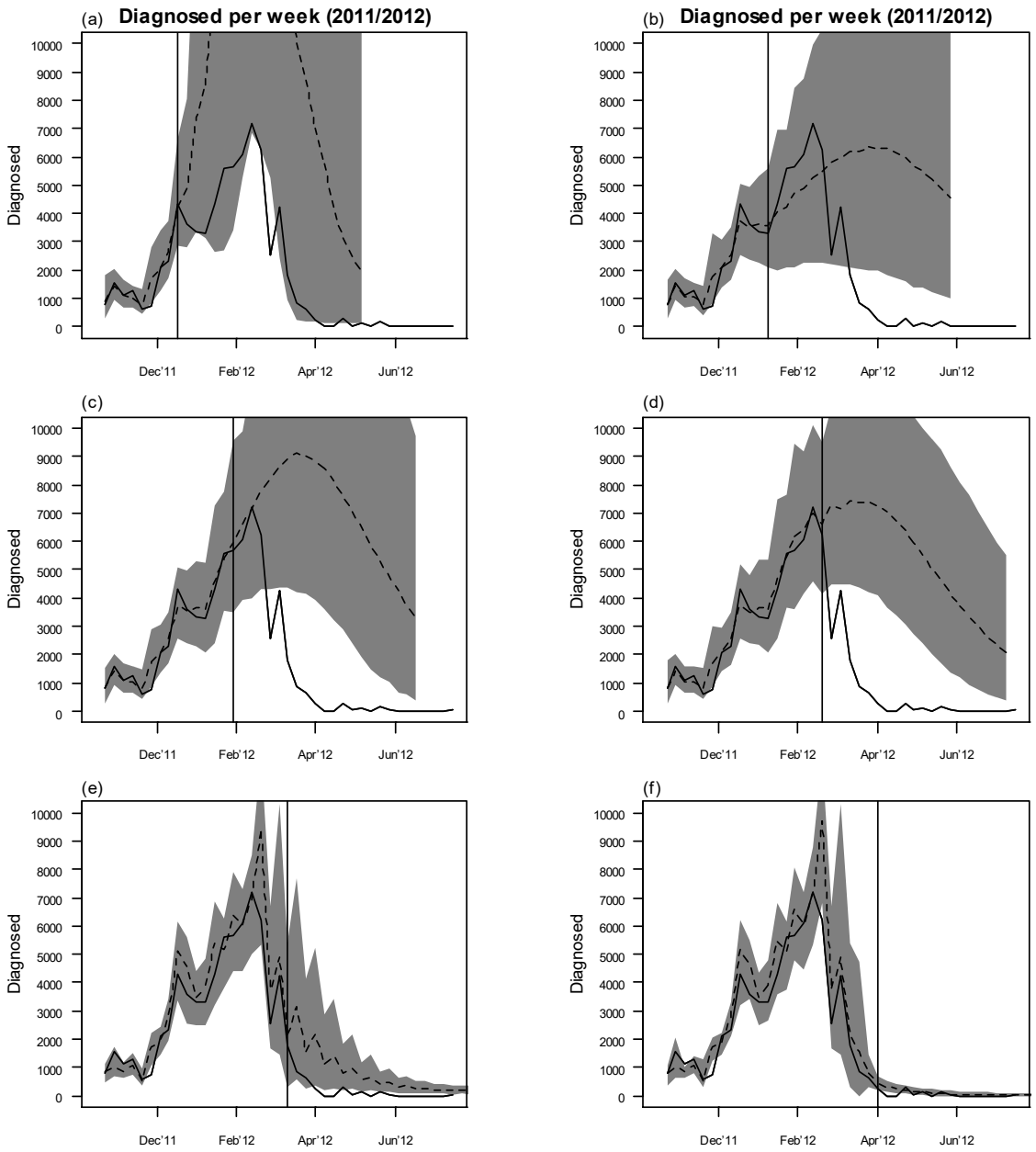


Figure 5.5 – Prediction plots at different time points when the model was fitted for the diagnosed dataset (2011/2012) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

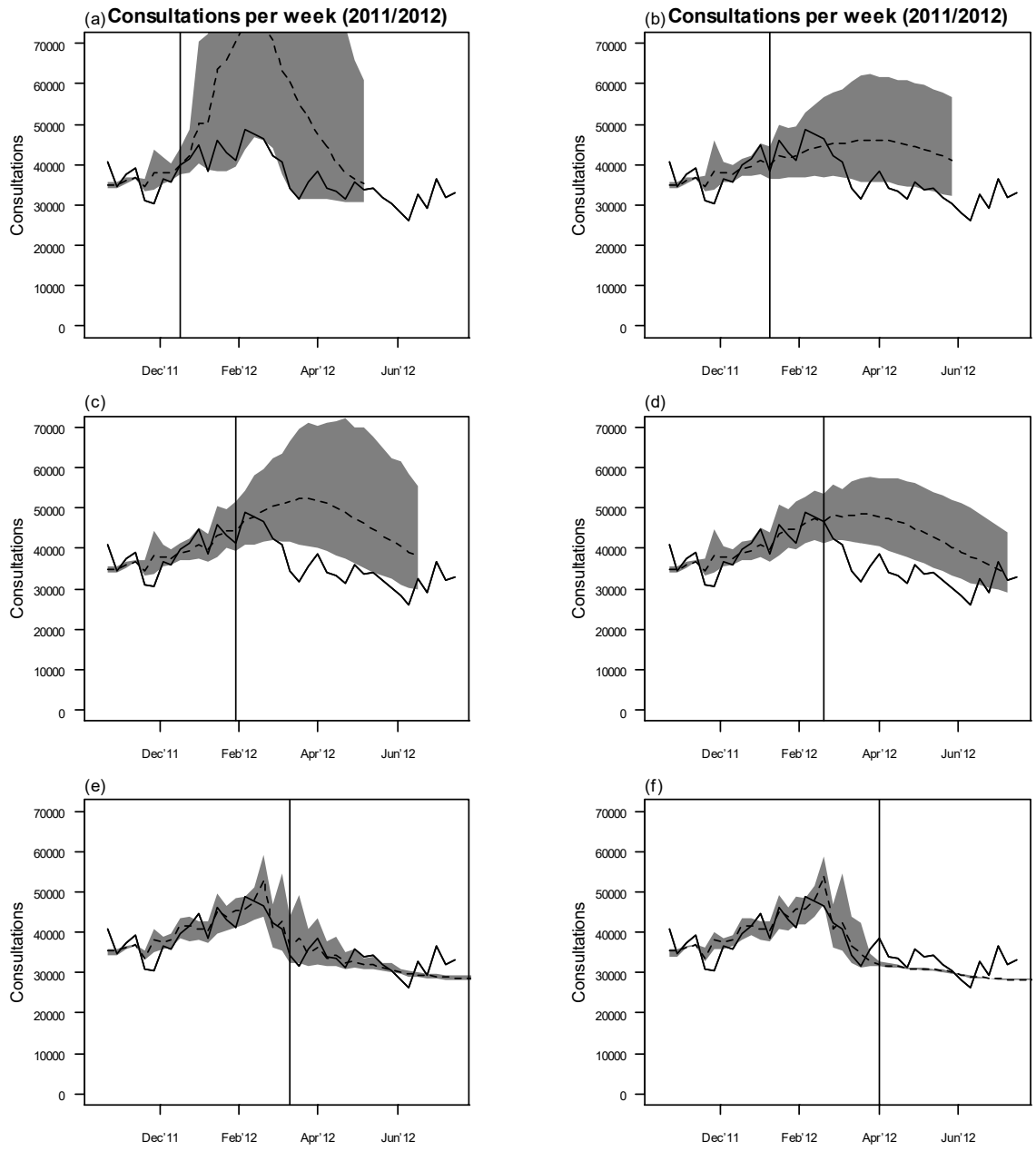


Figure 5.6 – Prediction consultation plots (2011/2012) at different time points when the model was fitted through the time-dependent linear regression model (joint model) between consultations (predicting variable) and Diagnosed ILIs (independent variable) as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

5.3.3 2012/2013 seasonal influenza data

Similar to the previous two seasons when assuming 9 known data points, the SEIR model forecasts for consultations were underestimated (Figure 5.7a). However, the forecasts

improved when assuming more known data points (Figure 5.7), except when assuming 15 known data points. For the latter case (Figure 5.7c), the number of consultations were overestimated. In general, the certainty in predictions also improved with time (resulting in narrower confidence intervals). For the diagnosed 2012/2013 dataset (using the joint model), the model produced weak predictions. The predictions (Case 1) were initially underestimated by a high degree (Figure 5.8a). For case 2, the model accurately predicted the next few data points, but overestimated the peak of the diagnosed cases substantially (Figure 5.8b). This overestimation proceeded for the next two cases (assuming 15 and 18 data points respectively) with very wide confidence intervals (Figures 5.8c and 5.8d). After assuming the peak of the diagnosed cases as known (Figure 5.8e), the forecasts declined, though still carrying a certain level of overestimation and wide confidence intervals. The last case (assuming 24 known data points) accurately predicted the remaining data points (Figure 5.8e), though the influenza was then during its final stages.

Unsatisfactory real-time forecasts for the diagnosed 2012/2013 dataset might imply bad forecasts for the consultations dataset through the joint model. This can be seen to a certain extent in figure 5.9. Initially, the consultation predictions commenced fairly accurate, but then deteriorated from case 2 (assuming 12 known data points), especially for figures 5.9b and 5.9d. Ultimately, for the last two cases (Figures 5.9e and 5.9f), the predictions improved once again. The accurate forecast of figure 5.9a (assuming 9 observed data points) can be attributed to the fact that the diagnosed predictions (Figure 5.8a) are flat and low, and the actual consultations data is also flat and stationary. Hence, these accurate forecasts were produced since the consultations data utilised the diagnosed data as the dependent variable. For the next three cases (Figures 5.9b, 5.9c and 5.9d), the forecasts deteriorated with a clear overestimation for substantial parts of the outbreak due to the overestimation of the diagnosed forecasts (Figure 5.8). Moreover, when assuming 21 known data points and 24 known data points, the forecasts improved (similar to the diagnosed 2012/2013 dataset). For the 2012/2013 consultations dataset, the joint model (Figure 5.9) did not produce better forecasts when compared to the SEIR model (Figure 5.7). The main reason for these results is due to the weak forecasts for the diagnosed dataset 2012/2013 (consultations dependent on diagnosed).

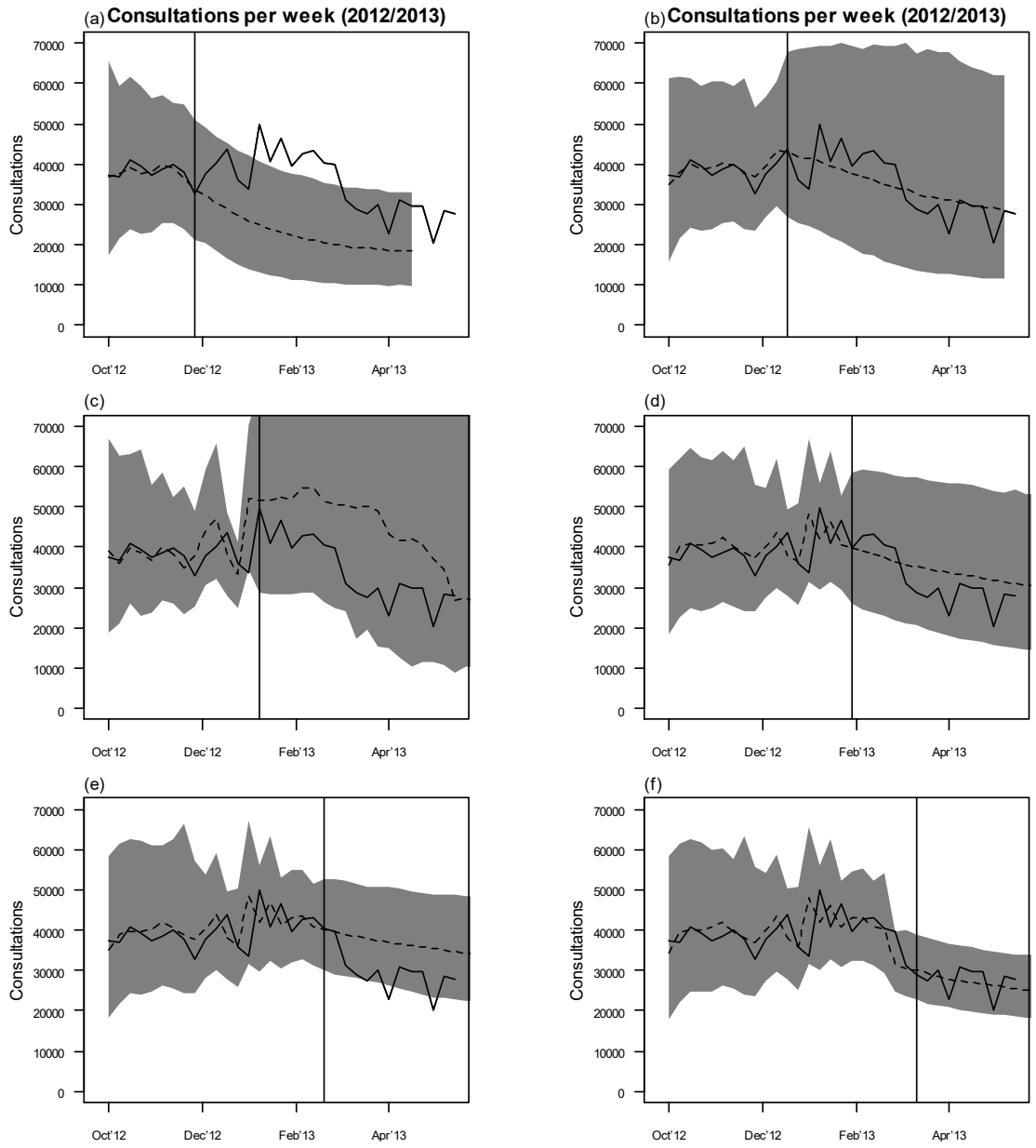


Figure 5.7 – Prediction plots at different time points when the model was fitted for the consultations dataset (2012/2013) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

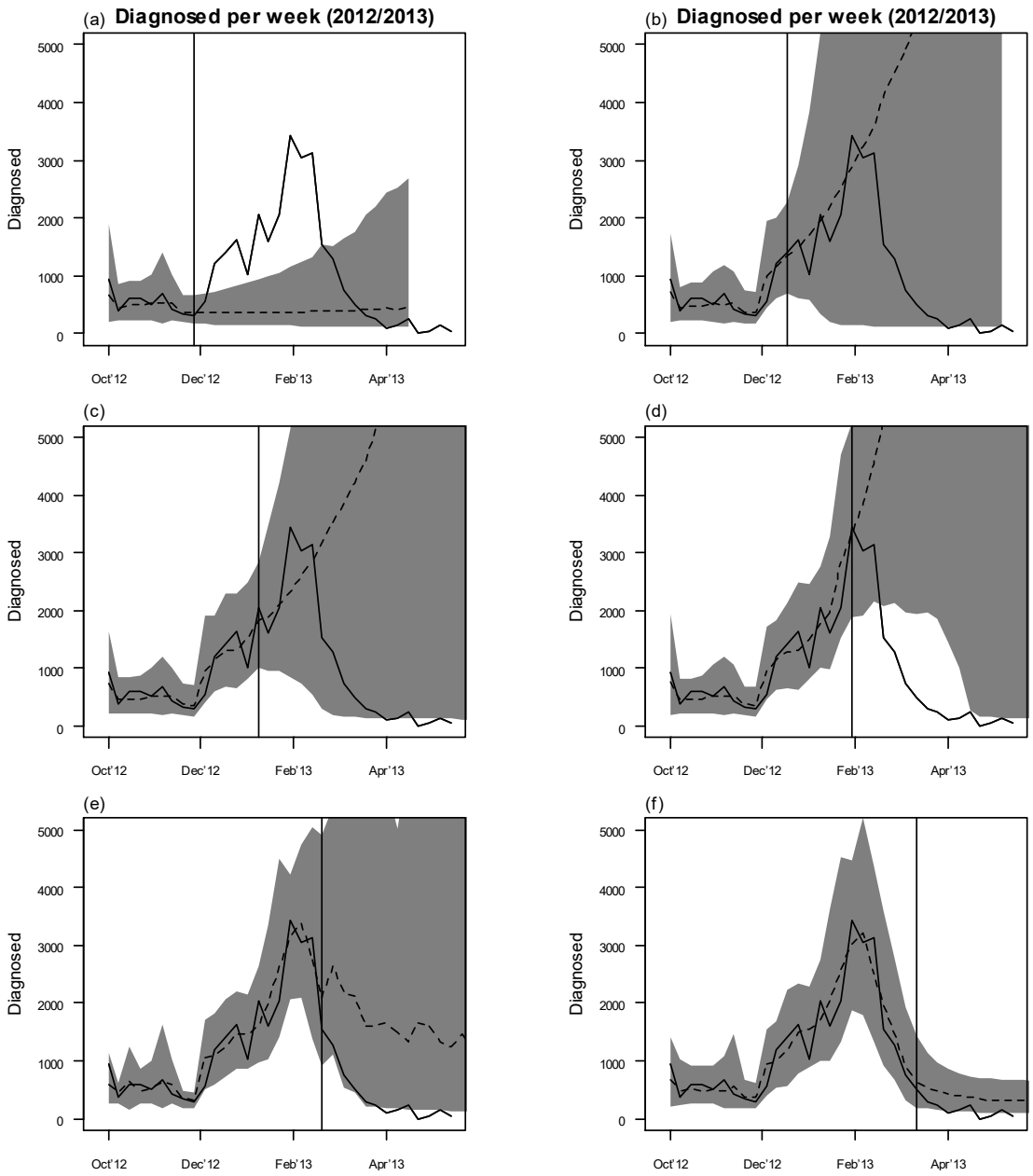


Figure 5.8 – Prediction plots at different time points when the model was fitted for the diagnosed dataset (2012/2013) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

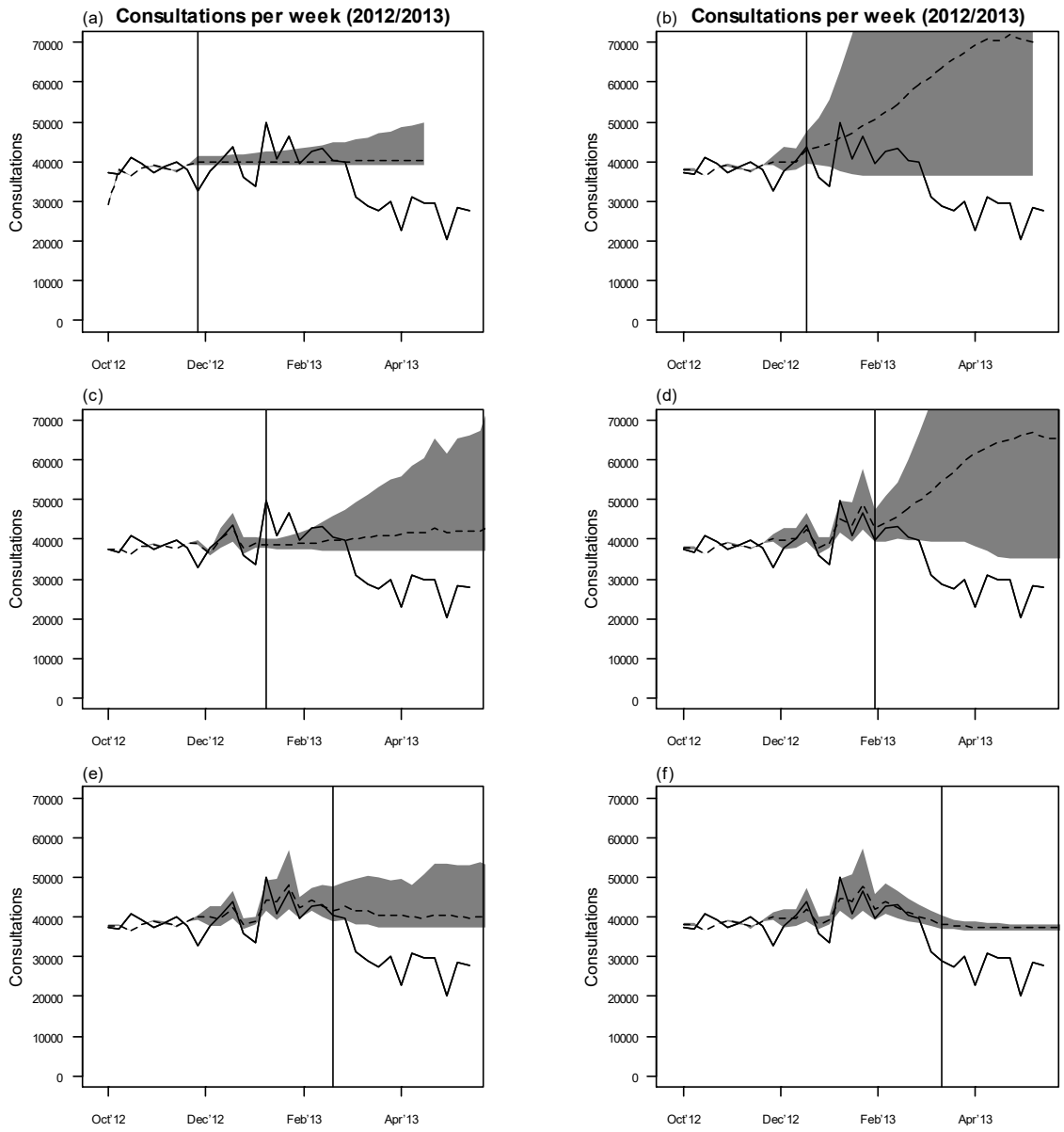


Figure 5.9 – Prediction consultation plots (2012/2013) at different time points when the model was fitted through the time-dependent linear regression model (joint model) between consultations (predicting variable) and Diagnosed ILIs (independent variable) as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

5.3.4 2013/2014 seasonal influenza data

The 2013/2014 consultations dataset is similar to the previous consultations datasets (SEIR model). In fact, when assuming 9 known data points, the forecasts are underestimated (Figure 5.10a). For the next two cases (Figures 5.10b and 5.10c), the

forecasts are also underestimated, however with narrow confidence intervals. On the contrary, figure 5.10d shows an overestimation of the number of consultation cases, while figures 5.10e and 5.10f are rather accurate with a clear signal of the number of consultation cases for the coming weeks, together with narrow confidence intervals.

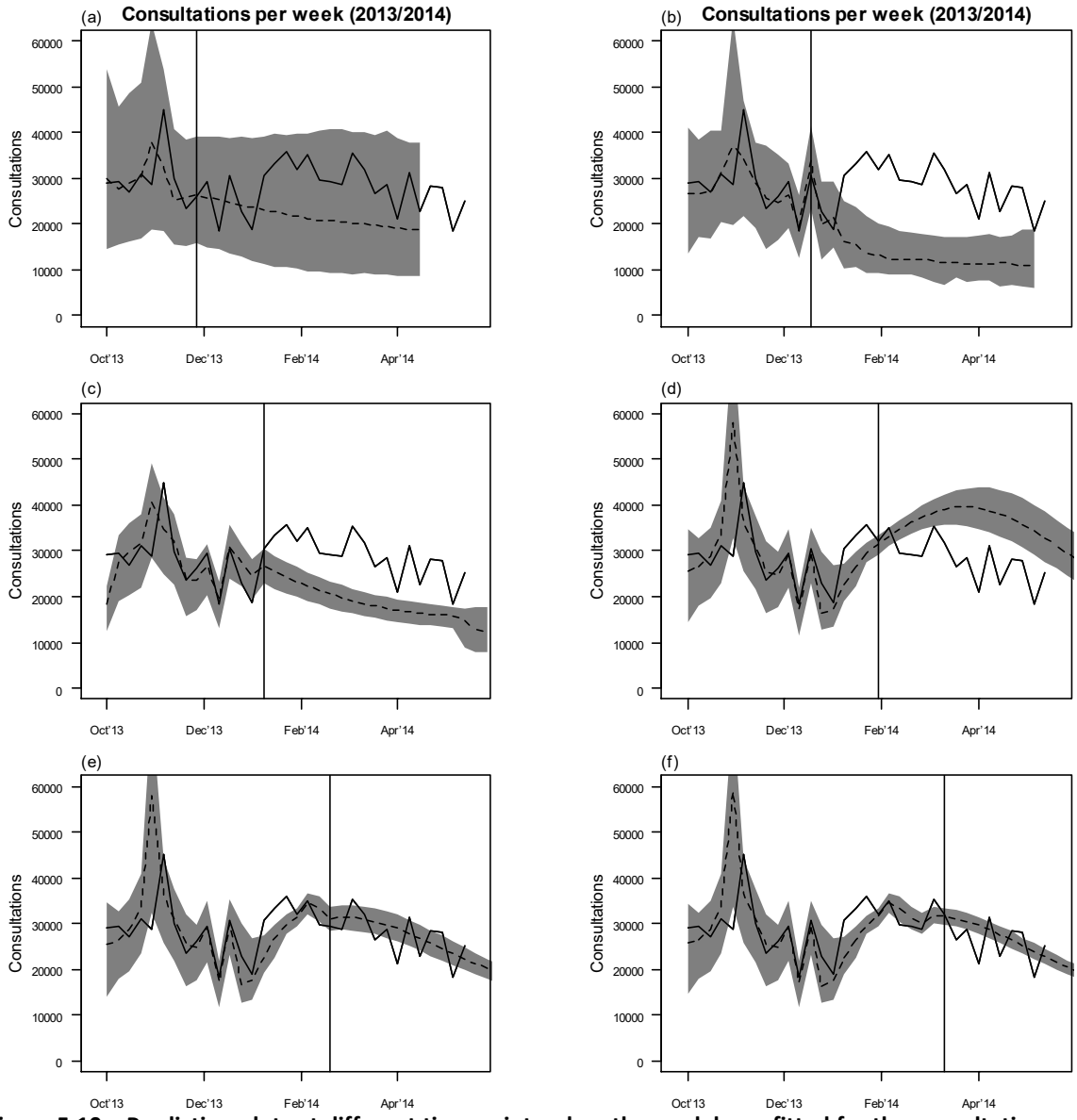


Figure 5.10 – Prediction plots at different time points when the model was fitted for the consultations dataset (2013/2014) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

The diagnosed 2013/2014 cases commence with a long period of a low number of diagnosed ILI cases. In general, the forecasts produced (Figure 5.11) are inaccurate and similar to the ones obtained in the previous influenza season (Figure 5.8, diagnosed 2012/2013). These forecasts (Figure 5.11) have wide confidence intervals.

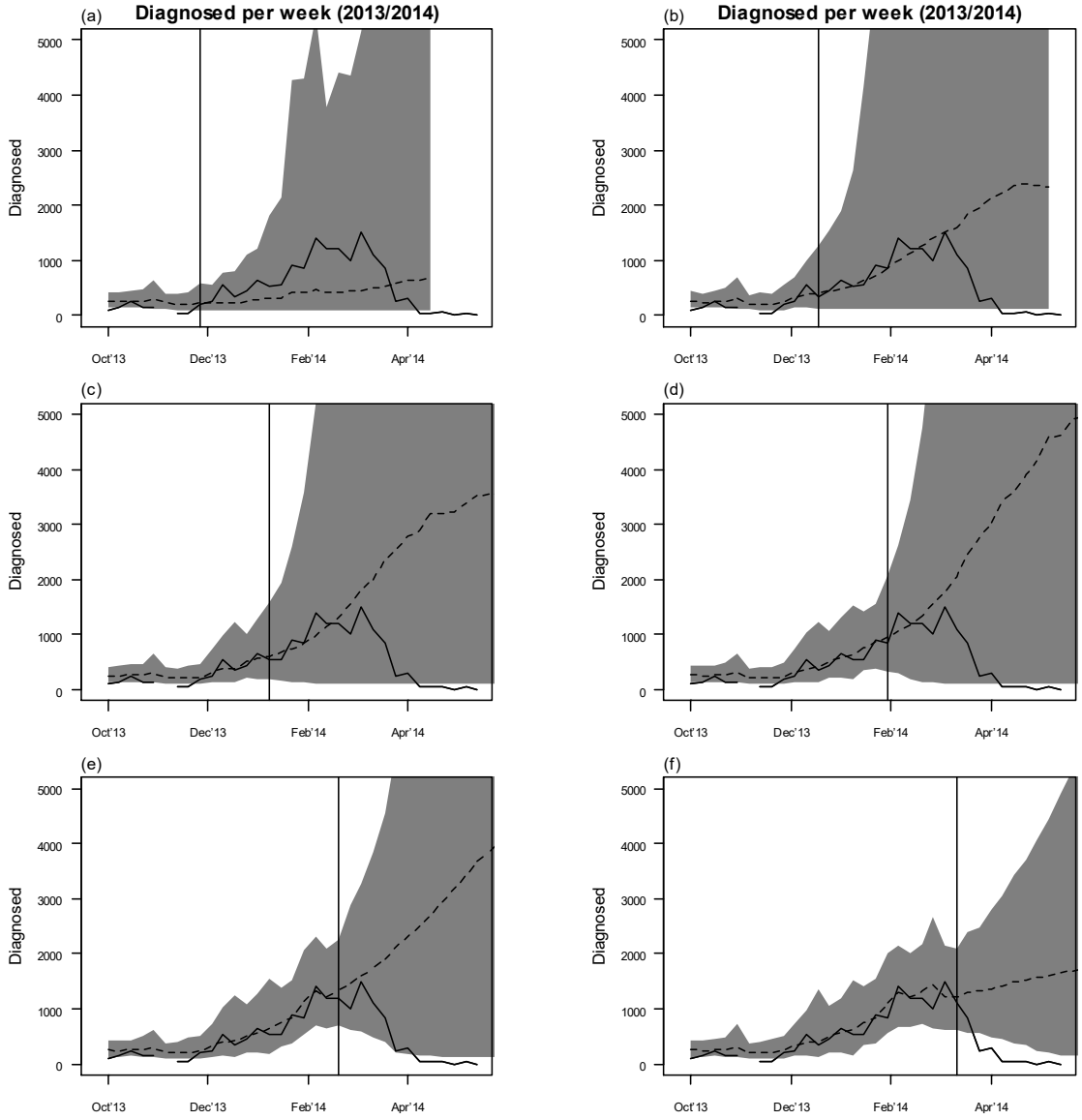


Figure 5.11 – Prediction plots at different time points when the model was fitted for the diagnosed dataset (2013/2014) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

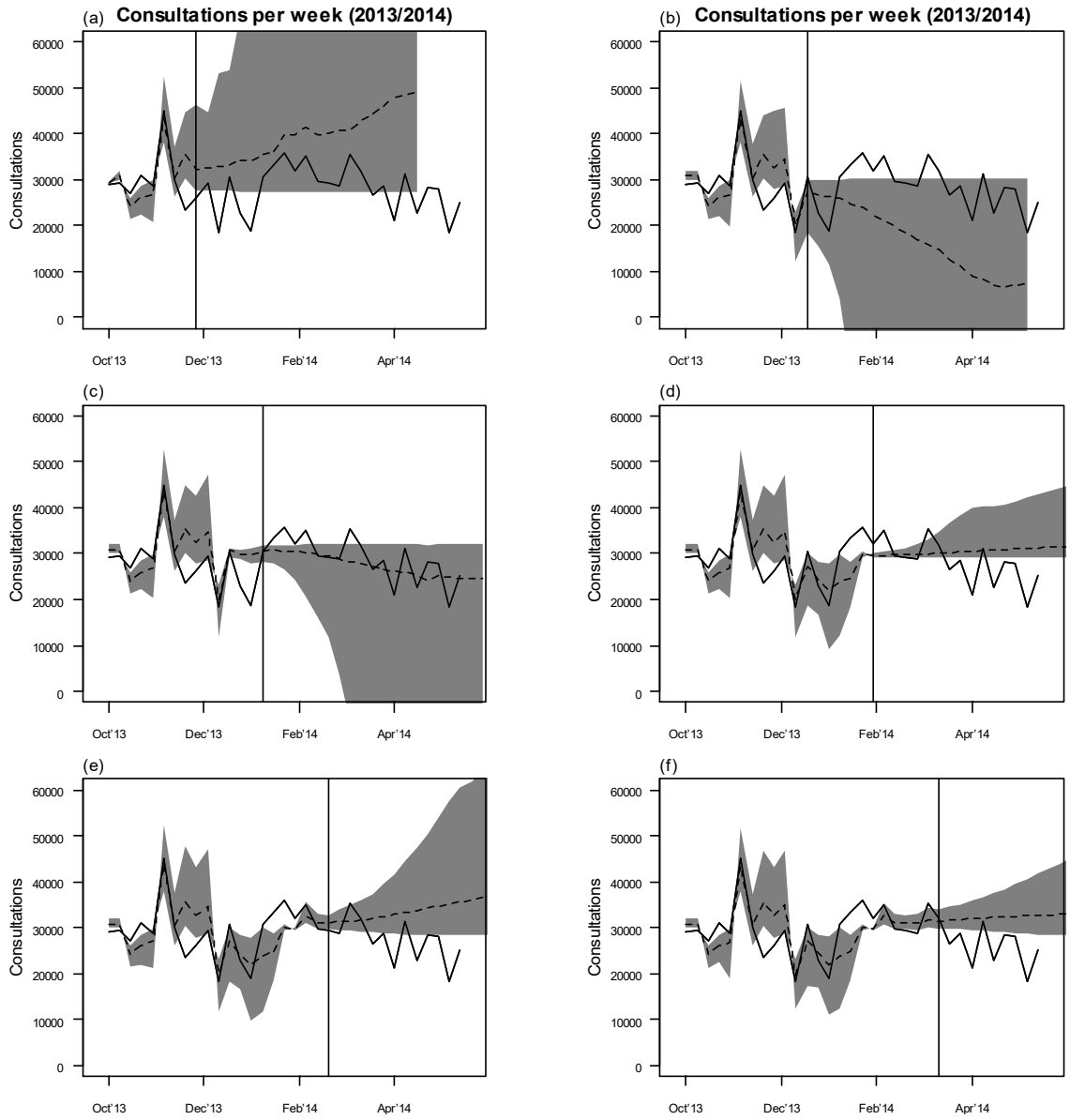


Figure 5.12 – Prediction consultation plots (2013/2014) at different time points when the model was fitted through the time-dependent linear regression model (joint model) between consultations (predicting variable) and diagnosed ILIs (independent variable) as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

Figure 5.12a shows an overestimation for the number of consultation cases with wide confidence interval. In contrast, figure 5.12b shows underestimation when compared with the actual dataset. For the first 16 weeks of the influenza season, the linear relationship between the diagnosed and consultation variables is inconsistent (Figure E.3, Appendix E). In fact, for some time points the gradient is negative (as discussed for Figure 4.11).

This influenced the confidence intervals for figures 5.12b and 5.12c. The lower part of the confidence interval is negative, which is unrealistic. Hence, this may be considered as a limitation in this method. However, this shows the level of uncertainty in the predictions for these two cases. Figure 5.12c shows accurate predictions, but wide corresponding confidence interval. When considering more known consultation cases (Figures 5.12d – 5.12f), the forecasts improved with narrower confidence intervals when compared with the previous forecasts (Figures 5.12a – 5.12c).

5.3.5 2014/2015 seasonal influenza data

As discussed in chapter 4, the consultations dataset for the 2014/2015 is rather stationary with some short-term oscillations. The SEIR model picked up the signal of the stationary data, as the forecasts for all cases are fairly accurate, except for figure 5.13b (assuming 12 known data points and predicting the next 20 weeks). For the latter figure, the consultations were slightly underestimated. The confidence intervals are narrow for most of the consultation forecasts, thereby showing high certainty in the predicted values. Thus, such a stationary dataset produces low uncertainty forecasts due to the low variability in the dataset.

As shown in the previous three diagnosed influenza datasets, overestimation was also recorded for the 2014/2015 diagnosed dataset (Figure 5.14). The confidence intervals are very wide for all different cases. A decline of the seasonal influenza was never predicted, hence producing a lot of uncertainty in these predictions for this dataset.

Figure 5.15 shows accurate forecasts for most of the consultations dataset (joint model). For the first three figures (Figure 5.15a, 5.15b and 5.15c), assuming 9, 12 and 15 known data points, the forecasts are close to the real data, and with narrow confidence intervals. Although for the diagnosed dataset the confidence intervals are wide (Figure 5.14), the confidence intervals are narrow for the consultations dataset (consultations dependent on diagnosed). This is attributed to a gradient close to 0 for the linear relationship between diagnosed and consultations at the points (9/12/15 data points) when the forecasts are carried out (Figure E.4, Appendix E). Hence, in such cases, the consultation predictions (Figure 5.15) are mainly based on the y-intercept (baseline of non-influenza cases). For figure 5.15d, the forecasts include some overestimation of the number of consultations

cases, while for the last two figures (Figures 5.15e and 5.15f), the forecasts improved with reasonable predictions.

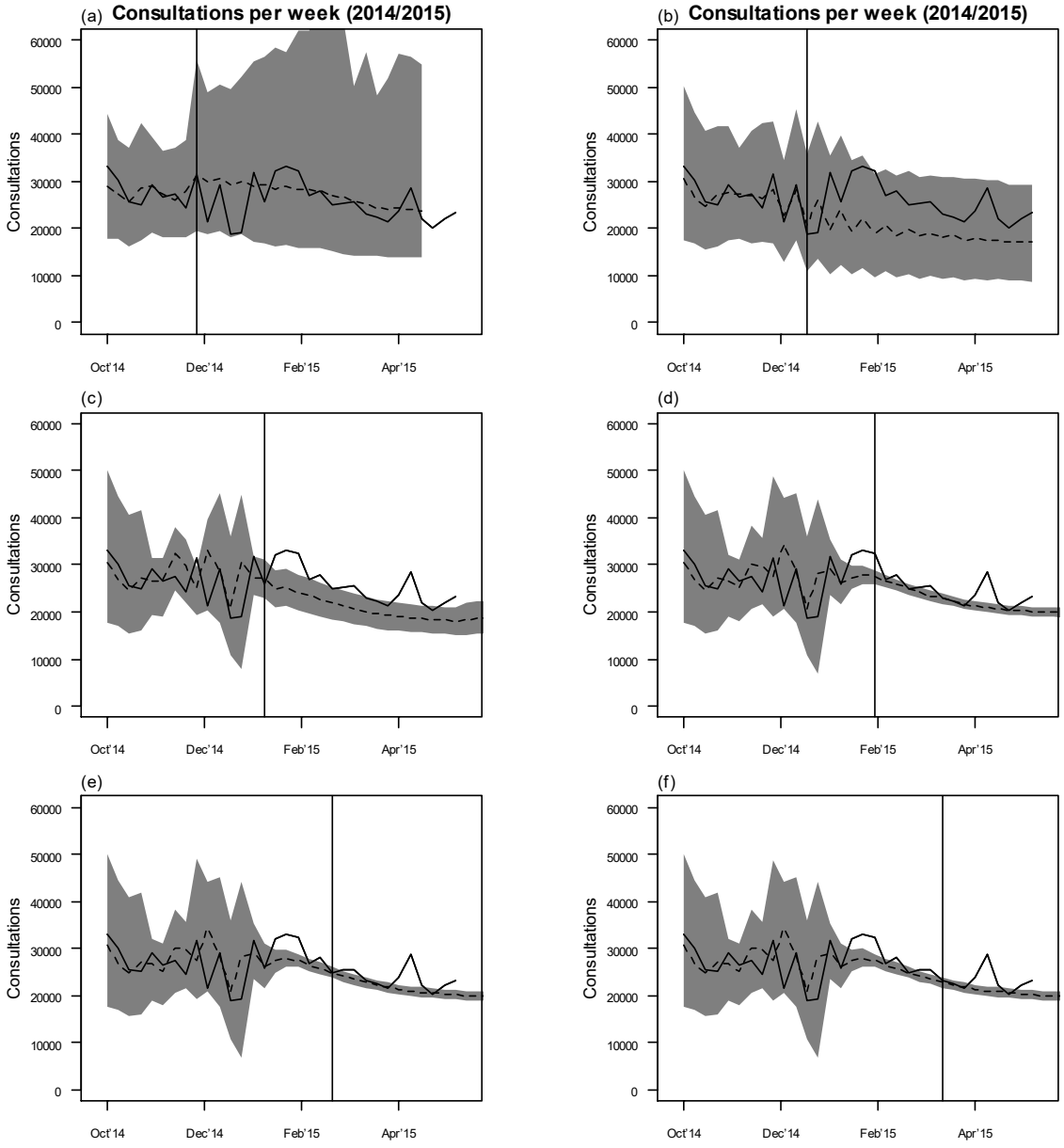


Figure 5.13 – Prediction plots at different time points when the model was fitted for the consultations dataset (2014/2015) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

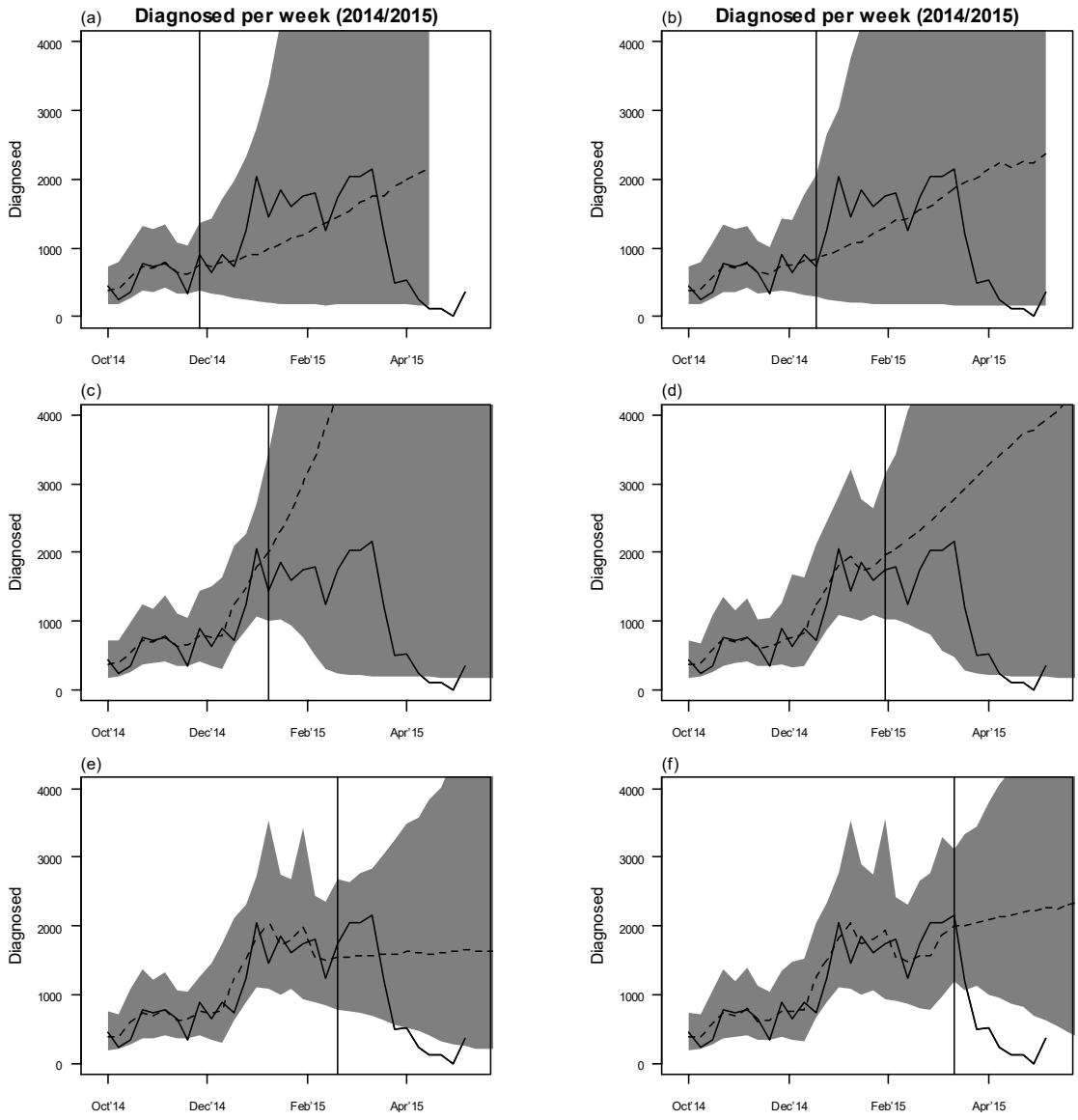


Figure 5.14 – Prediction plots at different time points when the model was fitted for the diagnosed dataset (2014/2015) using the SEIR model as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

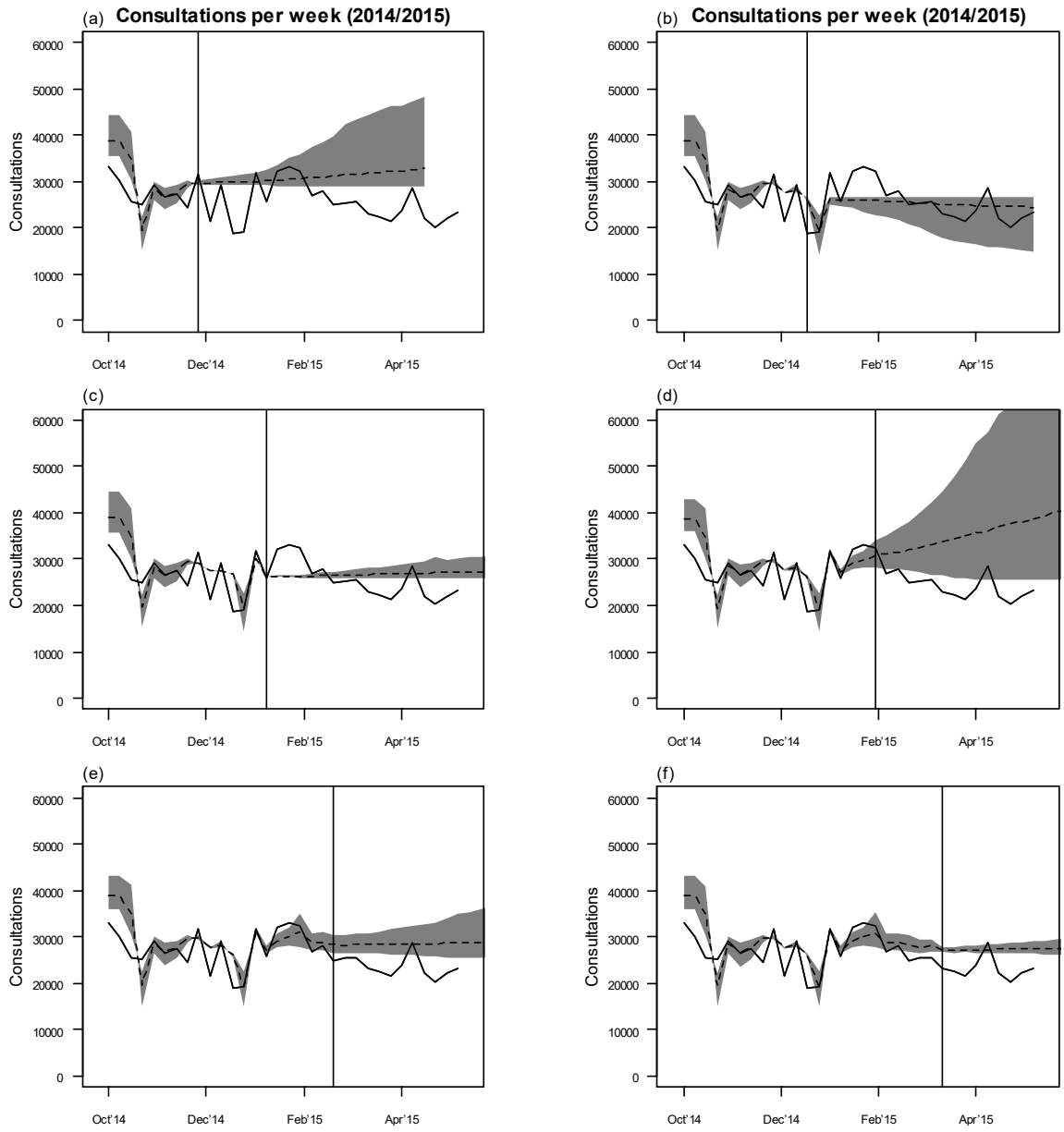


Figure 5.15 – Prediction consultation plots (2014/2015) at different time points when the model was fitted through the time-dependent linear regression model (joint model) between consultations (predicting variable) and Diagnosed ILIs (independent variable) as defined in Chapter 4, where chart (a) is using 9 data points (vertical line) and predicting the next 20, (b) is using 12 data points (vertical line) and predicting the next 20, (c) is using 15 data points (vertical line) and predicting the next 20, (d) is using 18 data points (vertical line) and predicting the next 20, (e) is using 21 data points (vertical line) and predicting the next data points up to the end of the influenza season, (f) is using 24 data points (vertical line) and predicting the next data points up to the end of the influenza season. Each plot includes the actual data (solid line) and the results of the model fit, dashed line (mean) and shaded area (95% high predictive density region).

5.4 Discussion

Real-time forecasting is challenging and is even more complex when data is limited. In our datasets, the number of reported consultation and diagnosed cases decreased over the years, thereby making the analysis even more difficult. In this chapter, we extended the

work of the previous chapters by producing real-time forecasting, thereby predicting the subsequent number of data points, based on a certain number of known weekly cases. In the previous chapters, we used all the data points to try to find the best SEIR model fit. This was done successfully, enabling us to understand the relationship between different datasets and to understand the parameter values of the influenza datasets. In this chapter we showed how different datasets and different models (SEIR model and joint model) perform when applying forward prediction.

Unfortunately, the above real-time forecasts are not always reliable. In fact, some forecasts are quite misleading when compared to the actual data. These forecasts are similar with the research published by Ong et al. (2010) [26], which also utilized the SEIR model and the particle filtering algorithm. The findings in the latter study show an overestimation for the reported ILIs. Similar to some of our findings, the end period of the outbreak was then modelled accurately [26].

The findings show that for higher number of consultation and diagnosed cases, the real-time forecasts are improved (2009/2010 datasets). Since the linear relationship between the consultations and diagnosed is strong for the pandemic data, this improved the (joint model) consultations real-time forecasts, as can be seen from the joint model predictions in figure 5.3. These results are similar with those obtained in chapter 4. It was established previously (in Chapter 4) that the higher the number of reported cases (by GPs), the better the relationship between consultation and diagnosed cases. When the relationship is strong between consultations and diagnosed, the level of certainty of the predictions is improved. In fact, the confidence intervals for the consultations data (through the joint model, Figure 5.3) are narrower when compared with the SEIR model fit of the consultations data (Figure 5.1).

The results of chapter 5 can serve as a good basis to decide when the joint model is producing satisfactory result. As established in chapter 4, when there is a higher ratio between diagnosed and consultations $\left({}^{diagnosed}/_{consultations}\right)$, this provides a stronger correlation value, hence predicting the dependent variable satisfactorily. For the last three datasets (2012/2013, 2013/2014 and 2014/2015), the general linear regression model provided a weak association (as described in Chapter 4, Table 4.1). However, since

the joint model is assuming a time-dependent linear regression model, the consultations dataset (2013/2014 and 2014/2015) still produced reasonable real-time forecasts.

Through the joint model, the diagnosed data can provide various signals to the consultations data. For example, since the slopes of the linear regression model are positive in general (see Chapter 4 for further information), then if the diagnosed data is predicting an outbreak, the consultations can easily follow with an outbreak as well, thereby predicting a high demand on doctors and the health sector. On the contrary, a low number and stationary forecasts for the diagnosed dataset also imply low number and stationary forecasts for the consultation cases (Figure 5.9 and Figure 5.15). A limitation for the joint model is that overestimation of the real-time forecasts for the diagnosed datasets (independent variable) are more likely to imply weak forecasts for the consultations dataset (Figure 5.9), since consultations are dependent on the diagnosed forecasts. Furthermore, negative relationship (slope) between consultation and diagnosed variables are more likely to imply unreasonable confidence intervals as well.

In chapter 4, we introduced an innovative extension (joint model) to the technique introduced in Ong et al. (2010) [26]. Following the findings in this chapter, this methodology does not always provide acceptable results. Hence, a further understanding is required as to whether it is possible to develop an improved and consistent framework that shall serve as a better tool for an early warning signal to predict the outbreak.

Chapter 6

Sensitivity Analysis

6.1 Introduction

In chapters 3 and 4, I showed that the SEIR model can accurately describe the observed datasets. Our findings showed that posterior parameter distributions are consistent between the diagnosed datasets for different years, while some posterior parameter distributions were similar for different proxies of the pandemic data (Chapter 3). However, my model used non-informative prior distributions for parameters and initial state values of the SEIR model (as described in chapter 2): the initial number of susceptible individuals, $S(0)$, the initial number of exposed individuals, $E(0)$, the initial number of infected individuals, $I(0)$ and the initial number of recovered individuals $R(0)$. $S(0)$, $E0)$, $I(0)$ and $R(0)$ are not directly observable [26] and so it is important to consider how sensitive the results are, to changes in these values. So far, I assumed that the individual mean values of $E(0)$ and $I(0)$ are equal to the initial number of consultations/diagnosed cases at the start of the outbreak. This means that $E(0) = I(0)$ since the same individuals who are infected but not infectious (E) eventually become infected but infectious (I). Moreover, the value of $R(0)$ was assumed equal to 65,000 (Chapter 3) as this resulted in a reasonable fit. The value of $S(0)$ follows from the previous three values (Population size (N) - $E(0)$ - $I(0)$ - $R(0)$). Several studies assume $R(0)$ equal to 0 [26, 56, 147], while other studies assume $R(0)$ equal to the number of infectious individuals recorded at the start of the analysis [148].

The effective reproduction ratio was one of the central points of the analysis in chapters 3 and 4, since this is the main measure of the severity of the outbreak. In chapters 3 and 4, we observed that the effective reproduction ratio started initially with a high value. This result has also been observed in Ong et al. (2010) [26]. However, such a high value appears to be unrealistic and so further investigation is needed. In this chapter, I will examine to what extent the initial prior mean values of $S(0)$, $E0)$, $I(0)$ and $R(0)$ influence the estimated value of R_t particularly for low t values. Throughout this chapter, reference to the 'prior value' implies the mean value of the prior distribution.

6.2 Sensitivity Analysis for $R(0)$

For seasonal influenza datasets, the prior value of $R(0)$ is the total number of removed (immune) individuals at $time = 0$, at the start of the outbreak. If the individuals are part of the removed compartment, then automatically they are not part of the susceptible

compartment. As discussed in chapter 2, the SEIR model is assuming that after an individual acquires and recovers from seasonal influenza, the individual becomes immune. However, immunity from the influenza can be acquired through vaccination for the same influenza virus [149], also leading to a low number of susceptible individuals. Thus, amendments in the $R(0)$ values directly influence the value of $S(0)$. Therefore, throughout this section, although we will be changing directly the value of $R(0)$, the sensitivity analysis is also applied on the $S(0)$ value.

For the scope of this analysis, for all influenza datasets which were defined in the previous chapters, the SEIR model in combination with the particle filter algorithm (as defined in chapter 2) is used. For parameters other than $S(0)$, $E(0)$, $I(0)$ and $R(0)$, the same prior values, as defined in chapter 3, are used. Similar to chapter 3 [79], throughout this section, the prior distributions of $I(0)$ and $E(0)$ are assumed equal to the number of confirmed cases (consultations or diagnosed) at the start of the influenza outbreak. The prior is assumed to be normally distributed, with mean and variance derived from the observed values of consultations and diagnosed (depending on the dataset being used). However, we vary the balance between $R(0)$ and $S(0)$ while keeping $S(0) + E(0) + I(0) + R(0) = N$ constant. For every dataset, the model is applied six times, that is, for $R(0)$=0 ($S(0)$=414,000-$E(0)$-$I(0)$), $R(0)$=50,000 ($S(0)$=414,000-50,000-$E(0)$-$I(0)$), $R(0)$=100,000 ($S(0)$=414,000-100,000-$E(0)$-$I(0)$), $R(0)$=150,000 ($S(0)$=414,000-150,000-$E(0)$-$I(0)$), $R(0)$=200,000 ($S(0)$=414,000-200,000-$E(0)$-$I(0)$) and $R(0)$=250,000 ($S(0)$=414,000-250,000-$E(0)$-$I(0)$). There was only one exception: for the 2009/2010 weekly diagnosed ILIs, we used $R(0)$=350,000 (($S(0)$=414,000-350,000-$E(0)$-$I(0)$) instead of $R(0)$=200,000 since a higher number of $R(0)$ was required to examine a larger difference between the R_t values. In the above calculations for different priors, 414,000 (N) is the population size of Malta.

As described in chapters 2, 3 and 4, different seasons of influenza varied in strength of their outbreak. In fact, for the pandemic season, a higher number of consultations and diagnosed cases were recorded, while for the diagnosed ILI datasets, the number of cases decreased across the years. It is important to note that for the 2009/2010, there was already a major epidemic before October 2009 such that during July 2009, the highest number of diagnosed cases were recorded by the GPs (Figure 1, Chapter 3).

The first two datasets that were taken into consideration were the consultations for the 2009/2010 season and the diagnosed dataset for the same season. As defined above, the SEIR model was run for every dataset for different $R(0)$ values, and the dataset for the effective reproduction ratio R_t was recorded for every single case. Figure 6.1 shows that the initial reproduction ratios R_t are highly dependent on the initial number of the removed individuals, but the dependence is largely diminished later in the epidemic. In fact, figure 6.1 shows that for the consultations data, the first three values of the effective reproduction ratio varies between 0.8 and 3.3 for different values of $R(0)$, while for the diagnosed data, the initial three values vary between 0.8 and 6.9 (Figure 6.1). During the initial stages of the outbreak, data is very limited and hence, the model parameters are being estimated based on very limited information. Therefore, a lot of variation exists at this stage until the model starts to stabilize due to further knowledge of the performance of the outbreak. This is one of the main strengths of the particle filter algorithm, where the parameter estimates are further refined as the outbreak unfolds. For the consultations dataset, the 4^{th} reproduction ratio number declines substantially to a value between 1.3 and 1.9 for different values of $R(0)$, while the diagnosed data declines considerably to values between 2.4 and 3.5 for different values of $R(0)$. From the 4^{th} point onwards, differences in the effective reproduction ratios for different values of $R(0)$ are substantially small.

One of the arising questions centres on what value of $R(0)$ to choose for further parameter estimation and prediction. Such a question is challenging as one needs to consider this by placing the $R(0)$ value in context. For example, one cannot assume that $R(0)$=400,0000, as this implies that $S(0)$<14,000 (414,000-400,000-$E(0)$-$I(0)$). It is very unlikely that less than 14,000 individuals will be susceptible to acquiring the influenza during the season, as one can see from all influenza datasets in this thesis. On the other hand, very low values of $R(0)$ produce unreasonably high initial values of R_t, which are substantially higher when compared to any other value of the effective reproduction ratio throughout the outbreak. Very low $R(0)$ values can be seen as a worst case scenario when forecasting data [26], while very high values imply that few people will get infected from the disease. Hence, one needs to establish the right balance between the results presented above.

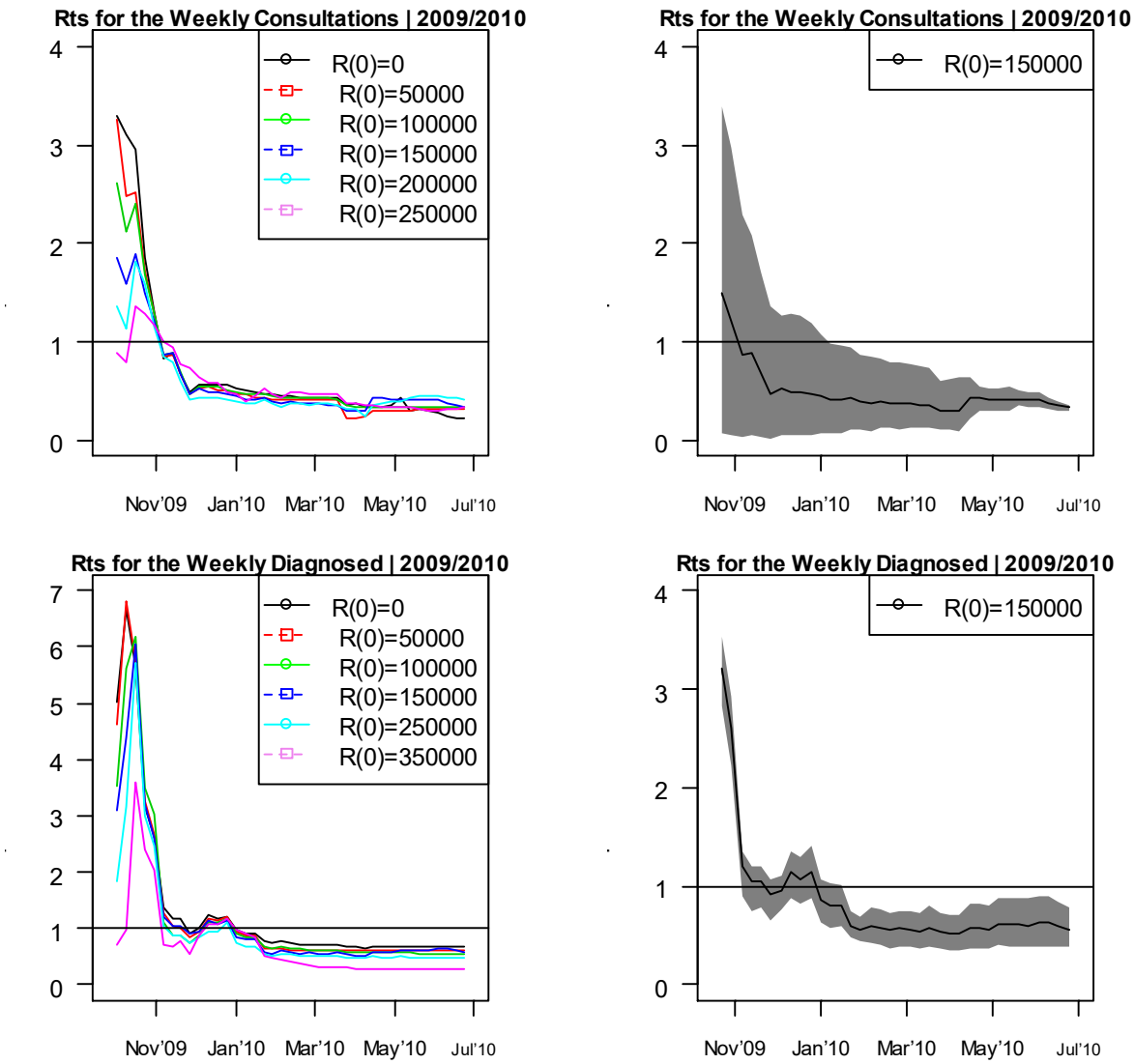


Figure 6.1–Sensitivity analysis for $R(0)$ in relation to the reproduction number for the consultations and diagnosed variables during 2009/2010. Charts on the right represent the R_t plot when $R(0)$=150000 without the first three R_t data points. From the 4th R_t point onwards, R_t values tend to stabilize for all different $R(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

In figure 6.1, one can clearly see that when $R(0)$=250,000 (for consultations data) and $R(0)$=350,000 (for diagnosed data), the R_t values lose some consistency (when compared with other $R(0)$ values) for a number of points throughout the outbreak (diagnosed from January 2010 onwards, Figure 6.1). For few diagnosed data points (January 2010) when $R(0)$=250,000, the R_t values are slightly lower than the other R_t values for different $R(0)$ values. Furthermore, assuming that 250,000 from a population of 414,000 are immune at the start of the outbreak is rather optimistic. Hence, based on the above observations and assumptions, $R(0)$=150,000 should be a fairly reasonable value for the 2009/2010 season (Figure 6.1). The charts on the right hand side of figure 6.1 portray the reproduction ratio

chart for $R(0)$=150,000 without the first three R_t values (as described above) but with the relevant 95% confidence intervals.

For the 2011/2012 datasets (consultations and diagnosed), the SEIR models were run again for all the different $R(0)$ values. Similarly as above, the initial R_t values appear inconsistent (Figure 6.2). In fact, for the consultations dataset, the first three initial values vary between 1 and 3.7. From the 4th point onwards, the effective reproduction ratio stabilize for different values of $R(0)$. For $R(0)$ =200,000 and $R(0)$=250,000, the effective reproduction charts vary when compared to the other charts. For the consultations dataset, most variation in the R_t values occur during the peak of the outbreak (March 2012), where for the highest two $R(0)$ values, the reproduction number is smaller. Therefore, up to $R(0)$=150,000 the effective reproduction ratios are relatively consistent for different values of $R(0)$.

The initial three R_t values for the diagnosed dataset are also inconsistent, varying between 1.3 and 7.8. In contrast to the previous cases, all R_t charts provide very consistent values from the 4th point onwards for all $R(0)$ values. As defined above, assuming high values of $R(0)$ would not be realistic; hence, in this case, our decision should be based on the consultations dataset ($R(0)$=150,000, Figure 6.2). The number of removed individuals for the same disease should be identical irrespective to the type of proxy being utilised.

Similarly to the 2011/2012 consultations dataset, figure 6.3 shows that for the consultations 2012/2013 dataset, the first three initial R_t values vary between 1.1 and 3.9, while the R_t value for the 4th point varies between 1.1 and 1.5 for different $R(0)$ values. From this point onwards, the effective reproduction ratio is more stable.

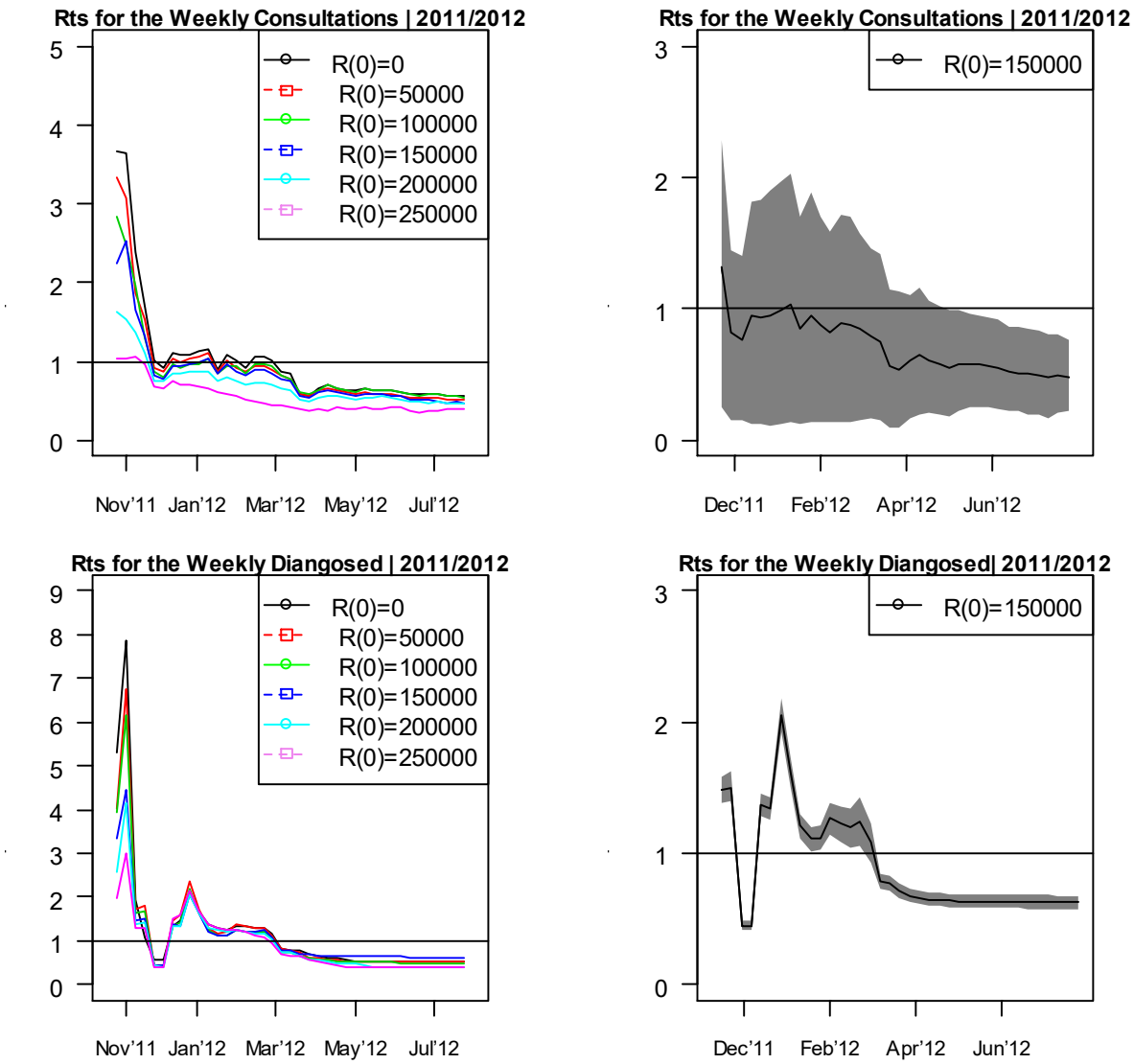


Figure 6.2– Sensitivity analysis for $R(0)$ in relation to the reproduction number for the consultations and diagnosed variables during 2011/2012. Charts on the right represent the R_t plot when $R(0)$=150000 without the first three R_t data points for consultations and without the first two R_t data points for the diagnosed dataset. After these, the initial R_t values tend to stabilize for all different $R(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

For the diagnosed data, the initial values are more inconsistent when compared with the consultations data (Figure 6.3). In fact the 4th reproduction ratio still varies between 1.6 and 3.6 for different $R(0)$ values. However, for the 5th point, the R_t values stabilizes between 1.1 and 1.8. From this point onwards, the effective reproduction ratio is consistent for different $R(0)$ values. For the consultations data, the effective reproduction ratio is lower for the highest two $R(0)$ values ($R(0)$=200,000 and $R(0)$=250,000), during the peak of the influenza (February 2013). Hence, up to $R(0)$=150,000, the R_t values are consistent. Thus, for this influenza season we will be assuming that $R(0)$=150,000 (Figure 6.3).

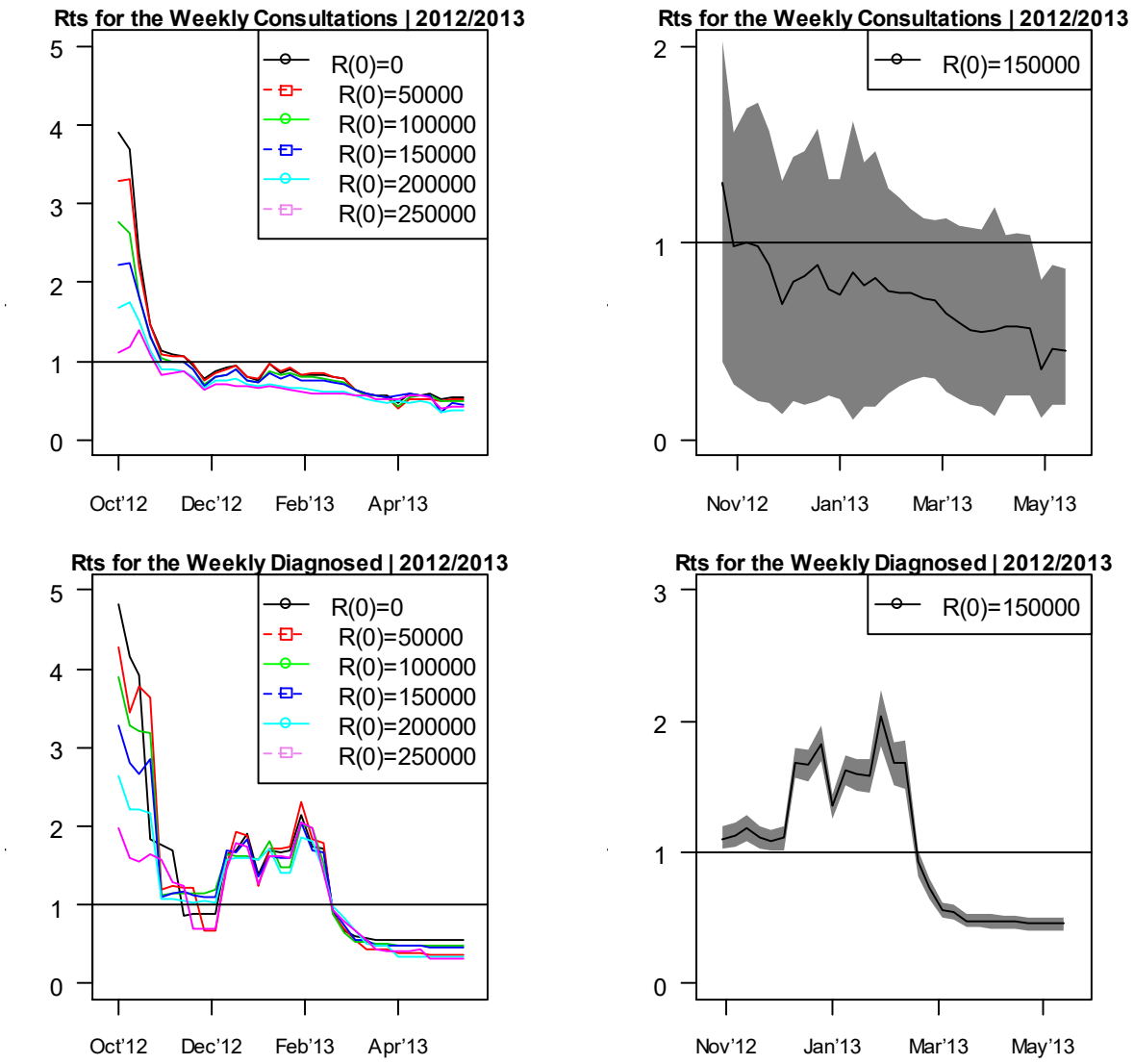


Figure 6.3– Sensitivity analysis for $R(0)$ in relation to the reproduction number for the consultations and diagnosed variables during 2012/2013. Charts on the right represent the R_t plot when $R(0)$=150000 without the first three R_t data points for consultations and without the first four R_t data points for the diagnosed dataset. After these, initial R_t values tend to stabilize for all different $R(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

Figure 6.4 shows the sensitivity analysis for $R(0)$ (and $S(0)$) for the 2013/2014 dataset. For the consultations dataset, the first two R_t time points for different $R(0)$ values vary between 1.3 and 4.1, while from the 3rd point onwards the R_t values are consistent for different $R(0)$ values until December 2013. During the January-February 2014 timeframe, there are some inconsistencies. For $R(0)$=150,000 and above, the inconsistencies (R_t) are more apparent. Furthermore, for $R(0)$=100,000 the effective reproduction ratio produced a rare peak for the consultations data. This is the only consultations data that produced an R_t>1 that coincides with the high R_t values of the diagnosed dataset (January-February 2014) occurring during the peak of the influenza season. The effective reproduction ratio

chart (Figure 6.4) for $R(0)$=100,000 seems to be a reasonable option to choose due to this particular result.

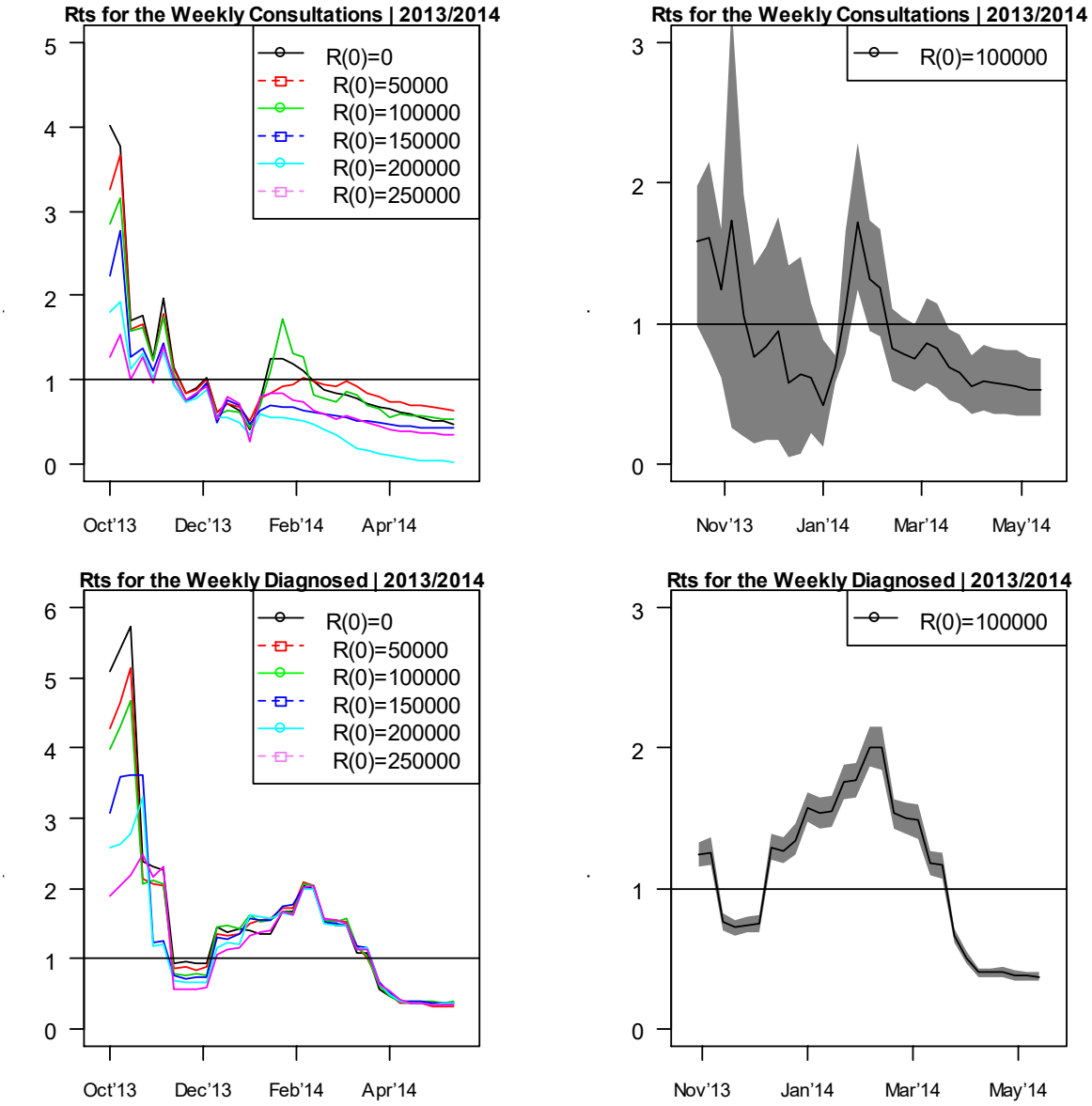


Figure 6.4– Sensitivity analysis for $R(0)$ in relation to the reproduction number for the consultations and diagnosed variables during 2013/2014. Charts on the right represent the R_t plot when $R(0)$=100000 without the first two R_t data points for consultations and without the first four R_t data points for the diagnosed dataset. After these, initial R_t values tend to stabilize for all different $R(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

Similarly to the previous cases, for the diagnosed 2013/2014 dataset, the first initial R_t values are inconsistent (Figure 6.4) such that up to point 4 the initial R_t values vary substantially for different $R(0)$ values. In general, the R_t charts (Figure 6.4, diagnosed dataset) for different $R(0)$ from the 5th point onwards are fairly consistent, except for November 2013 ($R(0)$=150,000 and $R(0)$=200,000).

The latest seasonal influenza datasets (2014/2015) is similar to the other datasets. In fact, for the consultations dataset, only the first two R_t values are inconsistent for different $R(0)$ values, while for the diagnosed dataset the first five data points are inconsistent (Figure 6.5). In general, the diagnosed datasets include more inconsistent initial R_t values when compared to the consultations datasets. However, the 95% confidence intervals for the consultations data are wider when compared to the diagnosed data. This shows that once the effective reproduction ratio through the diagnosed ILI datasets are stabilized, the R_t values incorporate more certainty than the R_t values through the consultation datasets. This can be attributed to the fact that the diagnosed dataset is a clearer signal of the strength of the influenza when compared to the consultations dataset. The latter dataset includes a substantial amount of background rate and sub-clinical cases, as discussed in chapter 4. Hence, this increases the uncertainty in the consultations dataset. Figure 6.5 does not show a clear direction on which best value of $R(0)$ to use. Hence, based on previous seasonal influenza outbreaks, it is reasonable to choose $R(0)$=150,000 as the best prior mean value for the removed compartment of the SEIR model.

In general, for most consultation datasets, only the initial values exceed the value of one, while for the diagnosed data, there is more variation of the effective reproduction number throughout the outbreak (Figure 6.5).

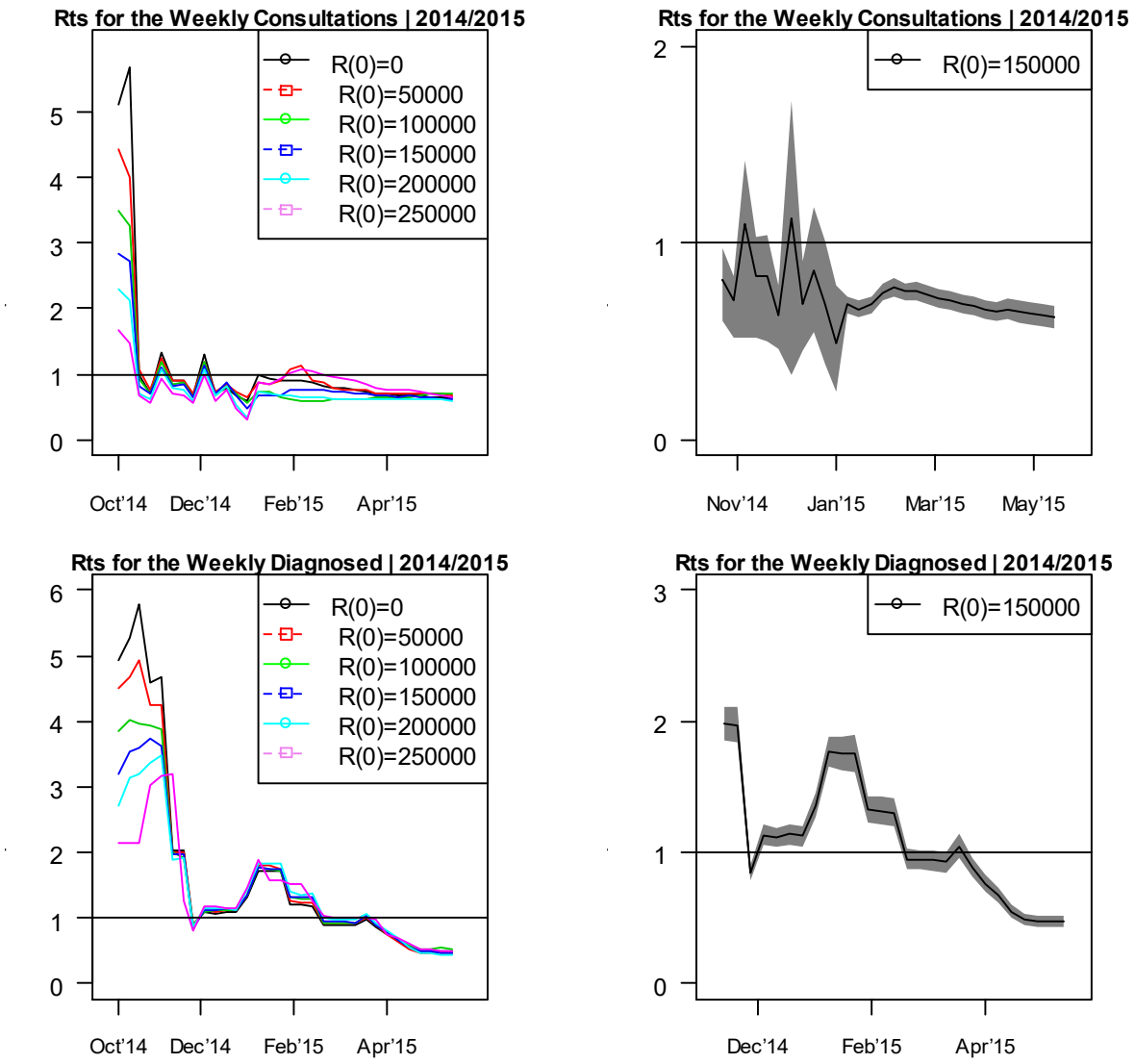


Figure 6.5– Sensitivity analysis for $R(0)$ in relation to the reproduction number for the consultations and diagnosed variables during 2014/2015. Charts on the right represent the R_t plot when $R(0)$=150000 without the first two R_t data points for consultations and without the first five R_t data points for the diagnosed dataset. After these, initial R_t values tend to stabilize for all different $R(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

6.3 Sensitivity Analysis for $I(0)$ and $E(0)$

Throughout this section, we aim to understand the sensitivity of the results to the mean prior values of $I(0)$ and $E(0)$ for the consultation and diagnosed datasets. As mentioned above, so far $I(0)$ and $E(0)$ were assumed equal to the number of observed cases at the start of the epidemic for both consultation and diagnosed datasets. As discussed in the previous section, a change in the values of $I(0)$ and $E(0)$ influence the value of $S(0)$. In fact, the higher the values of $I(0)$ and $E(0)$, the lower is the value of $S(0)$ (=Population size (N) - $E(0)$ - $I(0)$ - $R(0)$). Based on the previous section, the $R(0)$ will be assumed

equal to the final selected value for each dataset (Figure 6.1-6.5). Furthermore, we shall assume the initial value of $E(0)$ equal to the new selected value of $I(0)$. In order to test for the sensitivity of $I(0)$ and $E(0)$, several different values of $I(0)$ and $E(0)$ will now be considered. Note that throughout this section, any reference to the value of $I(0)$ also refers to the value of $E(0)$. Hence, in order to simplify the interpretation of the analysis in this section I will focus on the value of $I(0)$. As defined in the previous section, the same SEIR model and particle filter algorithm are used throughout the following analysis.

The number of consultations on a weekly basis is substantially higher than the number of diagnosed individuals (as it includes non-influenza and sub-clinical cases); hence the prior values of $I(0)$ and $E(0)$ for the consultation datasets will be tested for higher values. The initial number of weekly diagnosed ILI cases vary between 100 and 1500, as showed in the previous chapters. The initial number of consultations vary between 29000 and 42000, except for the pandemic 2009/2010 season, where these thresholds are exceeded substantially for the initial values (week 41 $\approx$ 79,000 cases, Figure 4.1). However, the consultations data include a substantial amount of background cases (as discussed in Chapter 4). In fact, even when modelling the consultations data, we assume a much higher number of background cases (compared with the diagnosed dataset). However, it is important to note that the mean prior for the reporting rate for the consultations is assumed to be 0.75, while for the diagnosed data it is being assumed equal to 0.25. Hence, this reflects the larger number of consultation cases (including the non-influenza cases) when compared to the diagnosed dataset.

For all the five different diagnosed datasets, the model is applied for $I(0)$=0 ($S(0)$=414,000-$E(0)$-$R(0)$), $I(0)$=1,000 (S(0)=414,000-1,000-$E(0)$-$R(0)$), $I(0)$=5,000 ($S(0)$=414,000-5000-$E(0)$-$R(0)$), $I(0)$=8,000 ($S(0)$=414,000-8,000-$E(0)$-$R(0)$), $I(0)$=10,000 ($S(0)$=414,000-10,000-$E(0)$-$R(0)$) and $I(0)$=15,000 ($S(0)$=414,000-15,000-$E(0)$-$R(0)$). For the consultation datasets, the model was applied for $I(0)$=0 ($S(0)$=414,000,000-$E(0)$-$R(0)$), $I(0)$=5,000 ($S(0)$=414,000-5,000-$E(0)$-$R(0)$), $I(0)$=10,000 ($S(0)$=414,000-10,000-$E(0)$- $R(0)$), $I(0)$=15,000 ($S(0)$=414,000-15,000-$E(0)$-$R(0)$), $I(0)$=35,000 ($S(0)$=414,000-35,000-$E(0)$-$R(0)$) and $I(0)$=50,000 ($S(0)$=414,000-50,000-$E(0)$-$R(0)$). However, sensitivity analysis for the 2009/2010 weekly consultations is applied for $I(0)$=0, 15000, 25000, 35000, 50000 and 60000, since the consultations for the pandemic season are substantially higher than the number of consultations for the

seasonal influenza datasets. Hence, $I(0)$ was required to be higher in order to observe differences between the R_t values for different $I(0)$ values. As mentioned above $E(0) = I(0)$ for all cases.

As observed in the previous section, a lot of variation exists in figure 6.6 for the initial values of R_t. In fact, for the consultations data, the first three R_t values vary between 1.5 and 4.1, while for the diagnosed data the initial three R_t values vary between 3 and 10.3. Similarly as before, the initial R_t values for the diagnosed dataset vary more than the R_t values for the consultations data. For the consultation dataset (Figure 6.6), there is no clear trend associated in relation to the change in $I(0)$ value. In fact, for the highest value of $I(0)$ (60,000), the R_t plot lies somewhere in between the other plots. Furthermore, for different values of $I(0)$, the R_t plots are rather consistent with some minor variations during January 2010 (for $I(0)$=0 and 60,000). Hence, since limited differences exist for different $I(0)$ values, then we can assume the mean prior value of $I(0)$ to be equal to the actual number of reported consultations (42,038) at the start of the outbreak (as assumed in chapters 3 and 4). One can interpret this value as the most informative prior of $I(0)$ as it is based on actual observed data.

The initial R_t values for the diagnosed dataset (Figure 6.6) shows that for the highest $I(0)$ value (15,000), the R_t values are substantially higher when compared to the other $I(0)$ values. For the lowest five $I(0)$ values, the first R_t values are rather close (between 3 and 3.6). This shows that the R_t values are not dependent on $I(0)$ for such lower values. Furthermore, for $I(0)$ values of 10,000 or lower, the R_t plots for the diagnosed data (Figure 6.6) are rather consistent. Thus, the initial observed diagnosed ILI value (2,700) is within the level of consistency for different $I(0)$ values between 0 and 10,000 (Figure 6.6). Hence, since the actual number of diagnosed individuals at *time = 1* is the most reliable available information, then we can assume this value to be the most reasonable mean prior value for $I(0)$ (Figure 6.6).

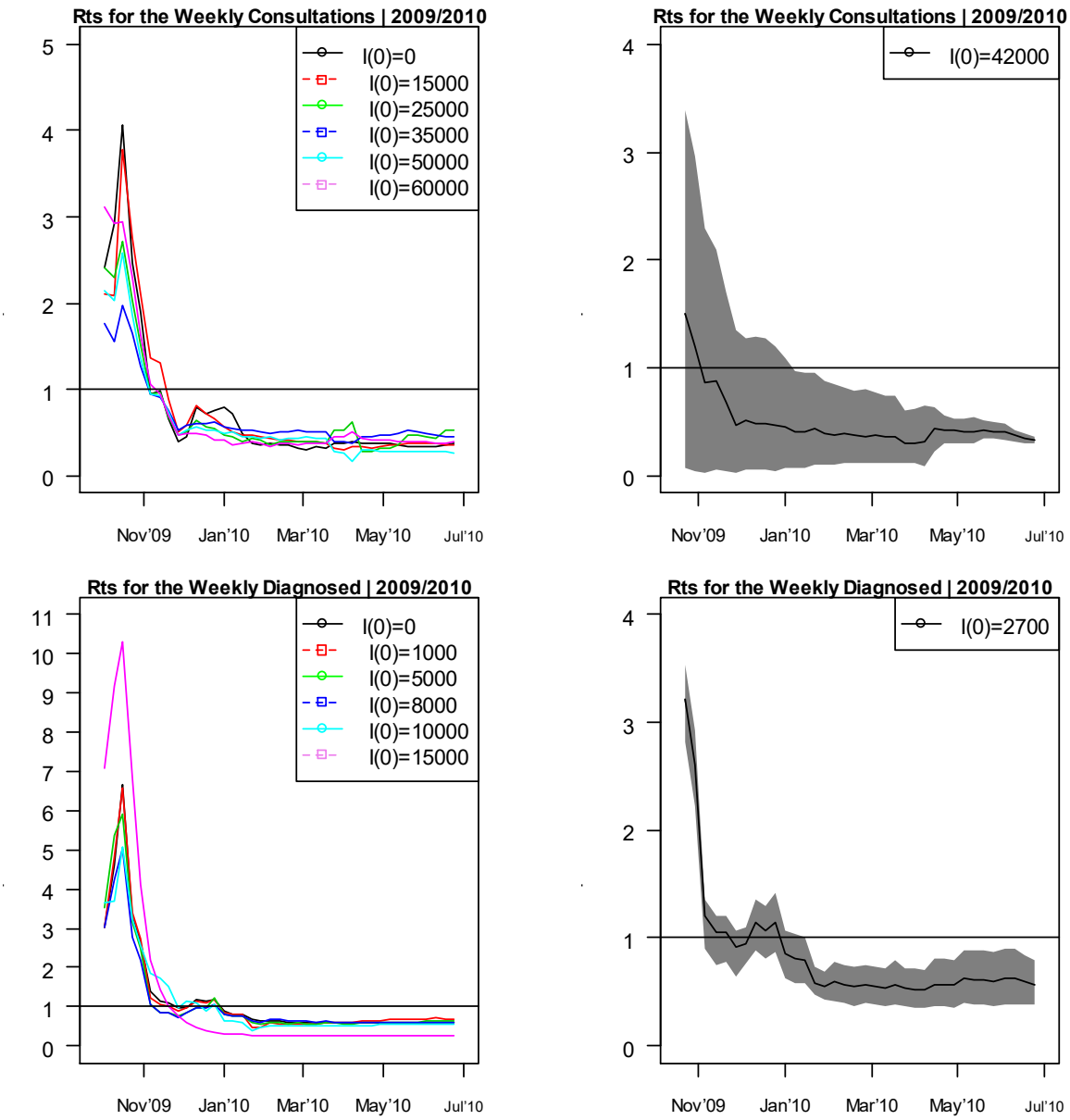


Figure 6.6–Sensitivity analysis for $I(0)$ and $E(0)$ in relation to the reproduction number of the consultations and diagnosed 2009/2010 datasets. Charts on the right represent the R_t plot when $I(0)$ and $E(0)$ are equal to the number of confirmed cases at the start of the outbreak, without the first three R_t points. From the 4th R_t point onwards, R_t values tend to stabilize for all different $I(0)$ and $E(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

For the 2011/2012 consultations data, little variation exists in the initial R_t values (Figure 6.7) for different $I(0)$ values. Different R_t charts are consistent; however, for lower values of $I(0)$, some R_t values are greater than 1 (February-March 2012), while for the two highest $I(0)$, R_t values are less than 1 for most of the outbreak. These highlight further the uncertainty of the R_t values for the consultations data. However, since the R_t values for $I(0)$ up to 15,000 are greater than 1 during the peak of the influenza, hence it is more reasonable to assume such a mean prior $I(0)$ value for this particular dataset.

Nevertheless, the R_t values for the consultations data (when $I(0)$=15,000) include substantially a wider confidence interval when compared to the diagnosed data.

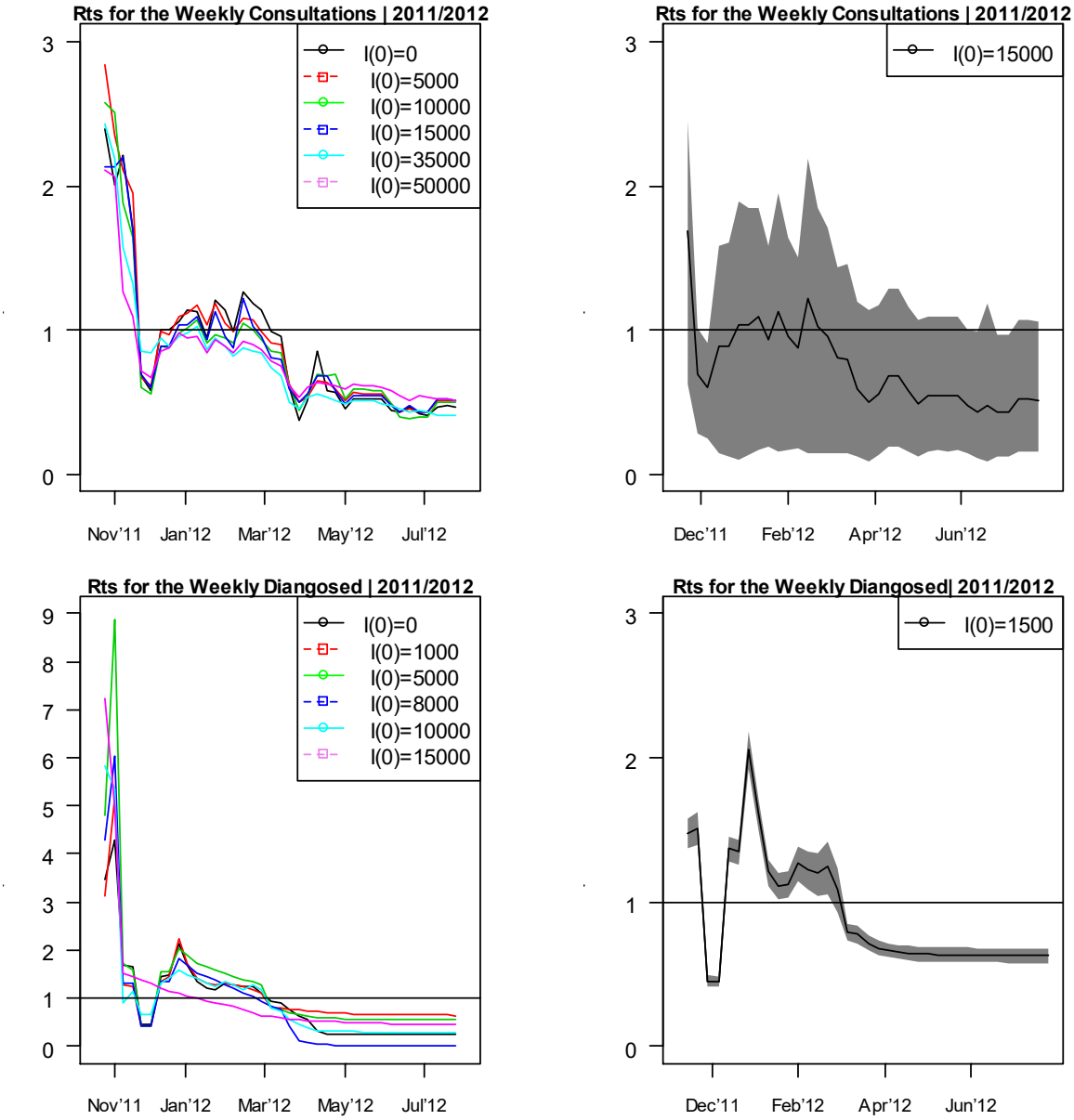


Figure 6.7–Sensitivity analysis for $I(0)$ and $E(0)$ in relation to the reproduction number for the consultations and diagnosed 2011/2012 datasets. Charts on the right represent the R_t plot when $I(0)$ and $E(0)$ are equal to 15,000 for the consultations dataset, without the first three R_t points, while $I(0)$ and $E(0)$ are equal to the number of confirmed cases at the start of the outbreak for the diagnosed dataset, without the first two R_t points. After these initial points, R_t values tend to stabilize for all different $I(0)$ and $E(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

For the 2011/2012 diagnosed dataset, the highest value of $I(0)$ produced different R_t values when compared with the other $I(0)$ values (Figure 6.7). In general, for lower values of $I(0)$, the R_t plots are consistent. Hence, we can assume that the number of observed

diagnosed cases at the start of the outbreak is the most reliable mean prior for $I(0)$ (Figure 6.7). Similarly to the above results, the initial R_t values vary substantially for the first two data points.

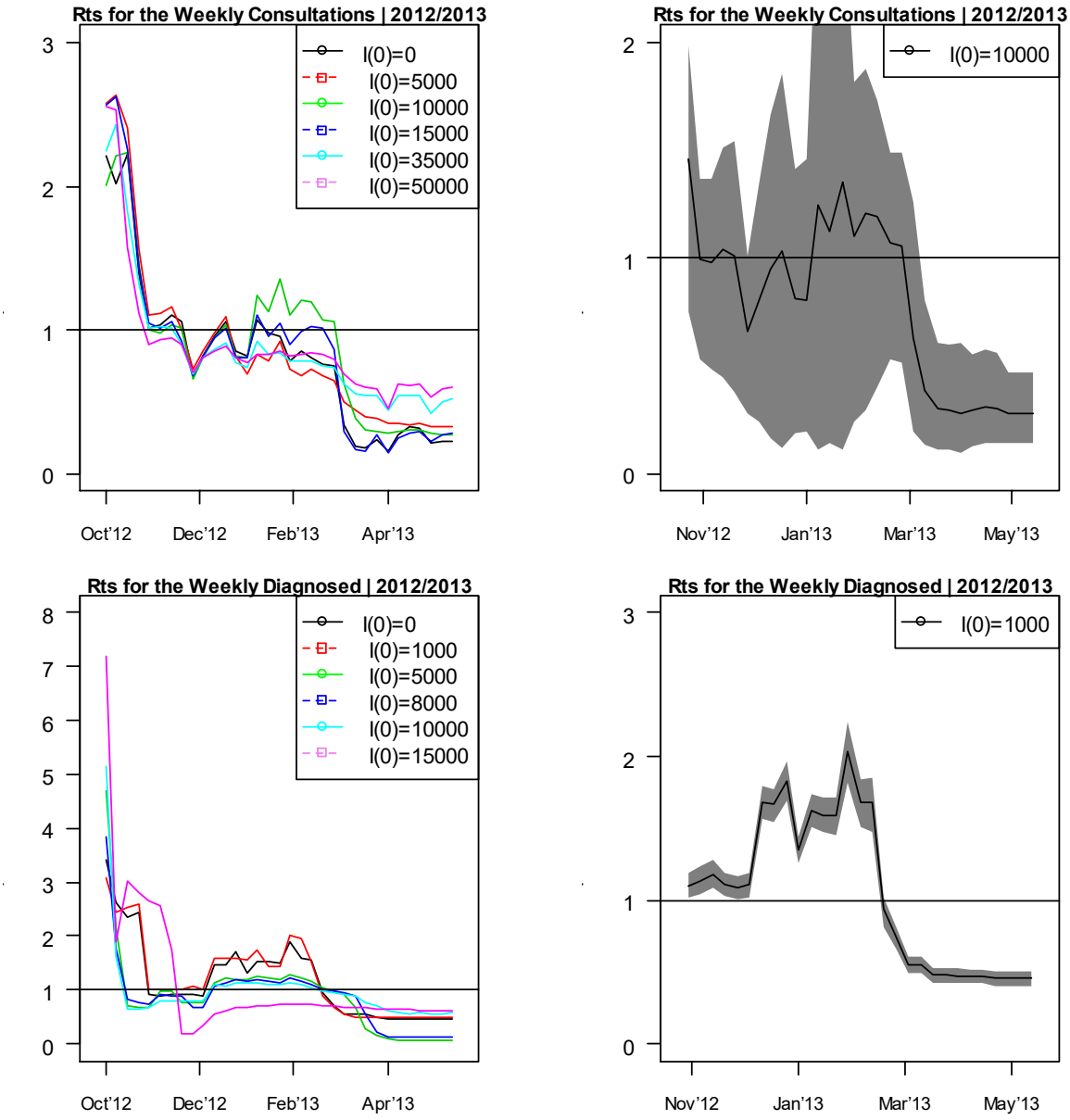


Figure 6.8– Sensitivity analysis for $I(0)$ and $E(0)$ in relation to the reproduction number for the consultations and diagnosed 2012/2013 datasets. Charts on the right represent the R_t plot when $I(0)$ and $E(0)$ are equal to 10,000 for the consultations dataset, without the first three R_t points, while $I(0)$ and $E(0)$ are equal to the number of confirmed cases at the start of the outbreak for the diagnosed dataset, without the first four R_t points. After these initial points, R_t values tend to stabilize for all different $I(0)$ and $E(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

Figure 6.8 shows that there is a substantial difference in the R_t plots during the peak of the influenza for different $I(0)$ values for both datasets. In particular, for the consultations data where $I(0)$=10,000, the R_t portrays a clear major outbreak during the peak of the

season. Hence, this prior mean value is a reasonable choice for this consultations dataset. Nevertheless, high variation exits during the peak of the influenza, as one can see in the 95% confidence interval for the final R_t plot for the consultations data (Figure 6.8). Furthermore, the initial R_t values stabilize after the 3rd point when compared with different values of $I(0)$.

For the diagnosed cases the higher the $I(0)$ value, the higher the initial R_t values (Figure 6.8). Figure 6.8 shows that there is some substantial variation in the R_t plots for different $I(0)$ values. When $I(0)$=15,000 the R_t plot seems to be entirely different when compared to the other plots (also seen in previous results). For the next two lower $I(0)$ values (10,000 and 8,000), the effective reproduction ratio shows a less powerful outbreak when compared to the two lowest $I(0)$ values (0 and 1,000). In such a case where there is substantial variation in the R_t between different $I(0)$ values, it reasonable to assume $I(0)$ equal to the number of confirmed cases (≈1,000) at the start of the outbreak (Figure 6.8).

Figure 6.9 which presents the consultations dataset, shows some inconsistencies for the first two initial R_t values. From the third point onwards the R_t plots are similar for different $I(0)$ values, except during January-February 2014 period. On the contrary, for the diagnosed dataset substantial variation exists in the first three R_t values (varying between 1.2 – 7.4). For the highest $I(0)$ values (15,000 and 10,000), the R_t plots are rather inconsistent when compared to other R_t plots for different $I(0)$ values, while for the three lowest $I(0)$ values, the R_t plots are more consistent. Hence, due to the above reasons it is more realistic to assume $I(0)$ equal to the number of observed cases at the start of the outbreak.

In contrast to the diagnosed data, in general the consultations data for higher values of $I(0)$ does not influence the initial value of R_t. This fact can clearly be seen in figure 6.10, where the initial values are rather close. After ignoring the first two R_t values, the R_t drops below the value of one, indicating that the initial R_t values are rather unrealistic when compared with the other R_t values. Most inconsistencies between the R_t plots appear for the highest $I(0)$ values (50,000 and 35,000), while for lower values, the R_t plots are consistent. For $I(0)$ equal to the number of confirmed cases (33,000) at the start of the outbreak (Figure 6.10), the final R_t plot results in the same shape for the first four lower $I(0)$ values.

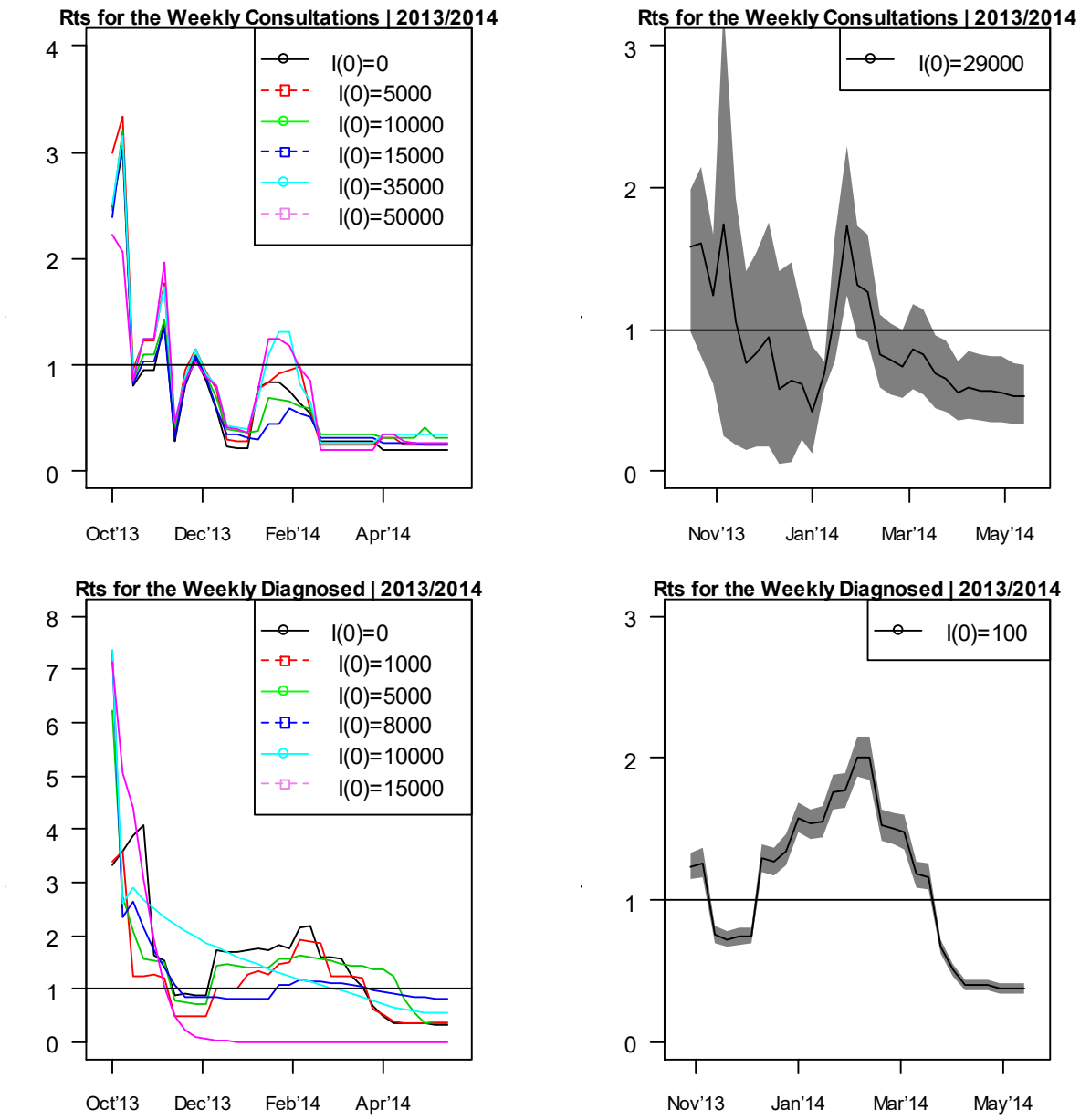


Figure 6.9– Sensitivity analysis for $I(0)$ and $E(0)$ in relation to the reproduction number for the consultations and diagnosed 2013/2014 datasets. Charts on the right represent the R_t plot when $I(0)$ and $E(0)$ are equal to the number of confirmed cases at the start of the outbreak, without the first two R_t points for the consultations data and without the first four R_t points of the diagnosed data. After these initial points, R_t values tend to stabilize for all different $I(0)$ and $E(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

Similarly to the previous diagnosed dataset, 2014/2015 dataset produced substantial inconsistencies in the initial R_t values (Figure 6.10). Figure 6.10 shows that for higher $I(0)$ values, the initial R_t values are also higher, where for $I(0)$ = 15,000, the initial R_t value exceeds the value of 7. Moreover, the general R_t plot for $I(0)$ = 15,000 is inconsistent when compared with all the other R_t plots for the diagnosed 2014/2015 dataset. The latter results were observed for all the diagnosed datasets throughout this section. The results in figure 6.10 suggest that the most reasonable value of $I(0)$ is equal

to the initial number of confirmed cases (450) at the start of the outbreak. The R_t plots are consistent for the lower values of $I(0)$.

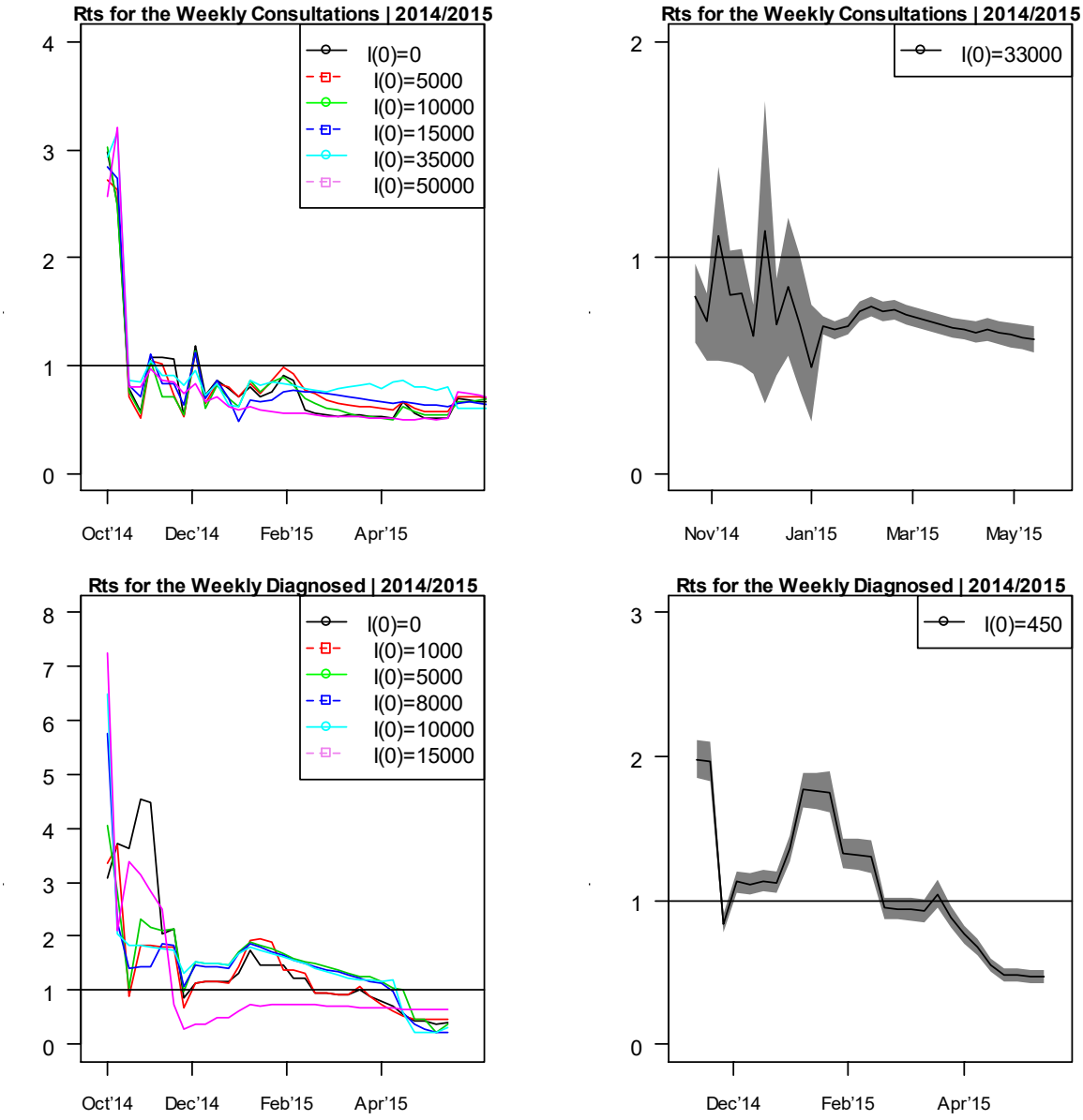


Figure 6.10– Sensitivity analysis for $I(0)$ and $E(0)$ in relation to the reproduction number for the consultations and diagnosed 2014/2015 datasets. Charts on the right represent the R_t plot when $I(0)$ and $E(0)$ are equal to the number of confirmed cases at the start of the outbreak, without the first two R_t points for the consultations data and without the first five R_t points of the diagnosed data. After these initial points, R_t values tend to stabilize for all different $I(0)$ and $E(0)$ values. Furthermore, the final R_t plots include a shaded area representing the 95% confidence interval.

The charts on the right hand side (Figures 6.6 - 6.10) include the 95% confidence interval for all chosen final R_t plots. Similar to the previous section, the final diagnosed plots have smaller variance in the confidence interval when compared to the consultations data.

6.4 Discussion

Throughout this chapter, we showed that the more the values of $R(0)$ are increased, the lower the resultant initial R_t values (for both consultations and diagnosed), while in general for the diagnosed data, the higher the values of $I(0)$ (and $E(0)$), the higher the initial R_t values. For the consultations data, the higher value of $I(0)$ does not influence the initial R_t values. In most of the above cases, between two to four initial R_t values were enough to remove most of the inconsistencies and unrealistic initial R_t values from all the different reproduction ratio charts for different $R(0)$s and $I(0)$s. Only for the 2014/2015 diagnosed ILI data we required a removal of the first five data points to obtain consistent values amongst the different R_t plots.

Stability was also assessed on the choice of the best value of the initial $S(0)$, $E(0)$, $I(0)$ and $R(0)$. In fact, for the above seasonal influenza datasets we were able to draw some conclusions about these priors. For $R(0)$, in general the most reasonable initial value is 150,000 (based on Malta's population size), since this is the maximum initial value of $R(0)$ where the R_t values across the whole outbreak period still tend to remain stable. For higher initial values of $R(0)$, the R_t values tend to be less consistent at different time points throughout the outbreak. There was only one exception for the 2013/2014 consultations dataset; for $R(0)$ equal to 100,000, the effective reproduction ratio provided some values greater than 1 during the peak of the outbreak. Hence, this was seen as more realistic for this dataset due to an accurate representation of the peak of the influenza outbreak.

For the prior mean value of $I(0)$ for the diagnosed dataset, we showed that the most reasonable choice is the number of observed diagnosed ILI cases at the start of the outbreak. The R_t values are more consistent for lower values of $I(0)$, for which it is consistent with the initial number of reported cases. In fact, for $I(0)$=15,000 (highest value considered for the diagnosed dataset), the plot of the effective reproduction ratio is inconsistent when compared to other $I(0)$ values. For the consultation datasets, R_t plots are more consistent for different values of $I(0)$ when compared to the diagnosed datasets. However, for two consultation datasets (2011/2012 and 2012/2013), lower values of $I(0)$ provided some R_ts>1, thus providing a signal which is associated with the peak of the influenza. Furthermore, these two datasets (2011/2012 and 2012/2013) are associated

with the highest number of consultation cases when compared between the four seasonal influenza datasets.

The 95% confidence interval of the final selected R_t plots provide further insight on the effective reproduction ratio. For the consultations data, the 95% confidence interval is substantially wider when compared to the diagnosed data. This further shows (as discussed in previous chapters) that the effective reproduction ratio of the diagnosed dataset is a clearer signal of the influenza outbreak. The consultation datasets include more uncertainties which merit further investigation in future research.

As published in several research papers [26, 65, 70, 150], the initial values of R_t are substantially high, similar to the initial R_t values in chapter 3 [79]. In these research papers, the reproduction ratio stabilizes to a value significantly lower than the initial value (as seen in this chapter). Other research papers only consider the median value and the 95% confidence interval for the effective reproduction ratio [80]. In fact, in a systematic review of the literature regarding the estimates of the reproduction number of seasonal, pandemic, and zoonotic influenza [80], I found that the median reproduction values were usually presented. Hence, the outliers were automatically ignored.

In a systematic review published by Biggerstaff et al. (2014) [80], the authors found that in 57 research papers related to the 2009 pandemic season, the median of the effective reproduction ratio was 1.46, while in another 20 research studies, the median effective reproduction ratio for 47 seasonal epidemics was 1.28. These values are further related to the effective reproduction ratio of the diagnosed datasets found in this thesis, as for the consultation datasets most of the R_t values are below one. Although other researchers analysed the initial phases of the seasonal influenza, they focused on the mean reproduction value for the initial period [151-152]. Hence, substantial research work ignores the initial value of the reproduction ratio and focuses more on the characteristics of epidemic. Essentially, the initial points of the reproduction values, as stated above, are not the true picture of the 'real' epidemic outbreak. This chapter provided a more holistic understanding of all the R_t values throughout the epidemic by examining the time series of the effective reproduction ratio for different initial mean prior values.

The above analyses suggests the importance of adopting a methodology when choosing the initial values of $R(0)$, $I(0)$, $E(0)$ and $S(0)$, especially since this has a direct impact on the most important epidemiological parameter, that is the reproduction number. The above method of analysis for the mean prior values of $R(0)$, $I(0)$, $E(0)$ and $S(0)$ were carried out when all the available data for each individual influenza season was observed. Hence, in future estimation of these values for an unfolding epidemic, one can either use the prior mean values of previous seasons for the Bayesian modelling or apply the model on the current available data. Hence, my research work is suggesting the following method when applying the sensitivity analysis for $S(0)$, $E(0)$, $I(0)$ and $R(0)$ as follows:

1. Apply the particle filter algorithm together with the SEIR model to fit the outbreak data for different values of $R(0)$. The prior mean values of $R(0)$ need to be selected in a realistic approach, for example, by selecting them according to the population size of the country under consideration. In my study I chose the zero cases for the low limit of $R(0)$, and $R(0)$ close to the population size as the upper limit. For different (and increasing) values of $R(0)$, one needs to monitor the R_t values and check whether they deviate substantially from the R_t values for different $R(0)$ values.

2. Plot all the respective R_t datasets for each individual model fit (for different $R(0)$ values) and remove the inconsistent R_t point/s sequentially from the left. These initial values are rather unrealistic and in general do not provide an accurate signal of the outbreak if the data corresponds to the initial cases of the influenza outbreak. These points represent inconsistencies due to the choice of the (unknown) $R(0)$.

3. Analyse all R_t plots for different $R(0)$ values and determine to which $R(0)$ value the R_t plot remains consistent. Hence, either:
    a. select the $R(0)$ value which corresponds to an R_t plot that has a peak related to the influenza season. If no peaks are observed,
    b. The value of $R(0)$ from which the R_t plot starts to deviate substantially from the previous R_t plots yields the ‘best’ estimate. Low values of $R(0)$ can overestimate the outbreak.

4. As regards to the mean value of $I(0)$, follow the same procedure as defined for $R(0)$. For the consultations datasets, select the $I(0)$ value which corresponds to an R_t plot that has a peak related to the influenza season. If no peaks are observed in the R_t plot (for the consultations data), then assume $I(0)$ equal to the number of observed consultation cases at the start of the outbreak, provided that the selected R_t plot is consistent with other R_t plots (for different $I(0)$ values). In general, for the diagnosed ILI dataset, the most realistic value of $I(0)$ represents the number of observed cases at the start of the outbreak. However, the same method as above (for the sensitivity of $R(0)$) needs to be applied to observe any inconsistencies in the effective reproduction ratios.

5. Assume $E(0)$ equal to $I(0)$. The number of exposed individuals and the number of infectious individuals can be assumed as equal. An infected person will eventually become infectious under normal circumstances.

6. The value of $S(0)$ follows from the values of $E(0)$, $I(0)$ and $R(0)$. $S(0)$ is equal to the population size under study without $E(0)$, $I(0)$ and $R(0)$ ($N - E(0) - I(0) - R(0)$).

The above method is a proposal based on the above datasets which warrants further testing for other populations. However, the above application of the sensitivity analysis on such prior mean values is a logical way to ensure that initial prior values are being discussed appropriately. However, if additional information is available on the outbreak that is directly related to $S(0)$, $E(0)$, $I(0)$ and $R(0)$, then these need to be tested and considered accordingly. For example improved information on the above priors can be found in serological studies and cross-sectional surveys related to any outbreak. As discussed in the literature review, several countries make use of such surveys to monitor the influenza disease progression [109, 112]. Other researchers use survey data to model the influenza through this information [118]. Throughout the next chapter, I will be examining the level of information that can be acquired from cross-sectional surveys.

Chapter 7

Probing into seasonal influenza: Exploring underlying factors

7.1 Introduction

In chapter 4, we showed that a substantial part of the consultations data is related to sub-clinical cases. This group of individuals carry a certain level of uncertainty since their illness is not clearly defined. In fact, according to the GPs data, this group might vary between 200 and 14,000 cases per week for the seasonal influenza datasets, and between 1,000 and 47,000 cases per week for the pandemic dataset. Furthermore, other individuals might opt for self-diagnosis, resulting in further uncertainty regarding the true number of infected individuals due to the seasonal influenza. It is believed that a significant proportion of the population do not visit their GP to be examined [79, 153-155] for their symptoms. These uncertainties all form part of the under-reporting rate in epidemiological studies, thus implying that there is limited information of the outbreak. Hence, in this chapter we aim to gain further underlying information about the influenza outbreak, rather than relying only on the 'standard' GP reporting data. The following cross-sectional survey aims to expand the analysis about the 'missing data' problem by acquiring further information on the true number of influenza cases within the Maltese population.

Furthermore, throughout this research I will be probing in detail on several important factors related to the seasonal influenza. In fact, information related to the symptoms of the seasonal influenza, the number of GP consultations throughout the year, hospitalisations due to the influenza, and medical information will all be examined throughout this chapter. Ultimately, such information aims to improve our knowledge of the influenza outbreak, to set better health strategies and to plan the appropriate interventions according to the needs of the population.

Throughout the following sections I shall analyse the survey related to the 2014/2015 influenza season. Subsequently, the results of a new survey (2015/2016) will be compared with the primary data (2014/2015 survey) of this chapter. For the first survey (Survey 1: 2014/2015) questions are related to the period August 2014 till July 2015 (Appendix B). For the second survey (Survey 2: 2015/2015) some general questions (such as GP consultations visit and regular medication) are related to the period May 2015 till April 2016 and influenza related questions are associated to time period August

2015 till April 2016. Note that further information about the methodology of this survey is found in chapter 2.

7.2 Ethical considerations

In order to carry out a cross-sectional survey in Malta about the above objectives, an application was submitted to obtain ethics approval (Appendix F) from the Psychology Ethics Committee, University of Stirling. My application was under the project title name: "Understanding the under-reporting of the Seasonal Influenza". The study was approved by the same committee on the 28th of August 2015.

Following an explanation of the main purpose of this research to the participants, individuals were invited to participate in the study through a telephone survey. Participants were given the option to opt out from this research study at any time during the 5-minute telephone survey. Furthermore, respondents were also assured that all the collected information would be processed anonymously and confidentially. Further information on the telephone interview introduction can be found in appendix F.

During this research survey, I administered the entire process thoroughly to ensure that the survey is in accordance with the above ethics application. For most questions, individuals were requested to answer to the questions retrospectively.

7.3 Representativeness of the sample

In order to ensure representativeness of the population, the sample (n=406) was stratified based on the demographics, gender, district and age. In fact, these demographics are fairly homogenous when compared with the study's population (Tables 7.1 – 7.3).

Malta's population is evenly distributed between females and males and this is reflected in table 7.1. From the Maltese population, 18.40% are 66 years of age or older, followed by those between 26 and 35 years of age (18.30%), those between 46 and 55 (17.80%), and individuals between 56 and 65 years of age (17.80%) (Table 7.2). Malta has six different districts (as defined by National Statistics Office, NSO) [16], which are defined in table 7.3. The most populated district is the Northern Harbour district (29.48%), followed by the Southern Harbour district (18.89%). For all the three demographical

variables, differences between the actual data and the sample data do not exceed the actual confidence interval of this study (±4.86%), which makes it representative based on the most important population's demographics.

Gender	Population	Sample	Difference
Female	50.50%	50.99%	0.49%
Male	49.50%	49.01%	-0.49%

Table 7.1 – Comparison of the population's gender against the sampled collected data. Percentages are very close, hence the sample is representative according to gender. Population data was retrieved from NSO's Demographic Review 2013 [16].

Age	Population	Sample	Difference
18-25	12.45%	13.05%	0.60%
26-35	18.30%	14.29%	-4.01%
36-45	15.19%	15.52%	0.33%
46-55	17.80%	17.49%	-0.31%
56-65	17.80%	22.41%	4.61%
66+	18.40%	17.24%	-1.16%

Table 7.2 – Comparison of the population's age, against the sampled collected data. Percentages are very close, hence the sample is representative according to age. Population data was retrieved from NSO's Demographic Review 2013 [16].

Districts	National Statistics Office (Actual Population)	Sample	Difference
Southern Harbour	18.89%	21.18%	-2.29%
Northern Harbour	29.48%	27.09%	2.39%
Southern Eastern	15.31%	19.70%	-4.40%
Western	13.80%	11.58%	2.22%
Northern	15.11%	16.01%	-0.90%
Gozo & Comino	7.41%	4.43%	2.98%

Table 7.3 – Comparison of the population's regions, against the sample collected. Percentages are very close, hence the sample is representative according to region as well. Population data was retrieved from NSO's Demographic Review 2013 [16].

7.4 Sample characteristics

All tables related to this section are found in Appendix G. Among participants that took part in this study (*n*=406), 70.4% are married. Furthermore, the majority (46.5%) of respondents are employees, followed by pensioners (21.8%), housewives (21.5%) and

students (7.5%). For the question regarding the educational level, 54.6% of the participants reached secondary level as their highest level of education, while 18.7% only reached a primary level of education, 13.5% reached diploma level and 13.2% reached tertiary level (Degree).

Our data show that on average there are 2.9 individuals inhabiting every Maltese household. The majority of the houses (33.3%) have 3 individuals living in one house, followed by those with 2 individuals (24.7%) and those with 4 individuals (24.7%). From the total sample, 11.1% of all participants live on their own, while 6.1% of all participants live in a household of 5 individuals or more. The top preferred means of transport in Malta is the individual's private car such that 64.2% of all the participants use their car as their main means of transport. This is followed by 21.5% of individuals who use public transport.

7.5 Results

7.5.1 Participants' general medical information

On average the participants visit their general practitioner (GP) 2.7 times in one year. The majority visit their GP twice a year (26.4%), followed by once a year (18.5%) and three times a year (16.3%). Following this, the number of visits per individual decreases (Figure 7.1).

In general, 41.2% of the participants take regular medication due to medical conditions such as asthma, diabetes, heart disorders or other. Predominantly for the older age group (66+, 89.9%), the proportion is significantly higher when compared to the younger generation (χ^2 (5) = 121.11, p-value < 0.01). For those between the age of 18 and 25 years, 17.0% take regular medication, and for those between 26 and 35 years, 8.6% take regular medication. Furthermore, results exceed the 50% threshold for the age group 56-65 (53.8%).

According to the same survey, one of every four Maltese citizens (25.9%) smoke on a regular basis. Males smoke significantly more than females (35.7% vs. 16.3%) (χ^2 (1) = 19.52, p-value < 0.01). Furthermore, on average Maltese smokers smoke 16.2 cigarettes per day. The majority (42.2%) smoke 20 cigarettes per day.

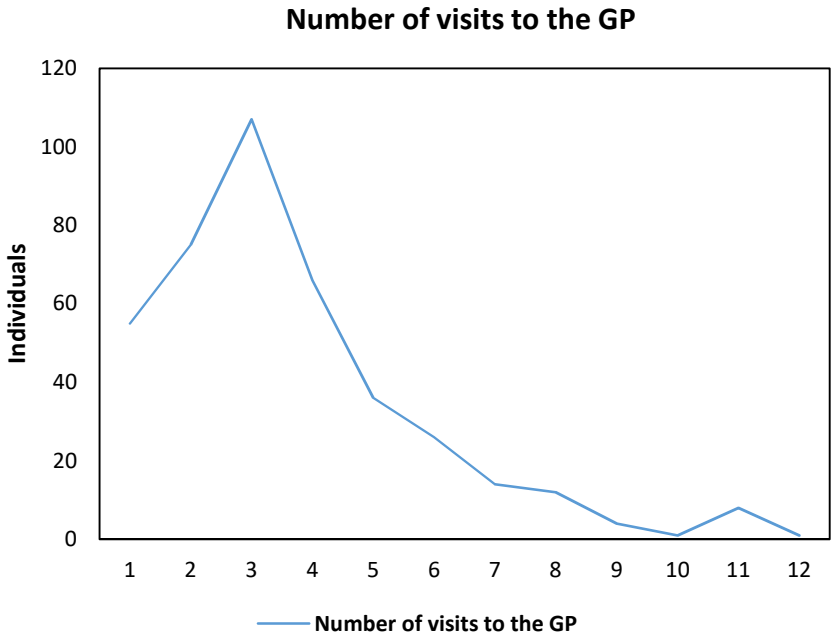


Figure 7.1 – The number of times individuals visit their GP. The most common number of visits per year is three visits.

7.5.2 The seasonal influenza vaccine

On a yearly basis, the Maltese Government offers the seasonal influenza vaccine free of charge to some groups of individuals as explained in chapter 2. However, others need to consult their private doctor to receive their influenza vaccination at a cost.

According to the survey results, during the 2014-2015 season, 43% reported that they had received the flu vaccine, while 55.3% had not taken the vaccine and 1.7% do not remember. Of those who received the vaccine flu, the only age group that exceeded the 50% uptake is the 66+ age group (73.9%). In relation to this, after applying a Chi-Squared test, it was found that there is a significant association between the different age groups when compared with the vaccine uptake (χ^2 (10) = 49.86, p-value < 0.01). This result is due to the above Government's inclusion criteria for the free vaccine. Furthermore, the latter result is similar to England's vaccine flu uptake rate for those aged 66+ [156]. Those between 18 and 25 years of age are the least age group who were compliant to take the flu vaccine (22.6%), while for those between 26 and 65 years the compliance to vaccination varied between 36% and 46% for the 2014/2015 influenza season. The European Council's recommendation is to reach the 75% among the higher risk groups of people [91].

The main reasons for those individuals that who did not take the vaccine flu were: 'not interested' (41.1% of individuals), followed by those who were afraid (24.1%) and 10.7% who said that they 'feel sick after taking the vaccine'.

7.5.3 Influenza-Like Illness (ILI)

In this research study, respondents were asked whether they had several symptoms from a whole list of ILI symptoms such as fever, cough, sore throat, headaches and others symptoms. Symptoms were mentioned to respondents one by one, and hence the respondents had to reply to every individual symptom. Respondents were asked to reply to this question retrospectively for the past year (August 2014 – July 2015).

The most common symptoms amongst the participants were 'runny or blocked nose' (61.6%), followed by a headache (60.6%), whilst the least common symptoms were vomiting (6.9%) and chest pain (11.6%). Figure 7.2 shows all the latter results in descending order. Furthermore, these symptoms were placed in three different groups; the most frequent (the left upper oval), the less frequent (the right lower oval) and the middle of the previous two groups (the middle oval). The middle group was mentioned by 323 individuals (79.6%) of the whole sample (n=406). The top two common symptoms were mentioned by 74.6% and the least common symptoms were mentioned by 48.3% of the survey respondents. According to my cross-sectional survey in Malta, 15.4% of the Maltese population did not have any of the above symptoms during the indicated one-year period. I interpret this number as individuals who can be considered as definite non-influenza individuals, however the 84.6% does not necessarily mean that they definitely had the seasonal influenza.

These results are in accordance to the findings of the 'UK Flu survey' (https://flusurvey.org.uk/en/results/) [109] which reports the most common symptoms as runny nose, cough, sneezing, headache, sore throat and feeling tired. However, the UK survey data is biased towards those individuals that have and use the internet and thus towards those with a higher level of education [109].

1	Runny or blocked nose	61.6%
2	Headache	60.6%
3	Sore throat	54.4%
4	Cough	50.0%
5	Sneezing	45.8%
6	Feeling tired or exhausted	41.1%
7	Muscle/joint pain	34.7%
8	Fever	28.6%
9	Loss of appetite	22.4%
10	Watery eyes	21.2%
11	Diarrhoea	18.0%
12	Shortness of breath	16.7%
13	Stomach ache	15.0%
14	Nausea	13.3%
15	Chest pain	11.6%
16	Vomiting	6.9%

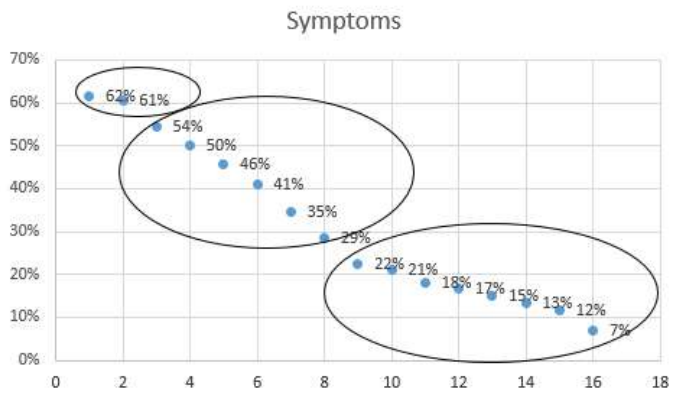


Figure 7.2 – This figure represents individual results for 16 different symptoms. The above results are sorted in descending order to elicit the most common symptoms amongst the participants from August 2014 till July 2015. Respondents were asked to reply for every symptom.

The most common month for the above symptoms was January 2015 (18.8%), followed by February 2015 (15.9%) and March 2015 (14.5%). The least popular months were August 2014 (0.3%), September 2014 (0.3%) and July 2015 (2.8%). These results are expected since the latter three months are the least common months for acquiring the influenza. Further details are given in the discussion section, where the above data will be compared with the ILI diagnosed cases as reported by the GPs for the 2014-2015 season (Chapter 4).

Months	Frequency	Result
Aug-14	2	0.3%
Sep-14	2	0.3%
Oct-14	38	5.8%
Nov-14	41	6.3%
Dec-14	70	10.7%
Jan-15	123	18.8%
Feb-15	104	15.9%
Mar-15	95	14.5%
Apr-15	70	10.7%
May-15	48	7.3%
Jun-15	43	6.6%
Jul-15	18	2.8%
Total	**654**	**100.0%**

Table 7.4 – This table represents the months that participants indicated as having any of the above symptoms. The top month for these symptoms was January 2015 (18.8%) and the least popular months were August 2014 (0.3%) and September 2014 (0.3%). Respondents were able to indicate more than one month for the occurrence of the symptoms.

On average, these symptoms persisted on the participants for 9.4 days. The most common duration for the above symptoms was 7 days (14.0%), followed by 3 days (13.1%), 4 days (11.6%), 2 days (9.4%) and 14 days (7.6%).

For the patients with the above symptoms, 56.5% claimed that they were restricted to stay at home, while 43.5% were not restricted to stay at home to recover from their ILI symptoms (Figure 7.3).

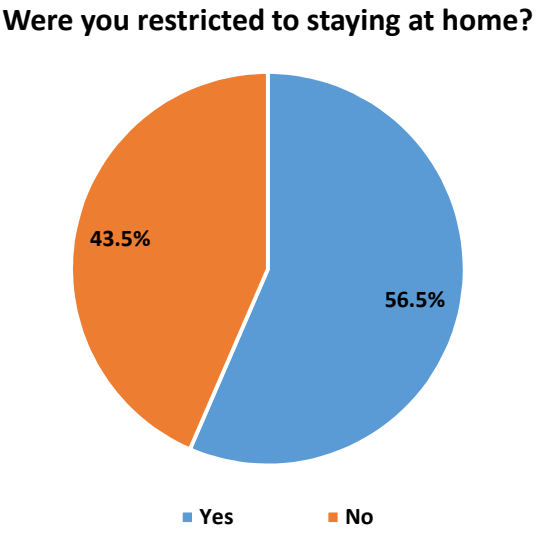


Figure 7.3 – Respondents were asked whether they were restricted to stay at home due to the above mentioned symptoms. 56.5% claimed 'Yes', while '43.5%' claimed 'No'.

7.5.4 Seasonal influenza 2014-2015

The previous section analysed symptoms related to the seasonal influenza (without mentioning the term 'seasonal influenza' to respondents). In this section we shall analyse items for which respondents were asked questions directly related to the term 'seasonal influenza'. As discussed in chapter 1, there exists a standard definition of seasonal influenza, however respondents were asked whether they had seasonal influenza during the past year, without actually giving them the standard definition. Hence, results are based either on their own judgement and understanding of seasonal influenza, and/or based on their GP's advice. Results from this survey showed that 29.8% of the individuals claimed that they had seasonal influenza. This contrasts significantly when compared with the 84.6% of individuals who claimed they had any of the above ILI symptoms. Furthermore, 67.0% claimed that they did not acquire seasonal influenza and 3.2% were unsure. The most common month (Table 7.5) for the seasonal influenza according to the respondents was January 2015 (28.4%), followed by February 2015 (23.0%) and

December 2014 (16.4%). The top two months are likewise the most common months for the ILI symptoms.

Respondents were able to mention more than one month for having the seasonal influenza. In fact out of those who claimed they had seasonal influenza (29.8%), the latter individuals indicated of having the influenza an average of 1.5 times during the year. There are several reasons for having seasonal influenza more than once. Primarily, people with a lower immune system might suffer from seasonal influenza more than once [84]. Secondly, during the seasonal influenza, individuals might suffer from influenza A (which is the common seasonal influenza) and influenza B[2]. Furthermore, since a significant proportion of respondents replied to the questionnaire based on their self-diagnosis, their ILI symptoms might have been adjudicated as another seasonal influenza. In reality, this might be incorrect or their understanding of influenza was in fact a common cold. Moreover, according to the Malta Health Promotion Department (MHPD), although there is a possibility that a person acquires other strains of the influenza virus, usually a person acquires the influenza once in a season, due to one of the viruses in circulation being most dominant (Appendix A).

Similarly, in reply to the question focusing on the duration of the above symptoms, individuals claimed that on average, the duration of the seasonal influenza was 9.9 days. The majority (42.5%) claimed that the seasonal influenza persisted for 7 days, followed by 14 days (18.6%) and 4 days (8.8%). This is similar to that stated by the World Health Organization (WHO) [82], which states that most people recover from the main seasonal influenza symptoms within one week. Furthermore, according to the Proprietary Association of Great Britain (PAGB) [158], complete recovery for seasonal influenza might take up to 10 days.

Respondents who claimed they had seasonal influenza during the past year were asked to identify any symptoms related to their seasonal influenza. Most of the respondents identified more than one symptom. On average, every respondent mentioned 5.4 symptoms. In total, the 121 respondents who claimed they had the seasonal influenza mentioned 658 symptoms (non-unique symptoms). Table 7.6 provides the percentages

[2] "Type B flu may cause a less severe reaction than type A flu virus, but occasionally, type B flu can still be extremely harmful. Influenza type B viruses are not classified by subtype and do not cause pandemics." [159]

based on the total number of symptoms mentioned (658). The most common symptom (Table 7.6) according to participants was cough (15.5%), followed by sore throat (14.1%), fever (12.2%), headache (10.6%), runny or blocked nose (10.6%) and sneezing (8.8%). The least popular mentioned symptoms were watery eyes (0.6%), vomiting (1.5%) and nausea (1.5%).

Months	Frequency	Result
Aug-14	0	0.0%
Sep-14	0	0.0%
Oct-14	15	8.2%
Nov-14	11	6.0%
Dec-14	30	16.4%
Jan-15	52	28.4%
Feb-15	42	23.0%
Mar-15	26	14.2%
Apr-15	4	2.2%
May-15	2	1.1%
Jun-15	1	0.5%
Jul-15	0	0.0%
Total	**183**	100.0%

Table 7.5 – The months indicated by participants for having the seasonal influenza. The top month for these symptoms was January 2015 (28.4%) and the lease popular month was June 2015 (0.5%). Respondents were able to indicate more than one month.

More specifically respondents were asked whether they had temperature. Out of the seasonal influenza individuals (including those who opted for the 'don't know' option), 64.2% claimed that they had temperature, 22.4% did not and 13.4% do not know. Furthermore, 68.7% visited a doctor due to their seasonal influenza, 18.7% did not and 12.7% they do not remember. Additional analysis showed that four out of every five seasonal influenza individuals took medicine to cure their influenza symptoms, while 13.4% did not remember. On the other hand, one in every five individuals were hospitalised due to the seasonal influenza. The hospitalised individuals spent an average of 6 nights at hospital. However, the majority (37.0%) spent 1 night, followed by those who spent 14 nights (22.2%) and 7 nights (11.1%).

The absolute majority of respondents (54.5%, Figure 7.4) claimed that at least one member from their household had acquired the seasonal influenza (excluding the respondent). However, when also taking into account the 29.8% from the total sample who claimed they had the seasonal influenza, 61.1% of all Maltese households had at

least one person with seasonal influenza. Furthermore, from those participants who had influenza cases amongst their household members, on average, 1.5 household members had the influenza (excluding themselves). On the other hand, on average there were 1.8 household members in Malta who had acquired the seasonal influenza, after taking into account the respondent's reply regarding their seasonal influenza.

Symptoms	Frequency	Result
Cough	102	15.5%
Sore throat	93	14.1%
Fever	80	12.2%
Headache	70	10.6%
Runny or blocked nose	70	10.6%
Sneezing	58	8.8%
Muscle/joint pain	35	5.3%
Feeling tired or exhausted	31	4.7%
Stomach ache	22	3.3%
Diarrhoea	20	3.0%
Loss of appetite	18	2.7%
Shortness of breath	18	2.7%
Chest pain	17	2.6%
Nausea	10	1.5%
Vomiting	10	1.5%
Watery eyes	4	0.6%
Total	**658**	**100.0%**

Table 7.6 – All the mentioned symptoms for seasonal influenza by Maltese participants. On average respondents mentioned 5.4 symptoms. In total, the respondents mentioned 658 symptoms, from 121 respondents who claimed they had seasonal influenza during the previous year. Hence, the percentages were calculated from the total number of mentioned symptoms by all respondents (n = 658). Unlike figure 7.2, respondents were not requested to reply for every symptom.

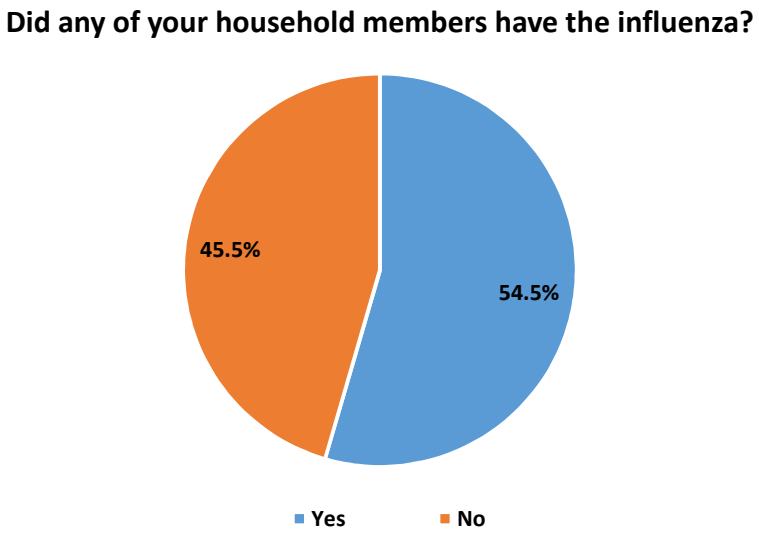


Figure 7.4 – The proportion of the number of respondents who claimed they had seasonal influenza patients within their household.

7.5.5 Seasonal influenza 2015-2016

Another survey was carried out for the 2015-2016 seasonal influenza period. The survey methodology was identical to that carried out in the previous survey, including accurate representativeness based on gender, age and district. However, this was carried out during the end stages of the seasonal influenza (April 2016), while the first survey was carried out three months after the seasonal influenza had ended (end of August 2015 and beginning of September 2015). Hence, the main scope of the 2016 survey was to test whether the information obtained from the 2015 survey had improved, when carrying out the survey at an earlier timeframe when compared to the first survey. Furthermore, I shall be comparing the results between both surveys to understand the consistencies and inconsistencies between different years.

7.5.5.1 Results of the 2015-2016 survey

On average, the participants visited their doctor 2.7 times during the past year. This is in full agreement with the first survey (2.7 times). According to the 2015/2016 survey, amongst the Maltese population, 38.2% take regular medication (41.2%, Survey 1) and 21.2% from the total sample smoke cigarettes (25.9%, Survey 1). On average, smokers smoke 13.7 cigarettes per day (16.2, Survey 1).

Respondents were given a list of symptoms for which they were asked to reply to every single one of them. The most common symptoms are presented in table 7.7. These include

runny or blocked nose (58.4%), headache (55.9%), sneezing (54.7%), sore throat (50.5%), cough (48.8%), muscle/joint pain (34.5%), feeling tired or exhausted (30.8%), watery eyes (26.1%), fever (25.4%), diarrhoea (16.0%), shortness of breath (15.5%), chest pain (12.1%), stomach ache (12.1%), loss of appetite (11.6%), nausea (9.1%) and vomiting (3.9%). These percentages are comparable to those in figure 7.2 for survey 1. The only differences are for sneezing (8.9% less, Survey 1), feeling tired or exhausted (10.3% more, Survey 1) and loss of appetite (10.8% more, Survey 1). These last two symptoms can be easily associated with the summer period, given that the first survey was carried out during some of the warmest days in Malta during the year. According to the 2015/2016 survey, 20% of the survey respondents did not have any of the above symptoms (15%, Survey 1). From those respondents who had at least one symptom, 54.9% were restricted to stay at home to recover (57%, Survey 1). On average, these symptoms persisted for 5.9 days (9.4 days, Survey 1).

Individual Symptoms	Frequency	Result
Runny or blocked nose	237	58.4%
Headache	227	55.9%
Sneezing	222	54.7%
Sore throat	205	50.5%
Cough	198	48.8%
Muscle/joint pain	140	34.5%
Feeling tired or exhausted	125	30.8%
Watery eyes	106	26.1%
Fever	103	25.4%
Diarrhoea	65	16.0%
Shortness of breath	63	15.5%
Chest pain	49	12.1%
Stomach ache	49	12.1%
Loss of appetite	47	11.6%
Nausea	37	9.1%
Vomiting	16	3.9%

Table 7.7 – Individual results for 16 different symptoms. The above results are sorted in descending order to elicit the most common symptoms amongst the participants for the 2015/2016 influenza season. Respondents were asked to reply for every symptom.

Respondents were then asked whether they had acquired seasonal influenza during the 2015/2016 influenza season. According to the survey data, 37.2% of the respondents (29.8%, Survey 1) claimed of having the seasonal influenza during the 2015-2016 period (until April 2016 which was the month of data collection). Furthermore, respondents

claimed of having the seasonal influenza 1.28 times during the same season (1.5 times, Survey 1). Respondents claimed that on average the duration of the seasonal influenza was 9.5 days (9.9 days, Survey 1). Most respondents claimed they had more than one symptom related to their seasonal influenza (Table 7.8). In total, respondents mentioned 1080 symptoms (non-unique symptoms) (Table 7.8). Hence, the percentages were calculated from this total (1080) in contrast to the individual symptoms analysed in table 7.7. From all symptoms mentioned by the respondents (Table 7.8), the most common symptoms are: sneezing (13.2%) followed by cough (13.2%), sore throat (13.1%), runny or blocked nose (12.8%), muscle/joint pain (6.4%), headache (9.4%) and fever (7.8%).

Furthermore, 55.0% of the respondents claimed that they had temperature (64.2%, Survey 1), 72.9% visited their GP due to their seasonal influenza (68.7%, Survey 1), 97.4% took medicine to cure from their influenza (80%, Survey 1) and 4.0% were hospitalized due their seasonal influenza (20.0%, Survey 1). The months associated with the above symptoms and the seasonal influenza will be analysed and compared throughout the discussion section.

What were the symptoms?	Survey 1	Survey 2	Difference
Sneezing	8.8%	13.2%	4.4%
Cough	15.5%	13.1%	-2.4%
Sore throat	14.1%	13.1%	-1.1%
Runny or blocked nose	10.6%	12.8%	2.1%
Muscle/joint pain	5.3%	11.7%	6.3%
Headache	10.6%	9.4%	-1.3%
Fever	12.2%	7.8%	-4.4%
Feeling tired or exhausted	4.7%	4.3%	-0.5%
Watery eyes	0.6%	4.2%	3.6%
Diarrhoea	3.0%	2.7%	-0.4%
Chest pain	2.6%	2.4%	-0.2%
Shortness of breath	2.7%	1.9%	-0.8%
Loss of appetite	2.7%	1.8%	-1.0%
Nausea	1.5%	1.1%	-0.4%
Stomach ache	3.3%	0.5%	-2.9%
Vomiting	1.5%	0.2%	-1.3%
Total	100.0%	100.0%	0.0%

Table 7.8 – A comparison (Survey 1 vs. Survey 2) between the symptoms related to the seasonal influenza as mentioned by the survey respondents. Responses between both surveys are similar within the ±4.87% margin of error, with the exception of the 'Muscle/joint pain' symptom. Since respondents were allowed to mention more than one symptom, in total the above symptoms were mentioned 1080 times, hence the percentages were calculated from this total. Unlike table 7.7, respondents were not requested to reply for every symptom.

From all respondents, only 24.9% claimed that have had one or more of their household members with the seasonal influenza (54.5%, Survey 1). However, after taking into account the respondents' replies for the seasonal influenza question, 43.8% of all Maltese households had at least one person who had suffered from the influenza (61.1%, Survey 1). On average, according to the survey data, 1.7 household members in Malta had the seasonal influenza (1.8 members, Survey 1).

7.6 Discussion

7.6.1 Validating the GPs data

Data presented in the results section provide important information for the scope of this dissertation. However, one can analyse such results from different perspectives, such as analysing the characteristics of those who were diagnosed with the seasonal influenza or constructing several scientific models to predict key variables. Factors such as age, gender, education status and district provide improved information to health promotion authorities to better plan their health promotion campaigns. However, it is beyond the scope of this dissertation to analyse such information, as we are more interested in the actual prediction of the seasonal influenza outbreak.

Data presented in this survey shed more light on different characteristics of the seasonal influenza. We showed different characteristics related to the health of individuals, symptoms related to the seasonal influenza, perceptions related to the seasonal influenza, the months in which respondents claimed of having several ILI symptoms and also the months when they acquired the seasonal influenza. The latter two variables can be directly compared with the GPs data (Chapter 4). Hence, throughout the next paragraphs we will be comparing:

1. The monthly occurrences of ILI symptoms (number of ILI symptomatic cases per month) as stated by the survey respondents against the monthly diagnosed ILI cases from the GPs reported data (as defined in Chapter 4). The survey question related to this analysis was, 'When did your symptoms appear for the above during the past year?' (Appendix B). 'Above' in this question corresponds to the list of symptoms as defined in figure 7.2.
2. The monthly occurrences of seasonal influenza as stated by the survey respondents against the monthly diagnosed ILI data from the GPs reported

data (as defined in Chapter 4). The survey question related to this analysis was, 'If 'Yes', when did you have the seasonal influenza?' (Appendix B). The 'Yes' reply corresponds to the respondents who claimed that they had experienced the seasonal influenza during a one year period.

Furthermore, the above two comparisons will be examined for both surveys. Hence, throughout this section, we also aim to compare the results of the first survey with the results of the second survey. For both comparisons, I will use the diagnosed data (GP data) which corresponds to the same year of the survey.

There is a good agreement between the monthly occurrences of the Influenza-like Illness (ILI) symptomatic cases (as stated by the survey respondents) (Figure 7.5) and the 2014/2015 diagnosed ILI cases (GPs reported data). This agreement can be explained by a strong linear correlation ($r = 0.90$) (Table 7.9). Such a strong correlation between the two variables was also found to be significant (p-value = 0.002, Table 7.9). This result validates the data collected by the GPs as a reliable source of information to model the seasonal influenza. Although the survey data was collected retrospectively, the respondents still remembered the actual months when they had the above symptoms. Nevertheless, the last three months of the survey data registered a higher number of cases when compared with the observed diagnosed ILI data (Figure 7.5).

Similarly as above, figure 7.6 compares the monthly occurrences of the seasonal influenza cases as stated by the survey respondents, against the 2014/2015 GPs diagnosed seasonal influenza cases. In the latter case only respondents who claimed of having seasonal influenza are analysed. While for the previous analysis all the respondents who claimed they had at least one ILI related symptom were analysed. The time dependence graphs are very close and the linear correlation between both variables can be summarized through a Pearson-correlation coefficient of 0.88 (Table 7.9). Such a strong correlation between the two variables was found to be significant (p-value = 0.004, Table 7.9).

It is interesting to note that figure 7.6 shows a lower number of seasonal influenza occurrences (survey data) when compared with the diagnosed ILI cases (GP data) for the late part of the influenza season. On the contrary, the occurrence of the symptoms (survey 1 data) overestimates the late part of the influenza season (Figure 7.5), when compared

with the diagnosed ILI cases (GP data). Furthermore, the peak number of cases for figure 7.5 for the survey data is 123, while the peak number of cases of figure 7.6 for the survey data is 52. Hence, for this peak value only 42% of the total symptomatic occurrences are seasonal influenza cases, according to the survey respondents.

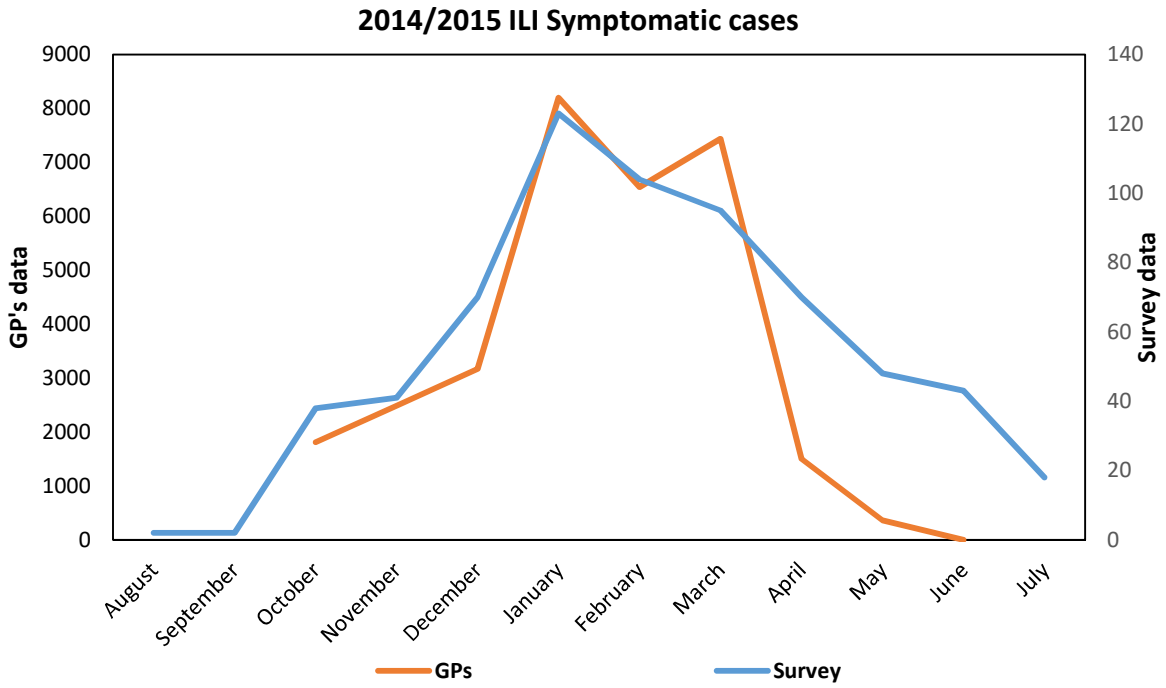


Figure 7.5 – Comparison of the monthly occurrences of the Influenza-like Illness (ILI) symptomatic cases (blue line) as stated by the survey respondents against the 2014/2015 GP diagnosed ILI cases (orange line) (Chapter 4). The y-axis represent the number of cases for both variables.

When comparing the monthly occurrences of the symptomatic cases as stated by the survey respondents (Figure 7.5) against the monthly occurrences of the seasonal influenza cases (Figure 7.6) as stated again by the survey respondents, this gives a Pearson-correlation value of 0.85, which is a strong correlation between the two survey variables. Hence, this means that the occurrence of the months for the above symptoms that were mentioned individually throughout the survey, are linearly associated with the same months that respondents claimed to have the seasonal influenza. Such a strong relationship was found to be significant (p-value = 0.008, Table 7.9). However, only around 30% of participants claimed they had the seasonal influenza, while around 84.6% claimed that they had any of the above ILI symptoms. Hence, based on these results, it is likely that respondents have a different perception of the definition of the seasonal influenza. Furthermore, it is important to keep in mind that a substantial proportion of the population opts for self-diagnosis to examine their ILI symptoms. Hence, illness

perceptions and health beliefs are rather subjective, although these are important predictors for health utilization [161-163].

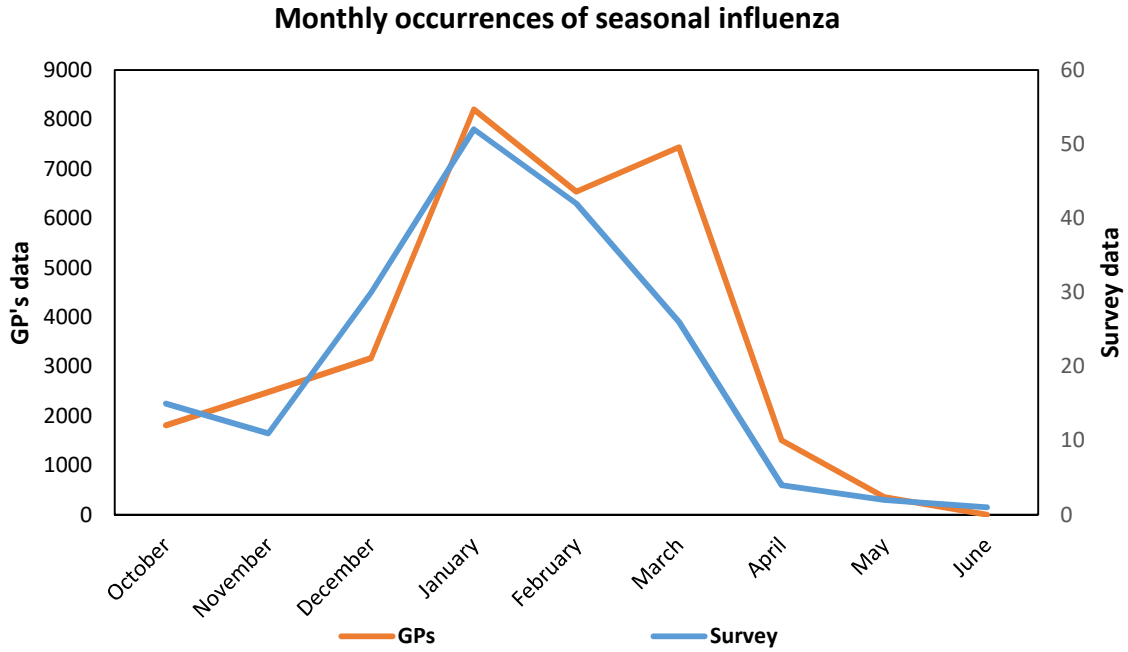


Figure 7.6 – Comparison of the monthly occurrences of the seasonal influenza cases as stated by the survey respondents (blue line) against the 2014/2015 GP diagnosed ILI cases (orange line) (Chapter 4). The y-axis represent the number of cases for both variables.

Correlations

		GPs_Influenza	Survey_Symptoms	Survey_Influenza
GPs_Influenza	Pearson Correlation	1	.899**	.882**
	Sig. (2-tailed)		.002	.004
	N	8	8	8
Survey_Symptoms	Pearson Correlation	.899**	1	.846**
	Sig. (2-tailed)	.002		.008
	N	8	8	8
Survey_Influenza	Pearson Correlation	.882**	.846**	1
	Sig. (2-tailed)	.004	.008	
	N	8	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Table 7.9 – Correlation analysis for the three variables related to the months of the influenza symptoms. 'GPs_Influenza' is the diagnosed seasonal influenza individuals collected by the GPs, while 'Survey_Symptoms' variable is the monthly occurrences of the Influenza-like Illness (ILI) symptomatic cases as stated by the survey respondents and 'Survey_Influenza' variable is the monthly occurrences of the seasonal influenza cases as stated by the survey respondents.

Similar results were obtained for the 2015/2016 survey (Figure 7.7). Hence, for the second time, the months being mentioned by the survey respondents are similar to the months that were recorded by the GPs for their reported diagnosed ILI cases. Both surveys recorded accurate results, even though they were carried out during different

timeframes (one survey carried out a few months after the end of the 2014/2015 seasonal influenza, and the second survey carried out during the end stages of the seasonal influenza). Similarly as above (2014/2015 survey), the number of symptomatic cases (2015/2016 survey) is higher for the late period of the seasonal influenza when compared with the diagnosed ILI cases (GP data). In contrast, the number of seasonal influenza occurrences (2015/2016 survey) is lower than the diagnosed ILI cases (GP data) for the late period of the seasonal influenza (Figure 7.7).

In order to further test the reliability of my cross-sectional survey, I shall analyse information related to the consultations data. According to the data obtained in the first survey, respondents visited their doctor around 2.7 times during the year for any type of consultation. Hence, based on the survey data, during August 2014 – July 2015, Maltese residents visited their doctor around 1.1 million times (after generalizing it to the whole Maltese population). GP consultations data only included the period of October 2014 to mid-May 2015 (total of around 835,000 cases, Chapter 4), while the survey is considering data for one whole year. However, the months where the data was not collected by the GPs (mid-May 2015 until the end of September 2015) are not synonymous with the seasonal influenza [84]. In fact it is not expected to have cases of seasonal influenza during the summer period [84]. Hence, one might consider the baseline number of non-influenza consultations (as described in Chapter 4) as the best estimate for those months where data was not collected by GPs. After extrapolating this data to the remaining months, this makes the number of consultations in Malta around 1 million over a period of one year, based on the GPs data. This latter estimate is not far from the estimated number of consultations that was obtained from the survey (1.1 million).

Data related to the general medical conditions of the individuals provided similar results between both surveys. In fact, variables related to the number of doctors' consultations, regular medication and number of smokers all provided consistent results between both surveys. Most common symptoms are also consistent. However, while in the first survey around 30% claimed of having the seasonal influenza, in the second survey around 37% claimed of having the seasonal influenza. In the second survey, every individual claimed of having experienced the seasonal influenza for an average of 1.28 times during the same season, while in the first survey an average of 1.5 times were recorded per individual. Furthermore, according to the second survey data, a lower number of Maltese households

registered seasonal influenza cases. Hence, these results provide further understanding that the seasonal influenza has different infection rates year-on-year. We already showed (in Chapter 4) that the number of infected individuals vary on a yearly basis and this may be attributed to the climate conditions [97].

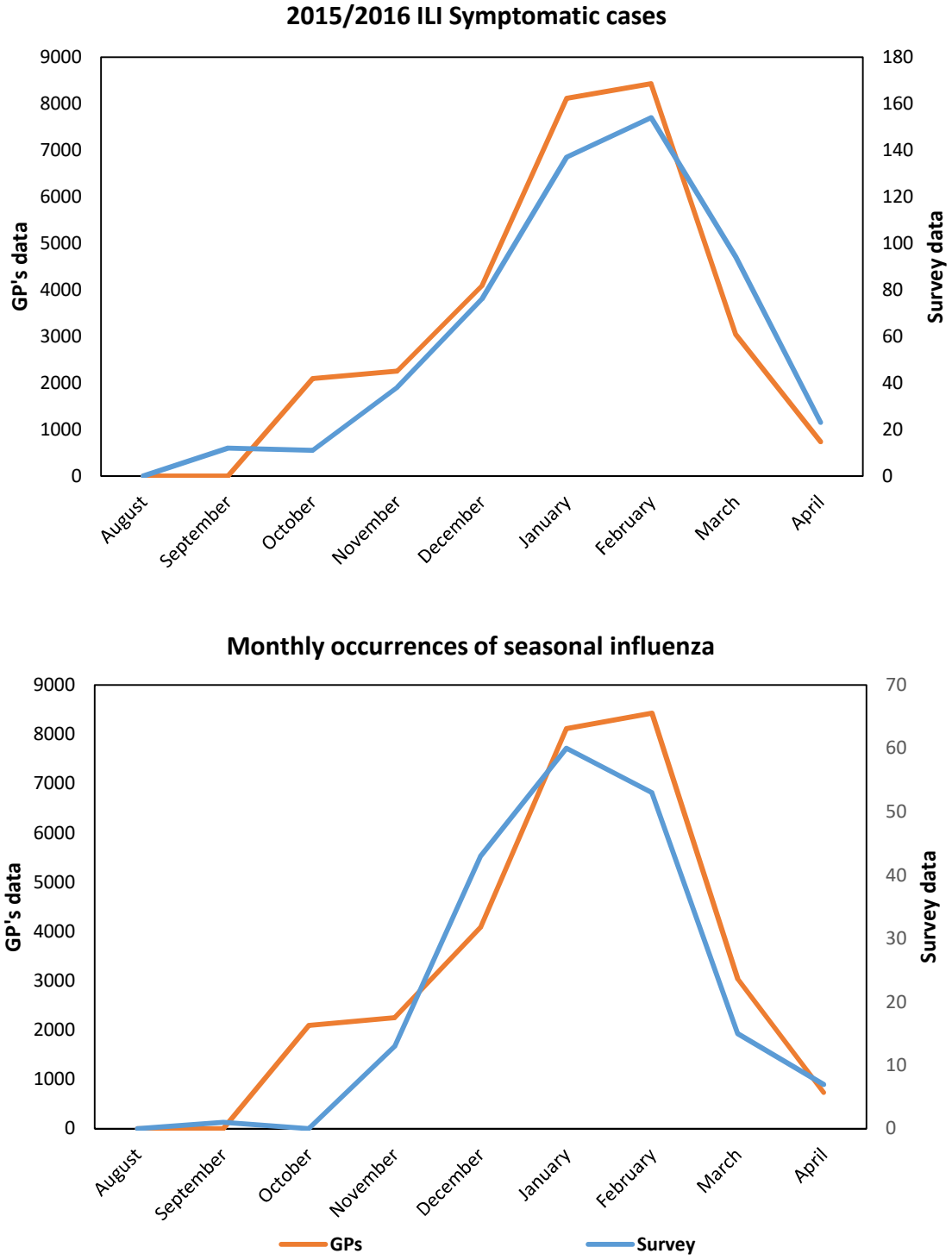


Figure 7.7 – Comparison of the monthly occurrences of the ILI symptomatic cases (upper chart) or seasonal influenza cases (lower chart) as stated by the survey respondents (blue line) against the GPs diagnosed ILI cases (orange line). The y-axis represent the number of cases for both variables.

The climate during 2015/2016 period in Malta varied substantially. Hence, this might have an influence on the seasonal influenza. Winter of 2015-2016 was registered as the driest winter on record in Malta [160]. During this period, rain was scarce and air temperatures were higher than usual. The temperature registered in February (23.6°C) was the highest recorded temperature in Malta for the past 93 years. When compared to the 2014-2015 season, the 2015/2016 winter was almost four times drier [160]. These phenomena were also experienced in Europe and United States. As discussed in the literature review, climate conditions may have a direct influence on the spread of seasonal influenza. Further analysis in relation to temperature data will be discussed throughout the next chapter.

7.6.2 Under-reporting

One of the main scopes of this study is to understand the under-reporting (GP cases vs. non-GP cases) factor. Throughout this section, I shall analyse the under-reporting rate of the seasonal influenza from different standpoints. In fact, I will consider five different measures in order to determine the reporting rate of the GPs data:

1. Diagnosed ILI cases from GP reporting (Chapters 2 and 4)
2. Respondents that had ILI symptoms based on survey data
3. Respondents that acquired seasonal influenza based on survey data
4. Individuals' temperature based on survey data
5. Seasonal influenza cases in households based on survey data

Furthermore, the reporting rate will be analysed based on the assumption that an individual might not acquire immunity after being diagnosed with seasonal influenza, and based on the assumption that individual acquire immunity after being diagnosed with seasonal influenza. Based on these two assumptions I will consider four different possibilities (cases) of reporting rate:

Case 1: Diagnosed ILI cases (GP data) against number of symptomatic cases (Survey data);

Case 2: Diagnosed ILI cases (GP data) against seasonal influenza cases (Survey data);

Case 3: Diagnosed ILI cases (GP data) against individuals' temperature (Survey data);

Case 4: Diagnosed ILI cases (GP data) against seasonal influenza cases in households (Survey data).

7.6.2.1 Case 1: Diagnosed ILI cases (GP data) against number of symptomatic cases (Survey data)

Since the survey was carried out amongst the population of 18 years of age and over, the above results were also assumed to be representative amongst those aged 17 years and younger. This was based on findings by Yang et al. (2015) [157], where for several influenza viruses, age groups revealed similar influenza patterns. Thus, it is possible to extrapolate the results amongst the whole population of Malta. By extrapolating my survey results, this means that the findings are assumed to be representative of the whole population; hence, for different results, the number of individuals can be calculated from the total Maltese population.

According to the GPs data (given in Chapter 4), in Malta there were around 32,000 seasonal influenza cases between October 2014 and Mid-May 2015. However, this does not mean that only these individuals had acquired the seasonal influenza.

According to the survey, after extrapolating the result over the whole population of Malta (425,384 [164]), around 360,000 individuals (84.6%) residing in Malta had a symptom directly or indirectly related to the seasonal influenza (ILI cases). From the above results, in total, the respondents reported that they had experienced any of these symptoms for 672 times (occurrences) during a one year period. Hence, one respondent might have had the influenza-related symptoms for more than one occurrence. In fact, on average the individuals claimed that they had experience these symptoms 1.9 times in different occurrences during the year. After taking this into consideration, the number of times Maltese citizens had experienced the above symptoms (as per survey 1) is estimated to be around 700,000 symptomatic occurrences. Hence, if we had to consider all these symptomatic cases (occurrences) as seasonal influenza cases, this would give a reporting rate of only 4.6% (32,000/700,000). However, this is a very crude estimation and in this case, a lot of assumptions are being taken into account. Primarily, we assume that all ILI individuals eventually acquired the seasonal influenza. Secondly, these individuals who had acquired seasonal influenza will not develop immunity from the same influenza virus (which is unlikely as defined in Chapter 1).

Based on the same variable (ILI symptomatic cases), we can work out a different percentage (survey data) by assuming unique ILI (symptoms only) individuals. Hence, in

this case we assume that these ILI symptomatic cases had resulted in seasonal influenza cases, and immunity from seasonal influenza was acquired. Thus, in total, there were around 360,000 individuals that who had at least one ILI symptom (survey 1 data). Based on the same logic, this implies a reporting rate of 8.9% (32,000/360,000).

The same analysis was carried out for the 2015/2016 survey. After comparing the latter reporting rates (non-unique ILI symptom individuals and unique ILI symptom individuals) with the 2015/2016 survey data by using the same methodology, this would result in the reporting rates equal to 5.1% (29,000/570,000) and 8.5% (29,000/340,000) respectively. Both percentages are similar to the 2014/2015 survey.

7.6.2.2 Case 2: Diagnosed ILI cases (GP data) against seasonal influenza cases (Survey data)

The data presented here allow us to analyse the number of seasonal influenza cases from different perspectives. In fact, according to the survey, it is being estimated and generalised for all the Maltese population, that around 130,000 people residing in Malta had the seasonal influenza (ignoring those who had responded with the 'don't know' option). This was based on the survey question which enquired whether respondents had the seasonal influenza (29.8%), and then scaled up to the whole population of Malta. However, all respondents mentioned that they had experienced seasonal influenza 183 times (Table 7.5). Hence, according to the survey results, one respondent might have had seasonal influenza more than once (≈1.5 times) (immunity is not being assumed). After taking this into account, the number of cases of seasonal influenza during the 2014/2015 season, is around 195,000 (through the use of Survey 1 data). By considering the total number of diagnosed ILI cases by GPs (≈32,000), the latter estimate (195,000) would give a reporting rate of 16.4% (32,000/195,000). This is a more conservative estimate when compared with the above under-reporting estimate. Primarily, we are assuming that only those individuals who claimed to have had the seasonal influenza actually had the same influenza. Furthermore, we are assuming that one individual might have acquired the seasonal influenza more than once (≈1.5 times).

If we had to assume immunity (more realistic, see Chapter 1 for details) from seasonal influenza (hence unique seasonal influenza individuals – 29.8%), this would imply a reporting rate of 24.6% (32,000/130,000).

By comparing the latter reporting rates (non-unique and unique seasonal influenza individuals) with the 2015/2016 survey data through the use of the same methodology, this would result in reporting rates equal to 14.5% (29,000/200,000) and 18.1% (29,000/160,000).

For case 2, there is a possibility that we are ignoring some of the individuals who claimed of having some of the above ILI symptoms, and also had the seasonal influenza (although replied negatively to seasonal influenza). Mainly, this is due to the fact that the above ILI symptoms are related to the seasonal influenza. Furthermore, we are also assuming that all individuals who claimed of having the seasonal influenza, actually had the same influenza.

7.6.2.3 Case 3: Diagnosed ILI cases (GP data) against individuals' temperature (Survey data)

One of the most significant symptoms of seasonal influenza is fever [159]. From all respondents, 28.6% (Table 7.2) claimed to have experienced fever during the year. If we had to consider this percentage as the number of seasonal influenza individuals, and consider that individuals had the seasonal influenza 1.5 times (according to survey results), we would have at total of 180,000 seasonal influenza cases during 2014/2015 (generalized through the survey), based on the total Maltese population. This result provides a reporting rate of 17.8% (32,000/180,000). Assuming immunity once again (i.e. acquiring the seasonal influenza only once), this provides a reporting rate of 26.7% (32,000/120,000). These results are similar to the previous two results of case 2 (16.4% and 24.6%).

The same calculations were carried out for the 2015/2016 survey through the use of the same methodology as above. This translates to a reporting rate of 20.7% (29,000/140,000) for non-unique cases and 26.6% (29,000/110,000) for unique cases.

7.6.2.4 Case 4: Diagnosed ILI cases (GP data) against seasonal influenza cases in households (Survey data)

From the above 2014/2015 survey results, 61.1% of all households in Malta had at least one household member with seasonal influenza. Furthermore, on average 1.8 members (according to survey 1) within the Maltese households had the seasonal influenza. According to the Maltese National Statistics Office (NSO), the total number of households in Malta is around 140,000 [164]. By using the latter data and taking into account that an individual had the seasonal influenza 1.5 times (Survey 1) during the same season (immunity is not assumed), we can estimate that there were around 230,000 seasonal influenza cases during the 2014/2015 season. Therefore, based on the GPs data, this result indicates that the reporting rate for the seasonal influenza is 13.9% (32,000/230,000). Similarly as above, when assuming immunity from seasonal influenza the reporting rate increases to 20.6% (32,000/155,000). When applying the same methodology but for the 2015/2016 dataset, this provides reporting rates of 21.5% (29,000/135,000) (no immunity assumed) and 27.6% (29,000/105,000) (immunity assumed).

7.6.3 Practical use

The above information is of interest to key people within the health authorities. After several meetings which I held with health authorities in Malta, it emerged that the above information is of high importance for their strategies, health promotion campaigns and planning (Appendix A). Information related to the size of outbreaks and characteristics related to the influenza are of interest to the Malta Health Promotion Department as it helps them to plan the level and strength of their campaigns. The estimate related to the occurrence of the peak of the influenza is one of the major priorities for the Department of Health Information and Research in Malta. Such information helps the department to submit refined information to different key health officials in decision-making positions. Health Ministry officials are mostly interested in the spread of influenza, specifically in predicting the demand on the local health care system. The implications of the seasonal influenza includes a huge cost on the health sector [165]. In fact, during the above meetings it emerged that, due to the seasonal influenza, a substantial amount of doctors are required and a higher number of hospital beds are occupied during this period

(Appendix A). By predicting the demand on the health sector, officials will be able to plan adequate bed management in hospitals.

By predicting the size of the outbreak, health promotion campaigns can be adjusted according to the size of the disease, and so will contribute to the control of the number of infected individuals, resulting in a lower demand on doctors and hospitals [40, 155]. In turn, this will reduce the number of illnesses, the mortality rates and lower the costs on the entire health sector.

The above results serve as a good basis to acquire further informative priors for the parameter estimation and predictive epidemiological modelling. Nevertheless, for future predictions of diseases, the above data may potentially provide improved prior values when compared to non-informative prior values (used in Chapters 3 and 4). Furthermore, this information will aid in designing a package of different sources of information in support to the prediction of current and future influenza outbreaks.

Further work is warranted to understand to what extent these surveys can contribute if they had to be conducted during an actual outbreak. This could lead to refined prior parameters during the course of the disease, providing even further refinements beyond this analysis. Throughout the next chapter we will use some of the above information to further improve our knowledge of outbreaks. It is clear that when another survey was carried out during the 2015/2016 influenza season, although a high number of results were similar, actual percentages of the seasonal influenza varied. As discussed in the literature review, one of the key variables that impacts on the influenza is the temperature. This might be clearly one of the main differences between some of the results obtained from the 2014/2015 and 2015/2016 datasets. Throughout the next chapter, we will be exploring this important variable (temperature) to understand the extent of use of the temperature data, together with the survey data as part of a package of information to predict the outbreaks.

7.7 Conclusion

There are limited studies that focus on a similar scope to this chapter, i.e. to estimate the under-reporting rate of the seasonal influenza. In fact, as defined in the literature review,

most of the research studies focus on surveys related to vaccine uptake and issues related to the seasonal influenza vaccine. Furthermore, there are few research papers that carried out cross-sectional nationwide surveys with similar research objectives.

Scientific surveys can provide detailed information to understand the real notion of seasonal influenza, and to offer an opportunity to improve the prior information for future epidemiological modelling. However, we need to treat such results with caution. To a certain extent, we are comparing self-diagnosis of individuals against the GPs influenza diagnosis. Hence, the baseline for both numbers is not necessarily the same. The self-diagnosis provides an estimate of the actual influenza cases based on personal perception. Nevertheless, there are several indicators throughout the survey that have shown that these results are a true representation of the actual population. The monthly data between the survey and GPs data (Figures 7.5-7.7) match fairly well, thus providing an extra level of confidence that the respondents are accurately remembering their medical history for the past year.

Based on the realistic estimates and lower number of assumptions, we have shown in the 2014/2015 survey that the reporting rate might vary between 13.9% and 17.8% when immunity from seasonal influenza is not assumed. When assuming immunity from seasonal influenza, the reporting rate varies between 20.6% and 26.7%. For the 2015/2016 survey, when immunity is not assumed, the reporting rate varies between 14.5% and 21.5%. For the latter survey (2015/2016), the percentages vary between 18.1% and 27.6% when immunity is assumed. The other estimated reporting rate of 4.6% (Case 1, 2014/2015 survey) was based on the assumption that any of the above mentioned symptoms resulted in acquiring seasonal influenza. This is rather a crude assumption, as some of the symptoms are related to a common cold (Case 1). In fact, only 56.5% of those respondents who had these symptoms felt the need to stay at home to recover. Estimated reporting rates between the two surveys are rather similar. Due to previous discussions in this chapter and in chapter 1, it is more likely that an individual experiences the seasonal influenza only once during the same season (thus acquiring immunity). Hence, the above reporting rates suggest that the reporting rate in Malta might vary between 18% and 28%, producing an average of 23%. Therefore, this implies an under-reporting rate of 77%. This is in accordance with the mean posterior reporting rate

parameter δ (23% - 29%) in chapter 4 (Table 4.4) for the diagnosed seasonal influenza datasets.

Through the survey data, we were able to estimate the number of unique individuals in Malta who had acquired seasonal influenza during 2014/2015 season, which is between 120,000 and 150,000. This means that between 28% and 36% of the Maltese citizens had seasonal influenza during the 2014/2015 period, while for the 2015/2016 season, this varied between 100,000 (24.5%) and 160,000 (37.2%) individuals. According to the Centers for Disease Control and Prevention (CDC), the seasonal influenza in the United States affects between 5% and 20% of the total population [165]. In Finland, it was estimated that 6% were infected during the first wave of the pandemic 2009/2010 season and 3% during the second wave [166]. None of these estimates were based on cross-sectional surveys, but rather based on on-line data [165] and national surveillance data [166]. However, one cannot really directly compare Malta's incidence rate with other countries, as Malta is an island and one of the most densely populated countries in the world. Given the contact between people is more likely to occur in Malta, the higher the frequency of face-to-face contact between individuals, the higher the incidence rate of the influenza [167].

Chapter 8

Forecasting seasonal influenza outbreaks: The new influenza model

8.1 Introduction

In chapter 5, I used the SEIR model and the joint model (SEIR and linear regression model) to carry out real-time forecasts. However, our main obstacle was to obtain early and consistent accurate forecasts that can provide real-time predictions [2]. This was due to various limiting factors such as limited information. In relation to this, when a substantial number of consultations and diagnosed cases were recorded in Malta, the relationship between consultations and diagnosed datasets was rated as strong (Chapter 4). Furthermore, it was established (Chapters 3 and 4) that posterior parameter estimates were a reliable source of information to employ in future influenza outbreaks. In fact, it was found that diagnosed posterior parameter values are consistent when compared across different influenza seasons (Chapter 4). Through a national cross-sectional survey (Chapter 7), we also showed that a significant portion of the population do not visit their GP to be examined for their ILI symptoms. Furthermore, we established that respondents might misinterpret the real meaning of seasonal influenza (Chapter 7), or that GPs might misdiagnose individuals with influenza (Chapter 4). All of these results shall converge in this chapter. Here, we ask the research questions whether we can find a better framework (than that in chapter 5) to predict future outbreaks and how early this can be done. At the end of this chapter, we will use the 2015/2016 seasonal influenza dataset as a model example to apply real-time forecasting through the use of the new influenza model that I will develop in this chapter.

The 2009/2010 pandemic season was not included as part of this analysis since the pandemic data has different characteristics when compared to the seasonal influenza datasets (as described in previous chapters). Furthermore, throughout this chapter I will focus on the diagnosed seasonal influenza cases rather than including the consultations data. The diagnosed dataset is a more direct proxy of the influenza outbreak, as it only includes individuals that were diagnosed with an ILI. Ultimately, by accurately predicting the number of diagnosed cases, we can predict the consultation cases as in chapters 4 and 5 (this will be explored in Chapter 9).

The weekly posterior parameter estimates obtained from the previous seasonal influenza datasets will be used for the scope of this analysis. These parameter estimates modelled

accurately the diagnosed datasets for 2011-2015 seasons (Figure 4.6) and hence can serve as a basis for analysing future outbreaks.

More specifically, in the following sections, the estimates of the effective reproduction ratio (R_t) (Chapter 6) will be used to understand the relationship between the diagnosed ILI data and the temperature data. We conjecture that temperature and particularly temperature changes can be used to predict the onset of the outbreak in a given season.

8.2 Results

8.2.1 Malta's temperature data

The temperature distribution in Malta during the seasonal influenza period tends to be rather consistent (Figure 8.1) across different years. During week 40, over four different seasons, the temperature varied between 20°C and 24°C. Subsequently, between weeks 6 and 9, the temperature in general reached the lowest levels. At this point, the temperature varied between 9°C and 14°C. By the end of the influenza season (week 20), the temperature was within the range of 18°C and 20°C. On average, the lowest mean temperature for the whole season was registered for the 2011/2012 season, with an average of 15.1°C (Std. Dev. 3.4°C). This was followed by 2014/2015 season (15.7°C, Std. Dev. 3.8°C), 2013/2014 season (15.9°C, Std. Dev. 3.4°C) and 2012/2013 season (16.2°C, Std. Dev. 3.6°C). The range of the average temperatures for the four different seasons is only 1.1°C.

The 2011/2012 diagnosed ILI cases are negatively correlated with the temperature data for the same period (Table 8.1), such that the Pearson correlation value between both datasets is -0.71 (p-value < 0.001). This shows that lower temperature values tend to provide higher values of diagnosed ILI cases. These results are echoed in the two subsequent datasets (2012/2013 and 2013/2014), however with lower correlation values. For the 2012/2013 season, the correlation value (Table 8.1) between both data sets is -0.60 (p-value < 0.001), while the 2013/2014 season registered a moderate negative correlation value of -0.59 (p-value < 0.001). For the 2014/2015 season, the correlation value reached once again a value equal to -0.71 (p-value < 0.001). The final Pearson correlation value is the same as that obtained for the 2011/2012 season.

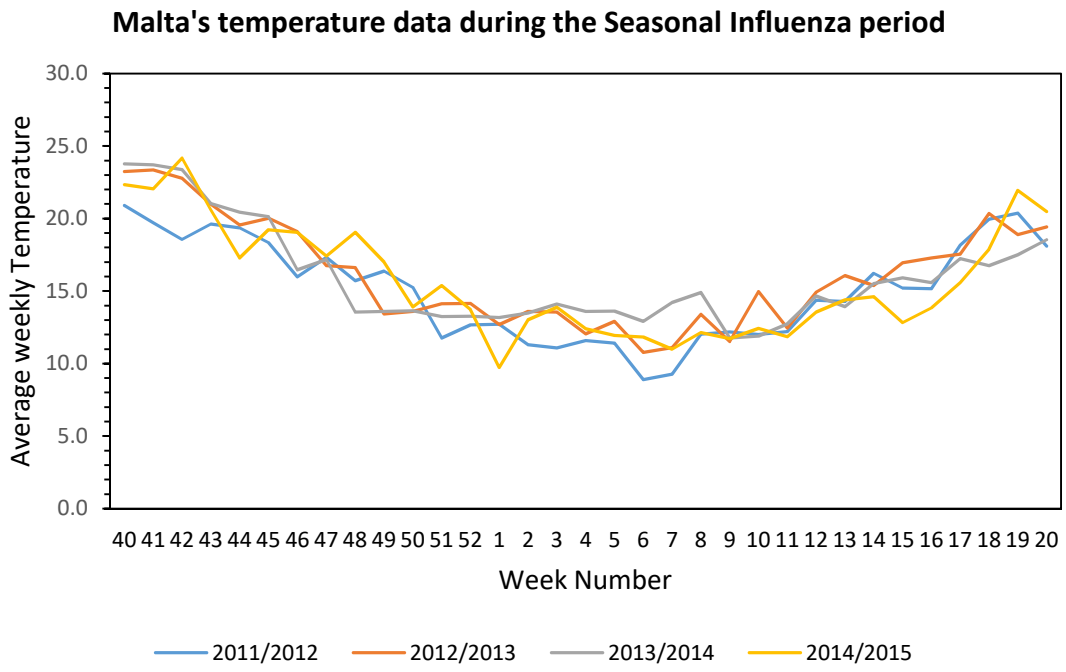


Figure 8.1 – Malta's temperature data during the four influenza seasons. The horizontal axis represents the week number, while the y-axis represent the average weekly temperatures in Malta. In general, over the years, temperature data have the same characteristics during the influenza season.

Influenza Season	Pearson Correlation Value	P-value
2011/2012	-0.71	< 0.001
2012/2013	-0.60	< 0.001
2013/2014	-0.59	< 0.001
2014/2015	-0.71	< 0.001

Table 8.1 – The Pearson correlation values when comparing the diagnosed ILI cases with the temperature data for the four individual seasons. The p-value is the test of associations between the two variables (as described in Chapter 2).

Figure 8.2 shows that during the 2011/2012 season, as temperature decreases below 14°C, the diagnosed ILI cases rise substantially and remain consistently high for around 3 months. During this period, the temperature remained lower than 14°C. Soon after the temperatures exceed the 14°C threshold, the number of diagnosed cases dropped to the same levels as before the temperature decreased below 14°C (Dec'11). Furthermore, the first drop in temperature below the 14°C was preceded with a 15.3°C and followed by 11.8°C. By comparing the difference in these two temperatures, this can be considered as a significant sharp drop in temperature (change of 3.5°C week-on-week). On average, the difference in week-on-week temperatures for the whole 2011/2012 influenza season

is 1.2°C. Thus, 3.5°C is almost three times as much higher than the average week-on-week difference. Furthermore, this is the largest drop in temperature throughout the whole influenza season.

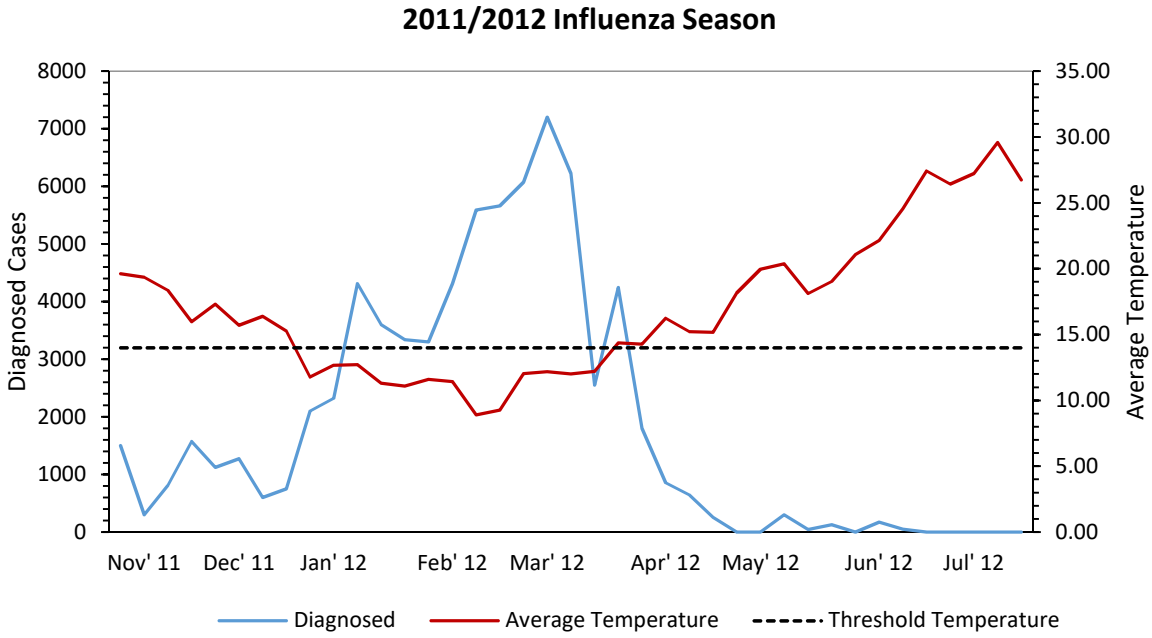


Figure 8.2 – This figure represents the 2011/2012 diagnosed ILI data (blue line) and the temperature data for the same period (red line). The diagnosed ILI cases are plotted on the left y-axis and the temperature data is on the right y-axis. The horizontal dashed line represents the threshold temperature of 14°C.

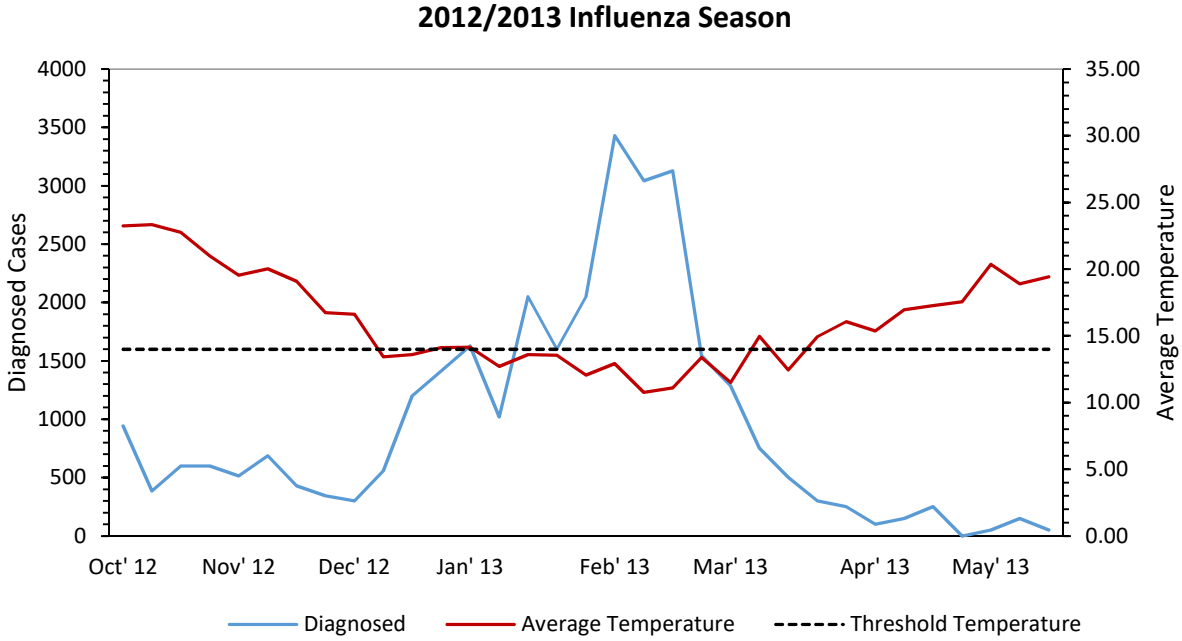


Figure 8.3 – This figure represents the 2012/2013 diagnosed ILI data (blue line) and the temperature data for the same period (red line). The diagnosed ILI cases are plotted on the left y-axis and the temperature data is on the right y-axis. The horizontal dashed line represents the threshold temperature of 14°C.

Figure 8.3 provides a parallel picture for the 2012/2013 season. In general, the first temperature below 14°C triggers a substantial increase in the number of diagnosed ILI cases. The temperature remains below this threshold (14°C) for around 3 months (with the exception of weeks 51 and 52 where the temperature exceeds the 14°C threshold by ≈0.1°C). When the temperature exceeds the 14°C, the seasonal influenza is at the declining stages (to the same levels as before the first drop below the 14°C threshold). Another comparable result to the previous influenza season is that before the first decline below the 14°C threshold, the previous temperature was 16.6°C, followed by 13.4°C. Hence, there is a difference of 3.2°C, which can be considered as a significant and largest sharp drop when compared to the average temperature difference on a week-on-week basis (1.3°C) for the same season.

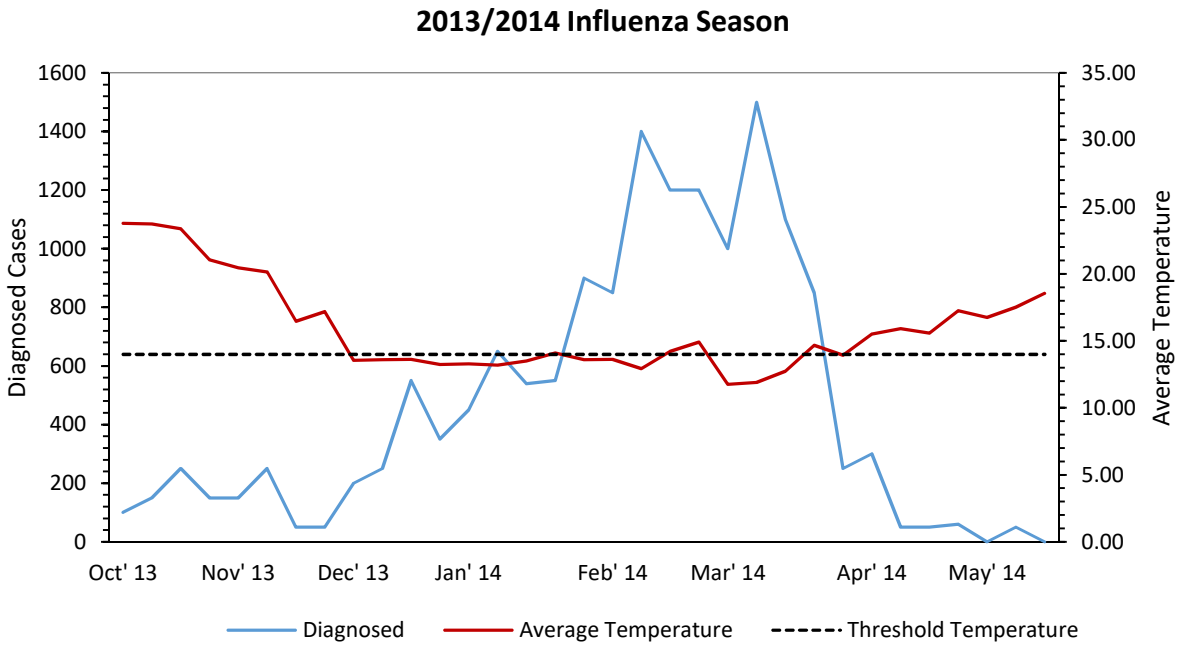


Figure 8.4 – This figure represents the 2013/2014 diagnosed ILI data (blue line) and the temperature data for the same period (red line). The diagnosed ILI cases are plotted on the left y-axis and the temperature data is on the right y-axis. The horizontal dashed line represents the threshold temperature of 14°C.

For the 2013/2014 season, the temperature remained below the 14°C threshold for almost 16 consecutive weeks (Figure 8.4) with an exception during week 4 (14.1°C), week 7 (14.2°C) and week 8 (14.9°C), where there was a slight temperature increase above the 14°C threshold. Similar to the above, during the period when the temperatures declined below the 14°C threshold, the diagnosed ILI cases began to increase. The number of influenza cases remained high until the air temperature started to become warmer (higher than 14°C). As shown earlier, the first drop below the 14°C (13.6°C) was preceded by a

significantly higher temperature (17.2°C). This is another substantial (and largest) difference of 3.6°C, while the week-on-week average temperature difference for the whole season is 0.9°C.

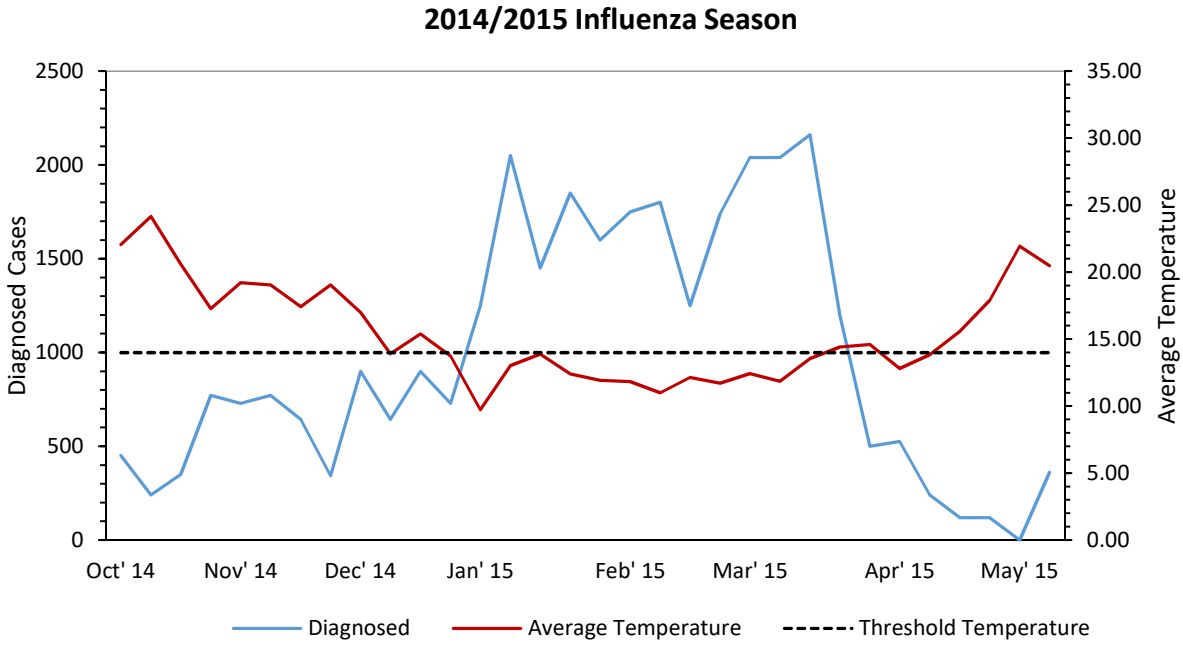


Figure 8.5 – This figure represents the 2014/2015 diagnosed ILI data (blue line) and the temperature data for the same period (red line). The diagnosed ILI cases are plotted on the left y-axis and the temperature data is on the right y-axis. The horizontal dashed line represents the threshold temperature of 14°C.

A similar pattern can be seen for the 2014/2015 season. As soon as the temperature dropped below 14°C, the diagnosed ILI cases increased significantly and remained high for 13 weeks. This coincides precisely with the first temperature which had exceeded the 14°C threshold following these 13 weeks. In fact, after these 13 weeks, the diagnosed cases declined sharply, and reached the same level of diagnosed cases before the temperature dropped below the 14°C threshold. The first drop below 14°C was registered with a temperature of 13.9°C, though preceded by a temperature of 17°C. Hence, the week-on-week temperature difference is 3.1°C. The latter difference can be considered as another sharp and largest drop, considering that the average week-on-week temperature difference for the 2014/2015 influenza season was 1.6°C.

We will use the 2015/2016 temperature data for out-of sample testing in this chapter; hence in this section we are not showing a similar analysis to the above (Figures 8.1-8.5). Figure 8.6 combines all the above relationships together through a scatter plot for all the above four seasonal influenza datasets. This figure shows that lower temperatures tend to

provide a higher number of diagnosed cases, while higher temperatures imply an exceptional low number of diagnosed cases.

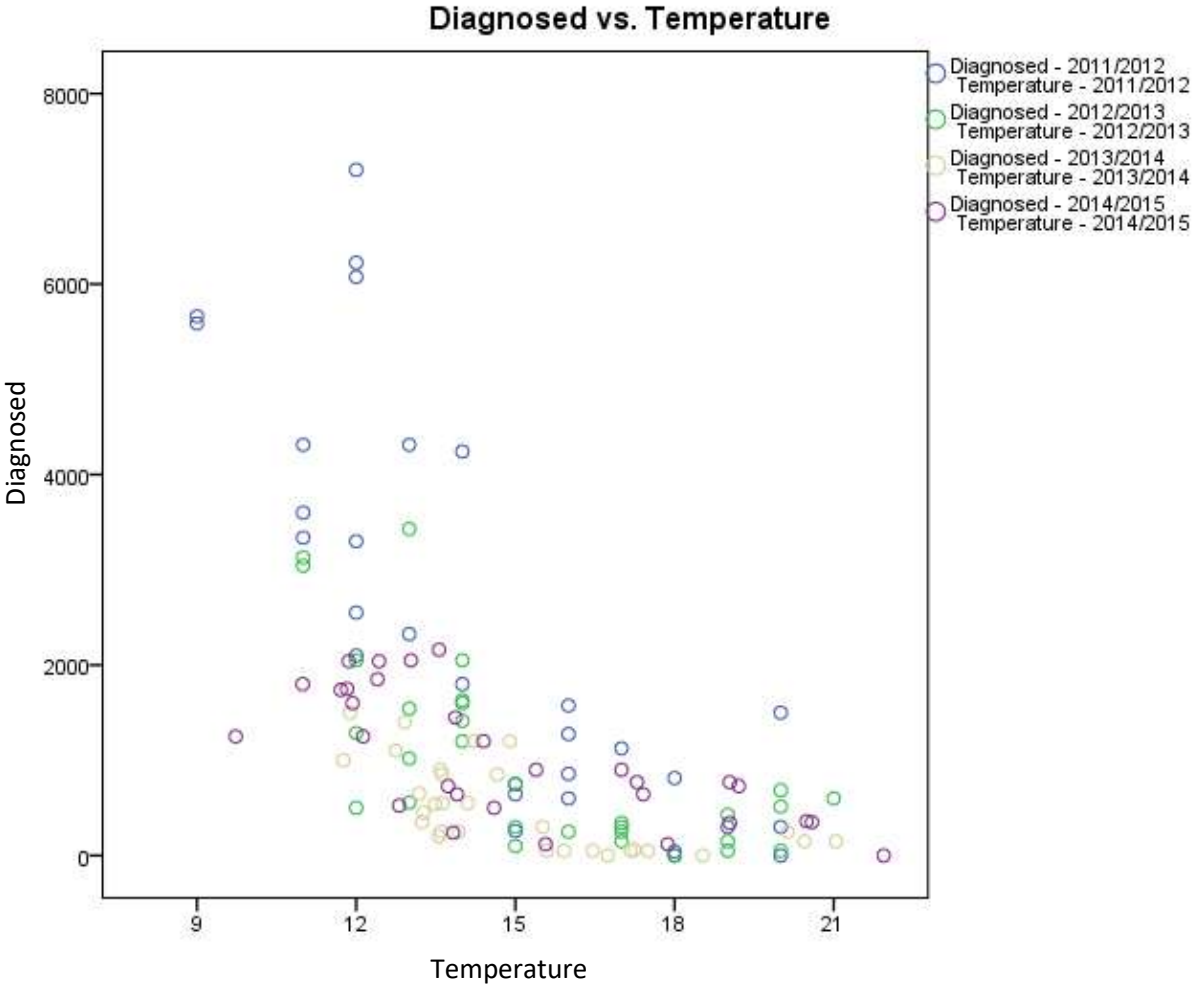


Figure 8.6 – Scatter plot for all the four seasonal influenza datasets. The y-axis represents the number of diagnosed cases while the x-axis represents the temperature data. This shows that lower temperatures imply a higher number of diagnosed ILI cases.

8.2.2 Malta's temperature data in relation to R_t

In chapter 4, we obtained four effective reproduction ratio datasets (Figure 4.7) for the four seasonal influenza periods. These were analysed collectively and in relation to the diagnosed ILI datasets. Furthermore, in chapter 6, the initial R_t values were analysed in detail and it was decided that some of these initial values are unreliable and would need to be excluded. Therefore, the improved R_t values (Figures 6.6-6.10) will be used for the scope of this analysis.

Figures 8.7-8.10 show that when the temperature is below the 14°C threshold, this coincides with R_t values greater than one. Moreover, the R_t values vary during this

particular period (temperature < 14°C). In contrast, when the R_t values decline below the value of 1, the temperature increases steadily while the R_t value remains almost constant for the late part of the season.

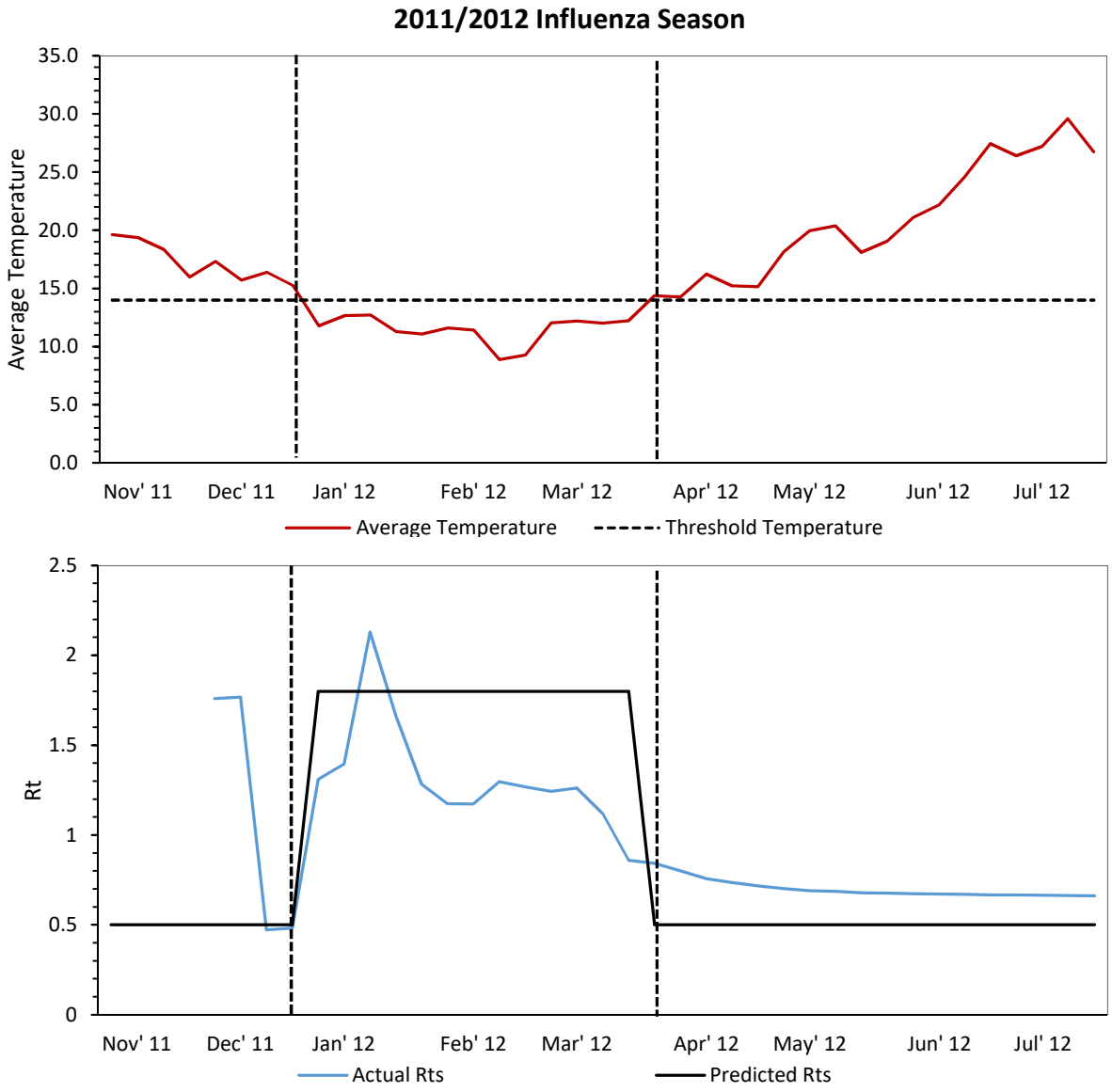


Figure 8.7 – The upper charts represent the temperature data as defined above, together with a horizontal line which represents the threshold of 14°C. The chart below compares the reproduction ratio (blue line) obtained from chapters 4 and 6, and the estimated reproduction ratio chart (black line) obtained through the temperature data. The first vertical dashed line represents one data point before the temperature declines below the 14°C, while the second vertical dashed line represents one data point after the temperature exceeds the 14°C threshold. The initial R_t values were eliminated from the R_t chart as defined in chapter 6.

The figures (Figures 8.7-8.10) indicate that, in general, a sharp drop in temperature triggers the reproduction ratio to exceed the value of 1 and hence corresponds to a sharp increase in the number of diagnosed cases (as described above). Almost in all cases

(Figures 8.7-8.10), when R_t is below one and then proceeded by another R_t value greater than one, the outbreak registers the first highest significant increase in the diagnosed cases. Furthermore, based on the latter results, when the temperature is greater than 14°C, the influenza is either not severe (R_t<1) or is at the termination phase (R_t<1). Hence, the initial R_t values which are greater than one, are rather unrealistic. As discussed in chapter 6, one has to treat the initial R_t values and parameter values cautiously anyway.

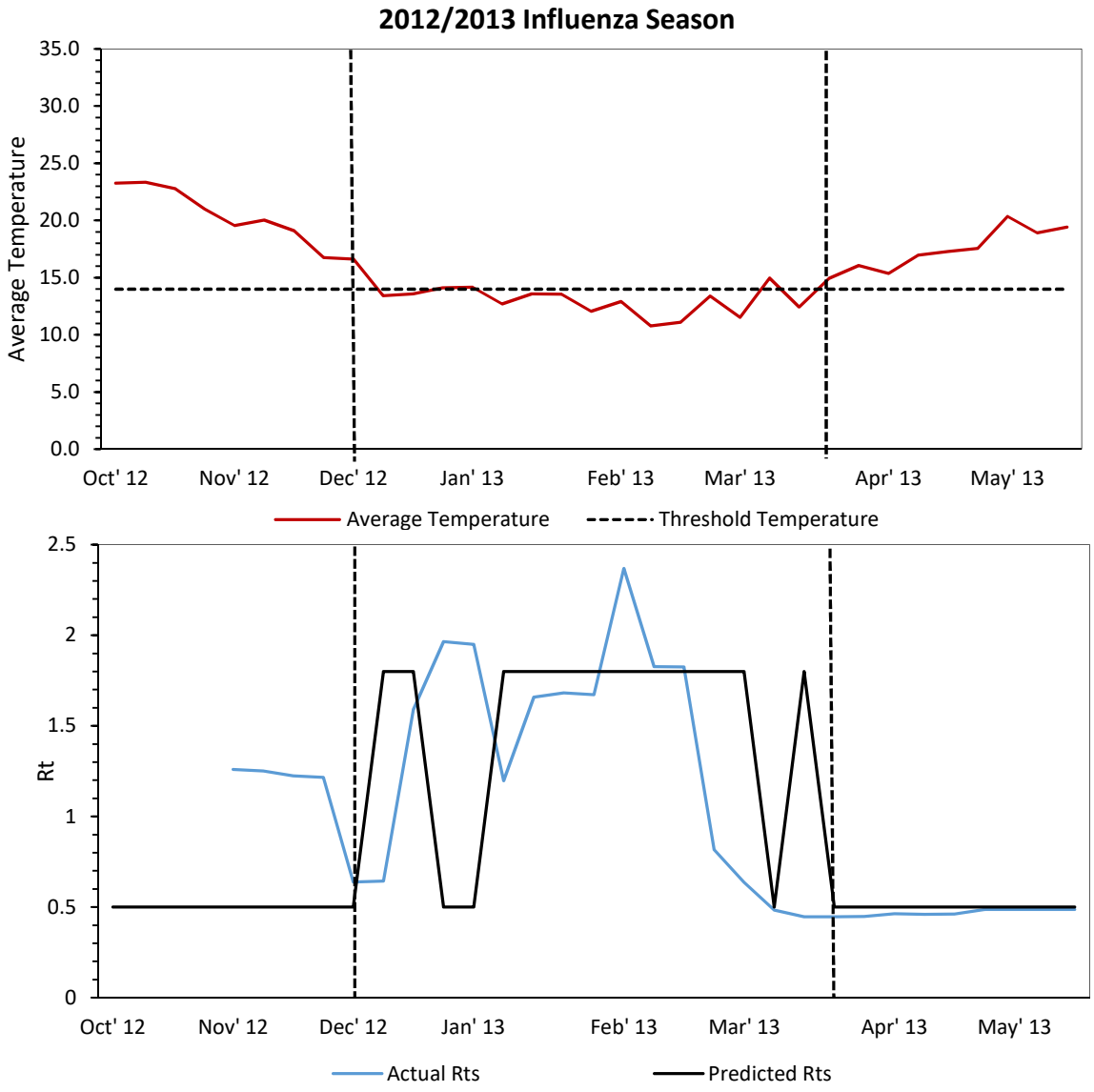


Figure 8.8 – The upper charts represent the temperature data as defined above, together with a horizontal line which represents the threshold of 14°C. The chart below compares the reproduction ratio (blue line) obtained from chapters 4 and 6, and the estimated reproduction ratio chart (black line) obtained through the temperature data. The first vertical dashed line represents one data point before the temperature declines below the 14°C, while the second vertical dashed line represents one data point after the temperature exceeds the 14°C threshold. The initial R_t values were eliminated from the R_t chart as defined in chapter 6.

Figure 8.7 compares R_t with temperature for the 2011/2012 season. For the period when the actual R_t values are greater than one (except the initial R_t values), these R_t values correspond to a temperature less than 14°C. Based on this threshold (14°C), we can propose a model for the reproduction ratio values. The newly constructed R_t chart (Figure 8.7, black solid line) accurately predicts when the actual R_t values are greater than one or less than one. Hence, the temperature data can be used as a strong basis to predict the reproduction ratio.

Figure 8.8 provides the predicted R_t values based on the temperature data for the 2012/2013 season. For weeks 51 and 52, the 14°C threshold was exceeded by 0.1°C (as described above), hence that is why the predicted R_t plot registered a dip during the initial part. Subsequently, during the peak of the influenza, the temperature data provide a good indication of the values of the effective reproduction ratio.

During 2013/2014 season, the temperature during the influenza season was consistently close to the 14°C threshold (Figure 8.9). Due to this, the predicted R_t plot registered some fluctuations and inconsistencies. However, the first sharp drop below the 14°C still produced a positive signal that the reproduction ratio will start to increase shortly, also resulting in an increase in the number of diagnosed cases.

The 2014/2015 season dataset is not an exception when compared to the previous seasonal influenza datasets (Figure 8.10). Some of the R_t initial values were accurately predicted, especially when predicting the R_t values which are greater than one. Subsequently, during an advanced period of the seasonal influenza period, the actual R_t values are slightly below the value of one, while the temperature is still below the 14°C threshold. The number of seasonal influenza cases was still high during the same period.

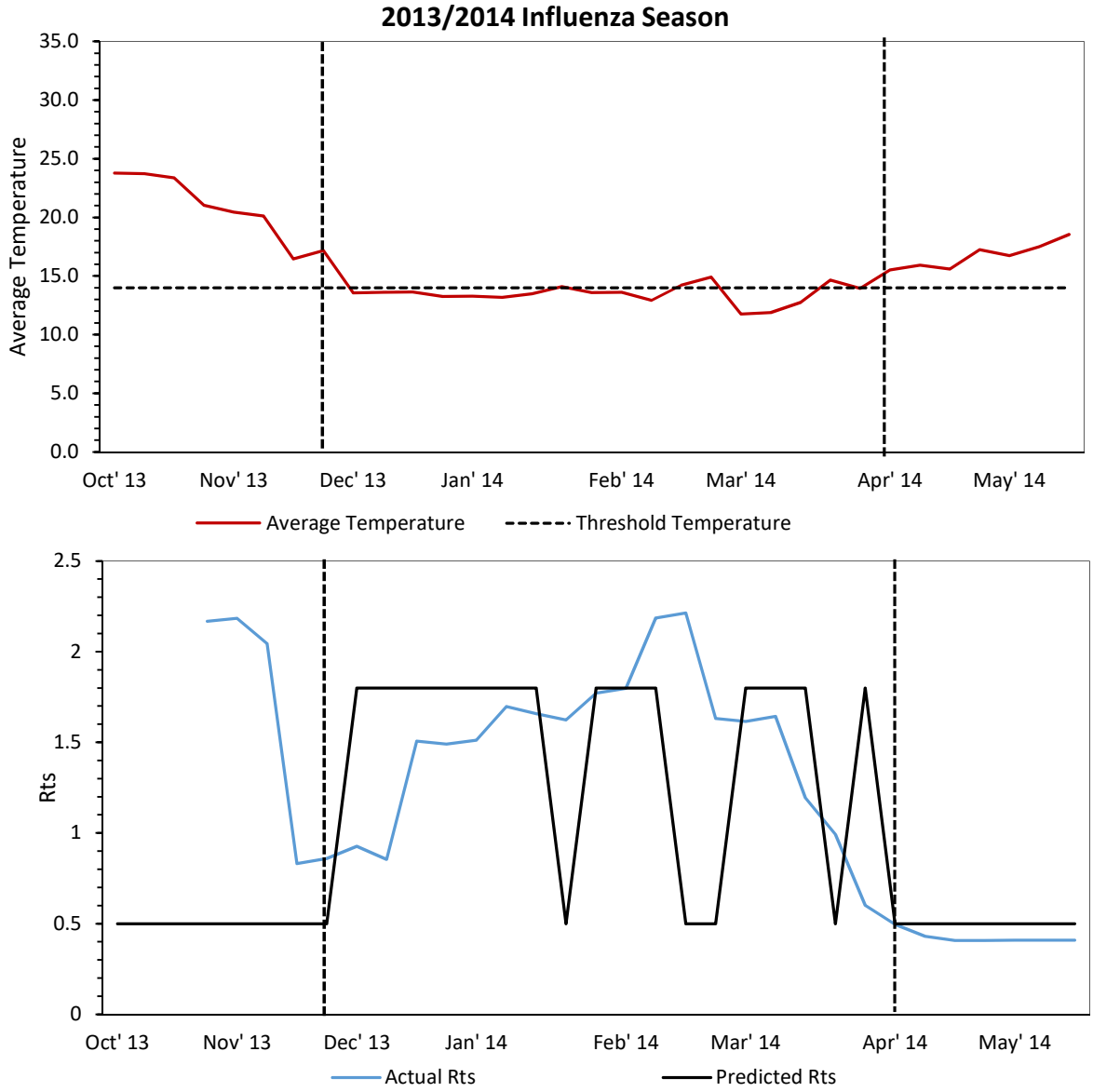


Figure 8.9 – The upper charts represent the temperature data as defined above, together with a horizontal line which represents the threshold of 14°C. The chart below compares the reproduction ratio (blue line) obtained from chapters 4 and 6, and the estimated reproduction ratio chart (black line) obtained through the temperature data. The first vertical dashed line represents one data point before the temperature declines below the 14°C, while the second vertical dashed line represents one data point after the temperature exceeds the 14°C threshold. The initial R_t values were eliminated from the R_t chart as defined in chapter 6.

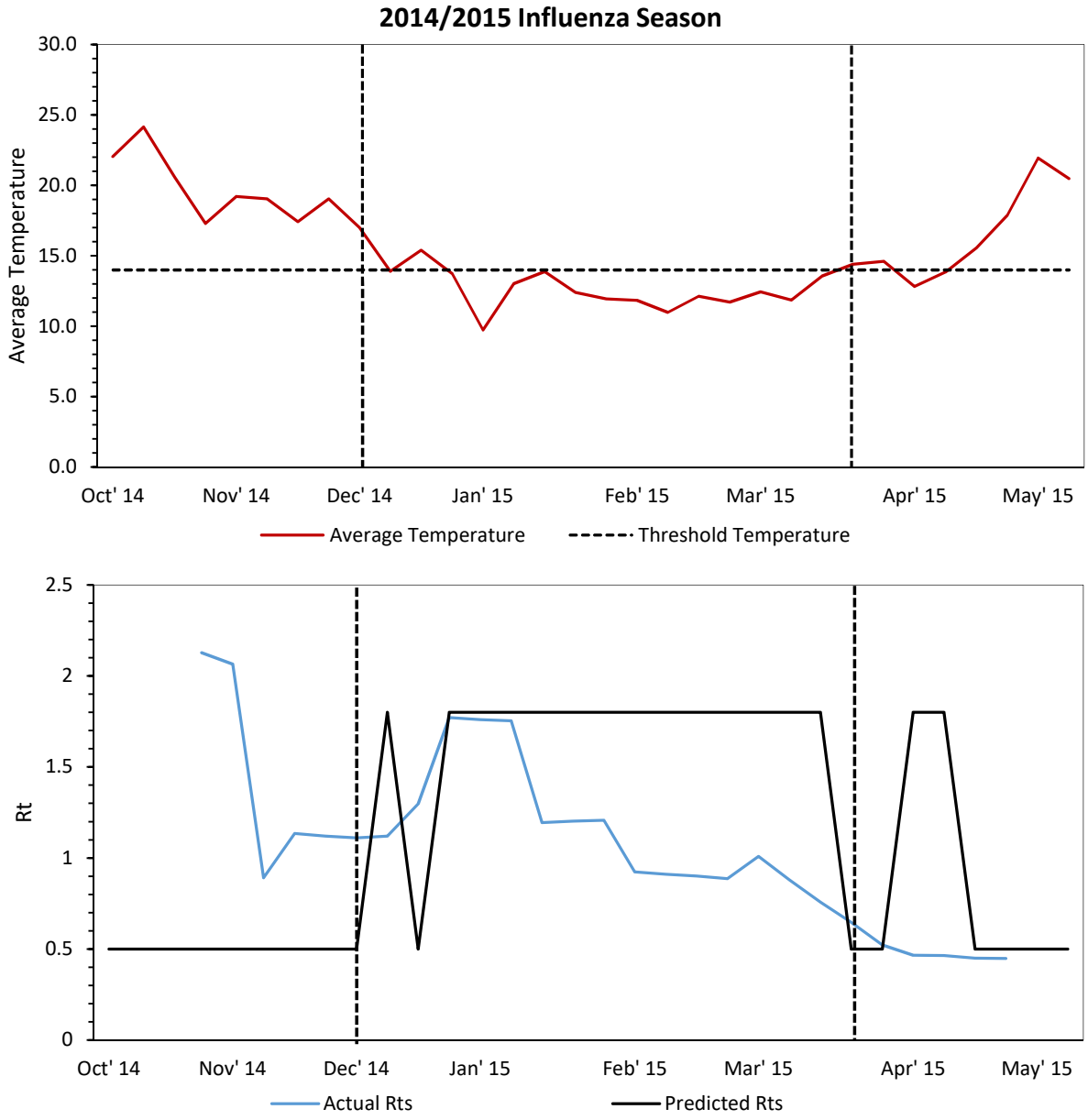


Figure 8.10 – The upper charts represent the temperature data as defined above, together with a horizontal line which represents the threshold of 14°C. The chart below compares the reproduction ratio (blue line) obtained from chapters 4 and 6, and the estimated reproduction ratio chart (black line) obtained through the temperature data. The first vertical dashed line represents one data point before the temperature declines below the 14°C, while the second vertical dashed line represents one data point after the temperature exceeds the 14°C threshold. The initial R_t values were eliminated from the R_t chart as defined in chapter 6.

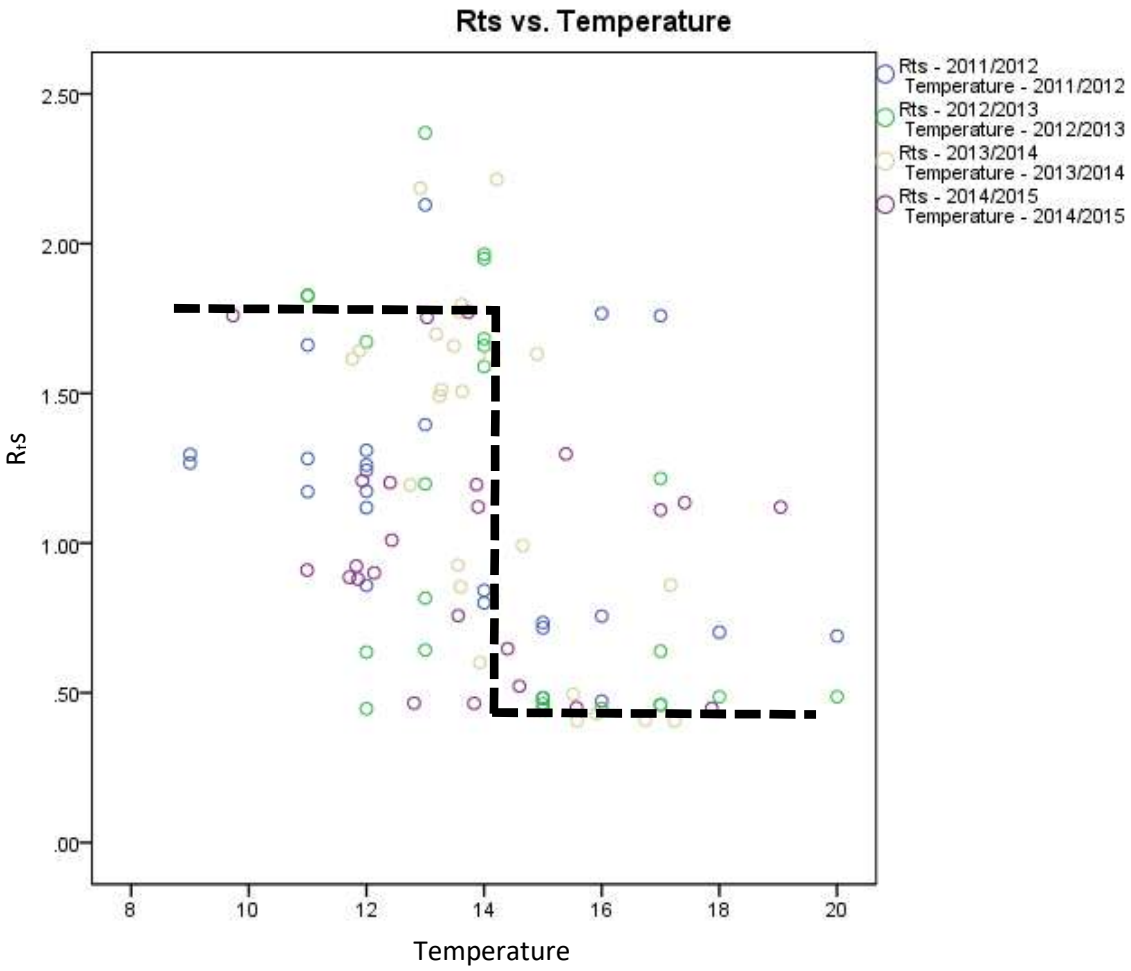


Figure 8.11 – Scatter plot between the reproduction ratios (y-axis) obtained through the analysis carried out in chapters 4 and 6 and the temperature data (x-axis). The vertical dashed line represents the temperature of 14°C. The horizontal lines are rough estimates for a reproduction ratio (1.8) when the temperature is less than 14°C, and the reproduction ratio (0.5) when the temperature is greater than 14°C.

Figure 8.11 shows that there is a weak (r = -0.4) negative correlation between the R_t values and the temperature data. However, during some periods of the influenza season, this relationship becomes stronger as shown in figures 8.7-8.10. The vertical dashed line in figure 8.11 represents the temperature of 14°C. The horizontal lines are a rough estimate of the reproduction ratio (1.8) when the temperature is less than 14°C, and the reproduction ratio (0.5) when the temperature is greater than 14°C. The main scope of the two latter reproduction values is to provide further meaning to the scatter points. These two latter ratios tend to be the most reasonable values to represent the reproduction ratio when it is greater than 1 (R_t = 1.8) and when it is less than 1 (R_t = 0.5), see figures 8.7-8.10.

8.2.3 The posterior parameter values

In chapter 4, we showed that the posterior parameter distributions accurately described the observed datasets (Figure 4.6). As discussed in chapter 4, posterior parameter values for different diagnosed datasets can be used for future outbreaks. Figure 8.12 represents all the time-dependent posterior parameters $(\beta, \varepsilon, \alpha, \tau, \emptyset, \delta)$ for the diagnosed ILI datasets for all the four influenza seasons. For these parameters, I used the new $R(0)$ values which were established in chapter 6.

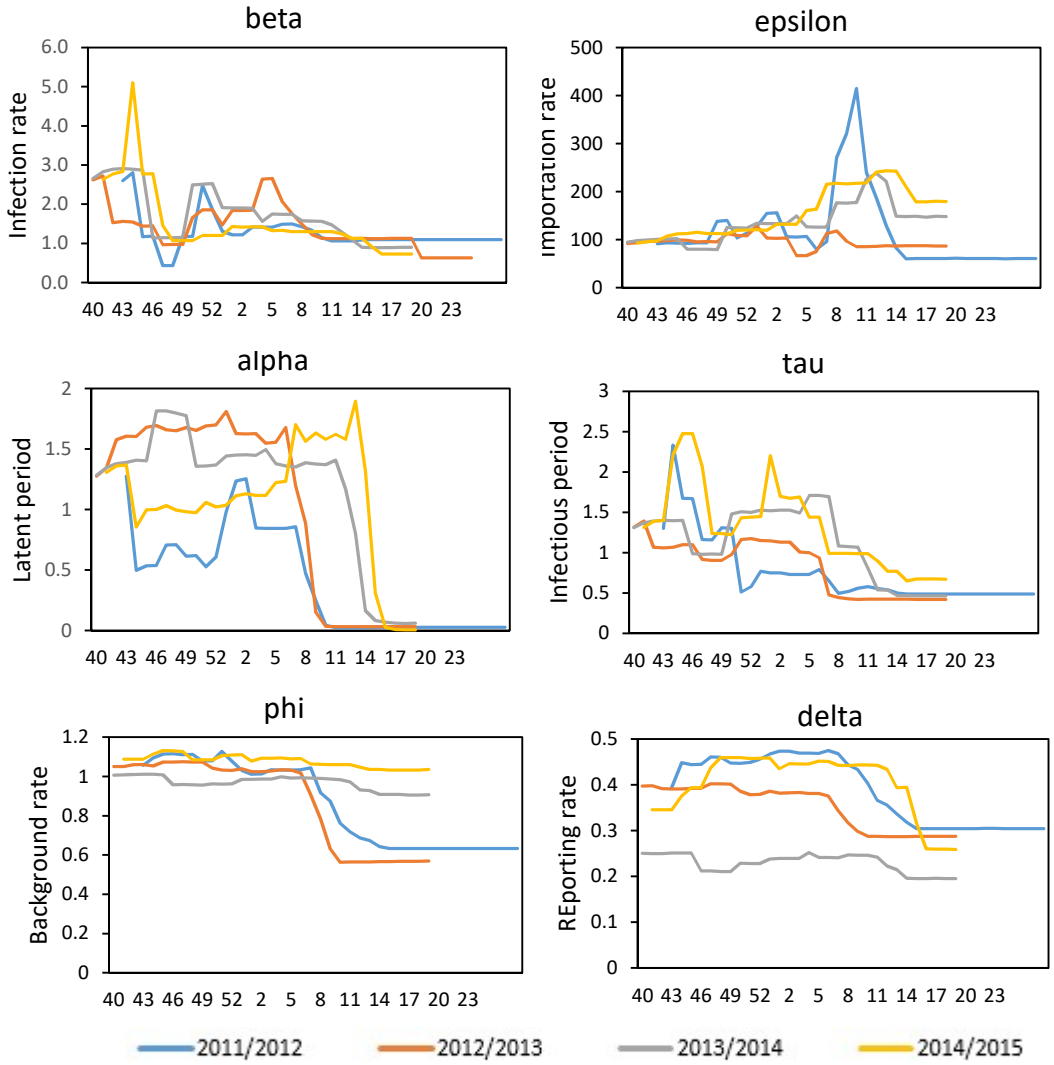


Figure 8.12 – Weekly posterior parameters for the diagnosed ILI datasets. The x-axis represents the week number. The blue lines corresponds to the 2011/2012 posterior parameters, the orange line corresponds to the 2012/2013 parameters, the grey line corresponds to the 2013/2014 parameters while the yellow line corresponds to the parameters for the latest dataset (2014/2015).

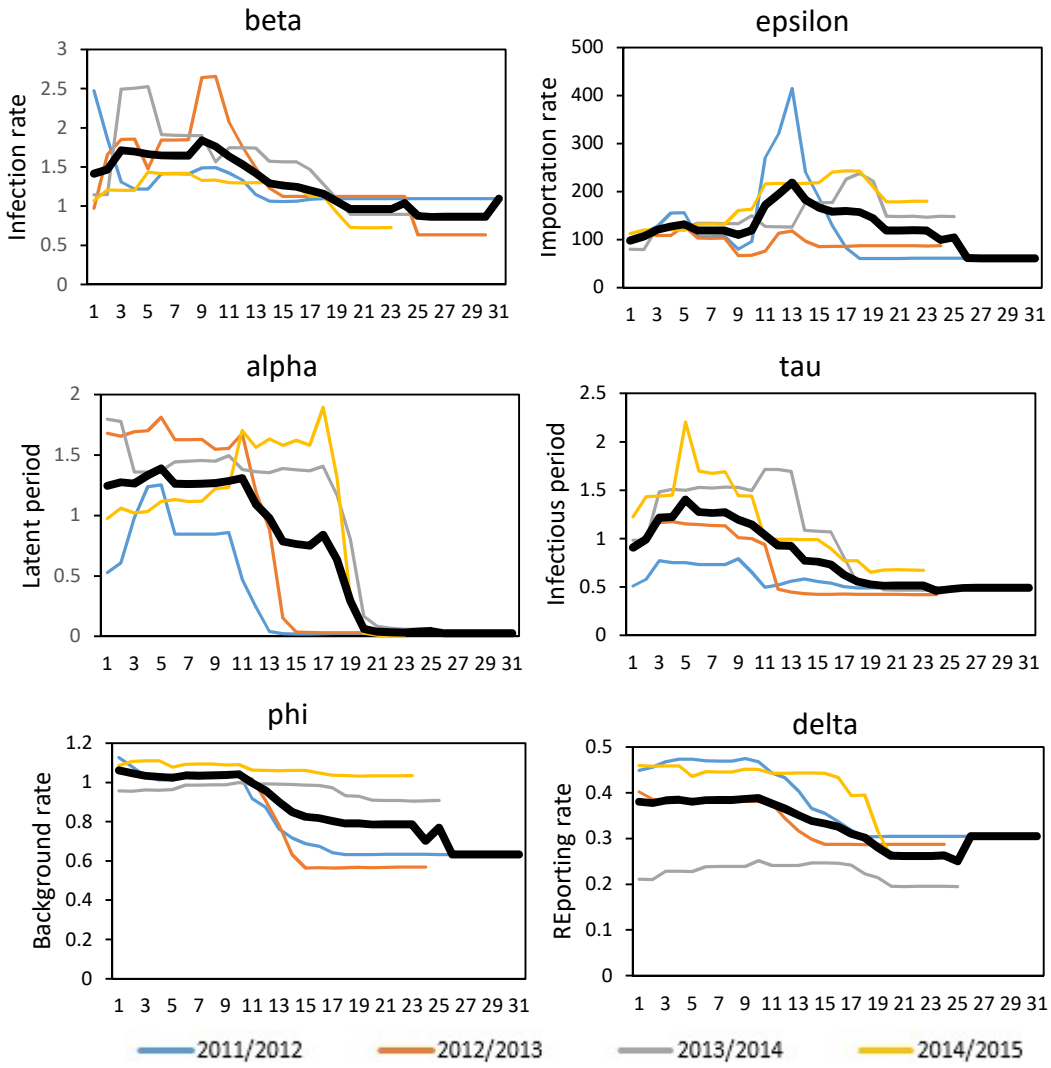


Figure 8.13 – The average parameter values for all the four seasonal influenza datasets. The first data point was assumed equal to the first drop in temperature below the 14°C threshold. Therefore, at $t'=1$ for all the average posterior distributions, this is equal to the average of all the first posterior parameter values when the temperature is less than 14°C for the first time. The blue lines corresponds to the 2011/2012 posterior parameters, the orange line corresponds to the 2012/2013 parameters, the grey line corresponds to the 2013/2014 parameters while the yellow line corresponds to the parameters for the latest dataset (2014/2015).

The time series of the parameters in figure 8.13 are now shifted by an interval that is dependent on the season, so that $t'=1$ always corresponds to the first drop below 14°C. For every parameter value $(\beta, \varepsilon, \alpha, \tau, \emptyset, \delta)$, the posteriors for all the influenza seasons were averaged to obtain one typical shape for every parameter of the SEIR model (Figure 8.13) and for the observation model D_t (as described in Chapter 2). All the previous posterior values before this particular data point (first drop below 14°C) are ignored.

The forthcoming methodology will forecast the outbreak based on the 14°C threshold. As explained above, the first temperature below the 14°C threshold, triggers the 'real' start (the first substantial increase in the number of diagnosed cases) of the epidemic. All the weekly diagnosed ILI cases prior to this starting point are low and stationary. However, when the temperature drops below the 14°C, the influenza starts to rise. Therefore, at $t'=1$ for all parameters we choose the average of all the first parameter values (for all four diagnosed datasets), corresponding in each seasons to the time point when the temperature is less than 14°C for the first time.

8.2.4 The 2011-2015 seasonal influenza datasets

Throughout this section I will be using the average time-dependent posterior parameters together with the SEIR model to predict the number of infected individuals by the end of the season. In this analysis the SEIR model will be used without the particle filter algorithm. In order to carry out this computation, a script in R was prepared for this analysis (Appendix H).

Since throughout this analysis we are assuming the 'real' start of the seasonal influenza when the temperature drops below 14°C, then the initial values for $S(0)$, $I(0)$, $E(0)$ and $R(0)$ are estimated from the number of diagnosed ILI individuals at this particular point. The number of infected individuals ($I(0)$) is assumed equal to the number of diagnosed ILI cases when the temperature drops below the 14°C threshold (Case 1), the number of exposed individuals is assumed to be equal to $I(0)$, the number of recovered individuals ($R(0)$) is equal to 150,000 except for the 2013/2014 dataset (as defined in chapter 6), and the number of susceptible individuals ($S(0)$) is the population size without all the previous initial values. However, a second case for analysis was also considered. In fact, for the initial value of $I(0)$, the average number of reported diagnosed ILI cases (by GPs) during the first two weeks, when the temperature was less than 14°C for the first time, was calculated (Case 2), thus resulting in a new value for $I(0)$. All the other initial values ($S(0)$, $E(0)$, $R(0)$) follow as defined above.

The number of diagnosed individuals (Case 1 and Case 2) at $I(0)$ are given in table 8.2. These initial values are related to the total number of seasonal influenza cases for the whole season. In fact, a higher initial number of influenza cases corresponds to a higher

number of diagnosed cases for the entire season (Table 8.2). For the 2011/2012 season, the high initial value (2,100) coincides with the highest number of diagnosed ILI cases (73,202) when compared to the other three seasons. On the contrary, for the 2013/2014 season, the low number of diagnosed ILI weekly cases (at $I(0)$) corresponds to the lowest number of total diagnosed cases for the entire season.

Forecasts were considered for the same time periods that were taken into consideration in chapter 4. Although for the 2011/2012 season data was collected till August 2012, only data till July 2012 was considered for the analysis, as almost no diagnosed ILI cases were recorded. For the other datasets, data until the middle of May was considered.

	Number of diagnosed ILI cases when the temperature < 14°C, for the 1st time (Case 1)	Average of the first two weekly data points (diagnosed ILIs) when temperature < 14°C, for the 1st time (Case 2)	Total number of diagnosed influenza cases being taken into consideration (as reported by GPs)
2011/2012	2,100	2,220	73,202
2012/2013	550	850	31,299
2013/2014	200	220	15,450
2014/2015	650	775	31,514

Table 8.2 – Column 2 represents the number of weekly diagnosed influenza cases as soon as the temperature drops below 14°C for each influenza season. Column 3 represents the average of the first two weekly data points as soon as the temperature drops below 14°C. The fourth column represents the total number of diagnosed ILI cases (as reported by GPs) for the entire influenza season.

Based on the above average 'shifted' parameters (as in Figure 8.13), the SEIR model (without the particle filter algorithm) was run for each individual season. For case 1, the average parameter values (Figure 8.13) from $t'=1$ onwards were used to run the SEIR model. For case 2, the average of the first two weekly data points (diagnosed ILIs) when temperature drops below 14°C was considered. Hence, for case 2 we used the average parameter values (Figure 8.13) from $t'=2$ onwards. The values of $S(0)$, $I(0)$, $E(0)$ and $R(0)$ were chosen differently for each individual season, as defined above. The SEIR model produced the weekly predicted values for different compartments. However, the number of infectious cases over time (I_t) were then incorporated in the observed model D_t (as described in Chapter 2) to be able to compare the GPs reported data against the new predicted data. When applying the model D_t the 'shifted' parameters $\emptyset$ and δ (Figures 8.13) were used (see section 2.6.1 for further detail about the model D_t).

Subsequently, the total number of predicted diagnosed ILI cases for the entire season was calculated for the four individual influenza seasons.

The model predicts rather accurately the total number of diagnosed ILI individuals as at the end of each individual season. In fact, when considering case 1, for the 2011/2012 dataset, the model predicted the total number of diagnosed ILI individuals with a precision of 84% (Table 8.3) when compared with the total number of actual diagnosed cases. However, when considering a more informative initial $I(0)$ value (Case 2), the prediction improved by nine percentage points (93%). For the 2012/2013 season, when considering case 1, the precision of the prediction reached 83%, while when considering case 2, the prediction almost matched the actual number of diagnosed ILIs accurately (Table 8.3). Similarly when applying case 2 for the 2014/2015 dataset, the prediction improved (99%), when compared with case 1 for the same diagnosed ILIs dataset (92%). The 2013/2014 dataset registered similar results between case 1 (105%) and case 2 (104%).

Both for case 1 and case 2, all the previous observed diagnosed ILI cases until when the above threshold is reached (14°C) were considered known. From then onwards, the forecasts were calculated. In general, using the 10^{th} and/or 11^{th} data points were enough to accurately predict the total number of diagnosed ILI cases by the end of the season.

	Case 1 Forecast	% of actual cases	Case 2 Forecast	% of actual cases	Total diagnosed influenza cases during the whole outbreak (GPs data)
2011/2012	61,642	84%	67,615	93%	73,020
2012/2013	26,102	83%	31,321	100%	31,299
2013/2014	16,225	105%	15,998	104%	15,450
2014/2015	29,125	92%	31,203	99%	31,564

Table 8.3 – The total number of forecasted influenza cases by the end of the influenza season for both cases defined in table 8.2. Columns 3 and 5 represent the precision of both cases when compared to the total number of diagnosed influenza cases as reported by the GPs (column 6).

8.2.5 Real-time forecasting of the seasonal influenza

Although the total number of cases can be predicted well by the SEIR model, the actual shape of the epidemic cannot (Appendix I). In this section I go round this problem by finding a 'typical' shape of an epidemic and then scaling it by the estimated total number

from the SEIR model (Table 8.3). Hence, in general, by mid-December the real-time forecasting was applied. For all cases, mid-December is still an early time point for the whole influenza season. Hence, predictions are being calculated for the subsequent five months.

The weekly ratio of the number of influenza cases were obtained from the total number of diagnosed cases (GPs) of each individual season (Figure 8.14). Hence, the ratio of time point 1 was calculated by the total number of reported ILI cases at $t=1$, over the total number of reported diagnosed ILI cases by the end of the season, and similarly for all the other time points and for each influenza season. The average ratios for each individual time point were then calculated, resulting in a 'typical shape' of the diagnosed ILI cases in Malta over time (Figure 8.14). In general, this 'typical shape' represents all seasons quite well.

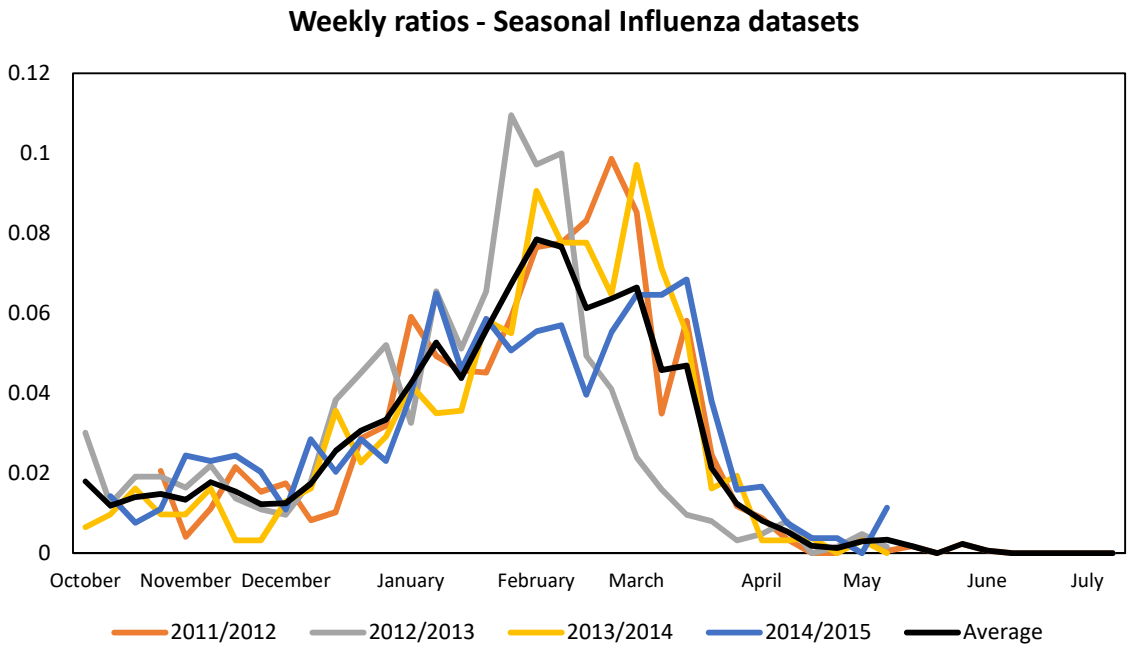


Figure 8.14 – Weekly ratios of the number of influenza cases relative to the total number of influenza cases for each respective year. The black solid line represents the average ratios at each individual time point for all the four influenza seasons.

Having established that the 'typical shape' represents each scaled diagnosed dataset (i.e. Diagnosed divided by a total), I conjecture that another good model representation of the data can be obtained by multiplying the 'typical shape' by the total number of diagnosed ILIs predicted by the model (Table 8.3). Since case 2 predictions are more accurate (Table 8.3), these model forecasts were used for the following analysis. Hence, for every influenza season, the total number of predicted diagnosed cases over-time were obtained.

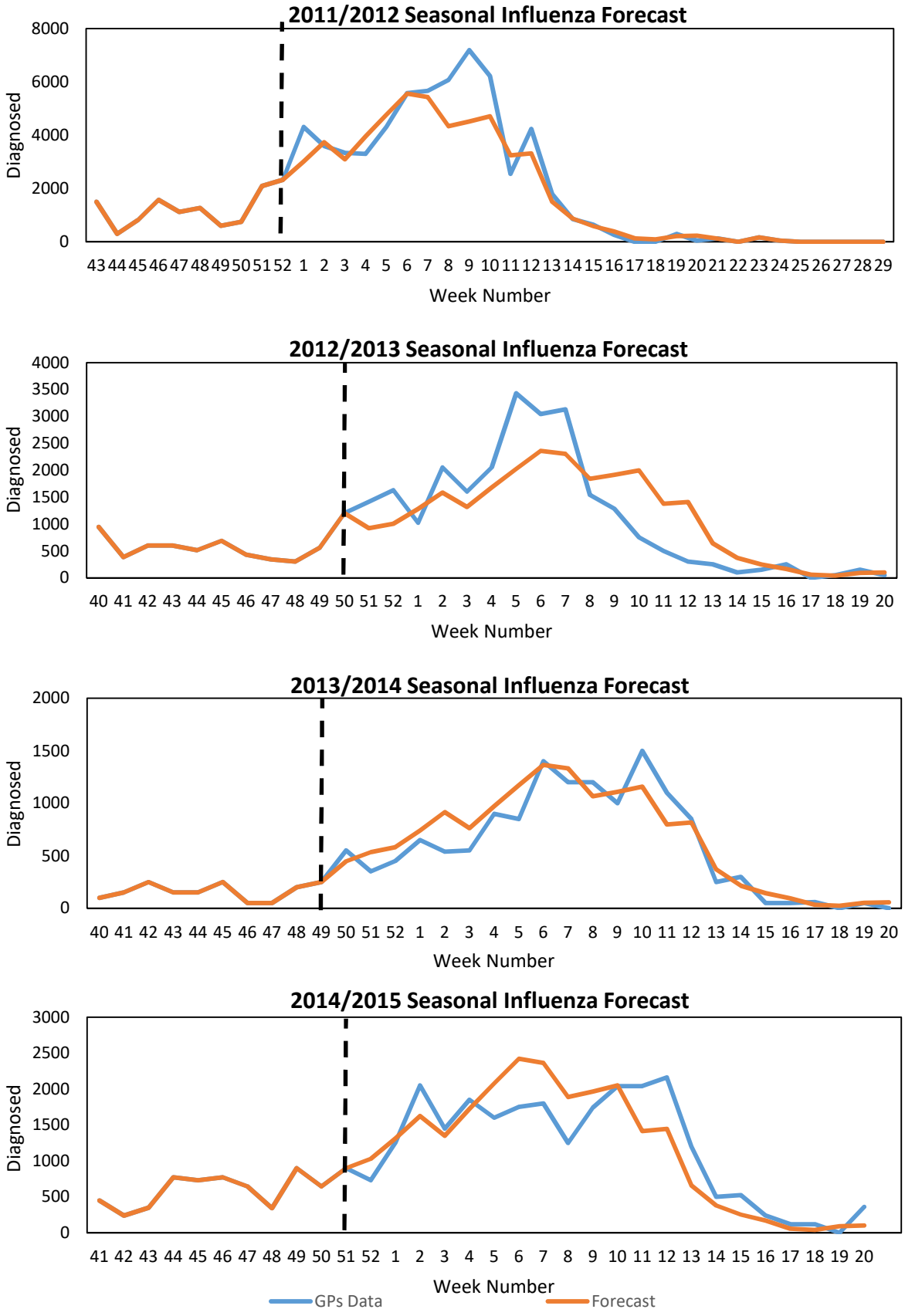


Figure 8.15 – Comparison of weekly diagnosed datasets for all the four influenza seasons. Data includes the actual data (blue line) and the results of the model fit for the diagnosed ILIs (orange line). The vertical dashed line is the time point when the model was applied.

Figure 8.15 provides all forecasts of the influenza seasons based on the typical shape of the number of diagnosed individuals in Malta (Figure 8.14). For all four seasonal influenza datasets, figure 8.14 together with the results of table 8.3 produced accurate forecasts (Figure 8.15).

8.2.6 The 2015/2016 Seasonal Influenza

As explained above, the temperature data offers a strong signal of the initial start of the seasonal influenza and the end stages of the same influenza. Thus, I will be using the temperature data to assist in forecasting the number of seasonal infected cases by the end of the 2015/2016 season.

The 2015/2016 influenza dataset was used to test the above methodological framework. The initial $I(0)$ was assumed to be equal to 550, based on the average two consecutive numbers of weekly diagnosed cases, when the temperature drops below the 14°C threshold for the first time. This happened at the 11^{th} data point (week 50) of the 2015/2016 influenza season (similar to other influenza seasons). All the other future diagnosed cases were assumed unknown, while previous parameter values of other diagnosed datasets were used. The first temperature below the 14°C threshold was 13.7°C. The latter was preceded by 14.1°C and followed by 11.7°C. The average week-on-week difference up till the latter point (which is the last known data point being considered in this example) was 1.3°C. Although the first difference (14.1°C-13.7°C) is rather small, the second drop in temperature (13.7°C-11.7°C) can be considered substantially higher than the average week-on-week difference (1.2°C) for the entire season.

Hence, the estimated time-dependent average parameters $(\beta, \varepsilon, \alpha, \tau, \emptyset, \delta)$ (as defined above, Figure 8.13) were incorporated in the SEIR model and the observation model D_t to predict the number of diagnosed ILI cases by the end of the season. $I(0)$ was assumed equal to 550, $E(0)$ equal to 550, $R(0)$ equal to 150,000 and $S(0)$ equal to 273,900.

	Forecast	% of actual cases	Total diagnosed influenza cases during the whole outbreak (GPs data)
2015/2016	26,784	92%	29,090

Table 8.4 – The total number of forecasted influenza cases by the end of the 2015/2016 influenza season. Column 3 represents the precision of the total number of forecasted influenza cases (Column 2) when compared to the total number of diagnosed influenza cases as reported by the GPs (Column 4).

The above model predicted 92% of diagnosed ILI individuals when compared with the actual total number of diagnosed ILI cases (Table 8.4) until the end of the season. In fact, the forecast of the total number of diagnosed cases is 26,784, while the total diagnosed cases during the whole outbreak (GPs data) is equal to 29,090. Furthermore, the wave of the outbreak can be predicted accurately by taking into account the above typical shape (Figure 8.14), based on the previous seasonal influenza datasets. Indeed, by the 13th of December 2015, we obtained rather accurate predictions (Figure 8.16) for the remaining weekly data points (till mid-May).

As discussed above, the new methodology obtained an improved model fit when compared to the prediction charts obtained in chapter 5. The estimated wave (Figure 8.15) is a reasonable representation of the 2015/2016 diagnosed ILI dataset.

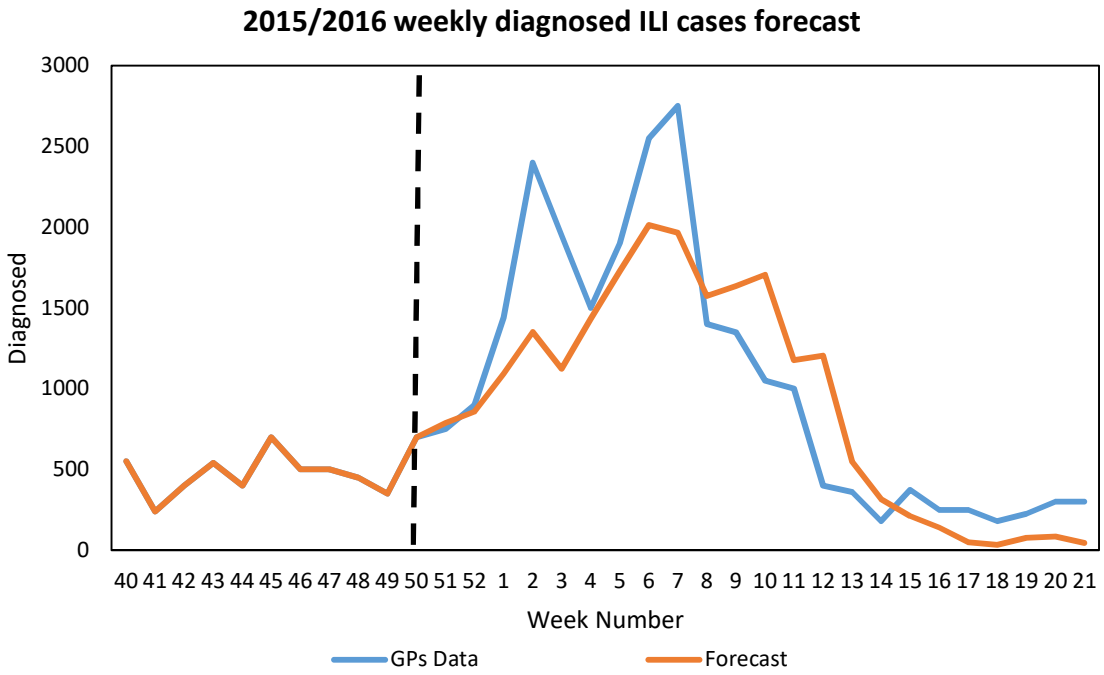


Figure 8.16 – Comparison of the weekly diagnosed cases for the 2015/2016 influenza season. Data includes the actual data (blue line) and the results of the model fit for the diagnosed ILIs (orange line). The vertical dashed line is the time point when the model was applied.

By using the estimated under-reporting rate (Chapter 7), we can actually calculate an estimated range of the total number of seasonal influenza cases for the entire season. Hence, based on the reporting rates (according to the survey findings) which vary between 20.6% and 26.7%, one can forecast the total number of influenza cases in Malta by mid-December. In fact, the range of the total number of influenza cases in Malta for

the entire season can be estimated to be between 99,000 and 130,000 influenza cases. In these calculations, we are assuming that once a person is infected, the patient acquires immunity from seasonal influenza during the same season.

8.3 Discussion

The above results provided a new methodology on how to combine different data sources into one integrated model to predict the seasonal influenza wave. Information from previous chapters was used to improve the modelling framework. The new defined methodology provides an improved way of how to predict the seasonal influenza outbreak, when compared to the SEIR model or the joint model (Chapters 5 and 6). Furthermore, the SEIR posterior parameters obtained from the previous datasets were also incorporated in the final model. The temperature variable was shown to be an important factor related to the seasonal influenza. For all datasets, there is a moderate negative correlation between the temperature and the number of diagnosed individuals. Hence, the lower the temperature, the higher the number of diagnosed seasonal influenza individuals. Furthermore, there is an important observed threshold of 14°C. For temperatures below this observed threshold, this corresponds to the first substantial increase in the number of diagnosed cases. In general, as soon as the temperature drops below the 14°C threshold, the number of diagnosed ILI cases is on the same levels as to when the temperature exceeds this threshold at the end of the influenza season. Thus, the first temperature drop in Malta below the 14°C threshold triggers an increase in the number of diagnosed cases, and hence triggers a rise in the reproduction ratio. In general, the number of seasonal influenza cases remains high for a period of 13 weeks. This corresponds to the number of weekly temperatures below the 14°C threshold for the above seasonal influenza datasets. Following these 13 weeks, the temperature exceeds the 14°C threshold again, and the diagnosed ILI cases stabilize to the same level before the temperature drops below the 14°C threshold. Therefore the latter result can predict an adequate estimate of the reproduction ratio.

When considering the average of the first two data points when the temperature drops below 14°C, the prediction improved due to more informative initial values. It is very clear that this average value, at this particular point, provides a strong signal of the severity of the influenza for the entire season. In fact, a low initial diagnosed ILI number

implies a mild influenza season, while a higher initial value results in an intense influenza season.

We also note that a sudden drop (≈3°C drop) in temperature is associated with the initiation of the epidemic. In contrast, the average change in temperature on a week-to-week basis was found to be rather low ((≈1.3°C). However when the temperature drops below the 14°C, in most cases this was preceded by a higher temperature of approximately 3°C. This drop in temperature represents the largest drop in temperature throughout the entire influenza season.

In general, the predicted number of diagnosed ILI cases (through the use of the new model and methodology) was 90-110% accurate when compared to the total number of actual diagnosed ILI cases. Such accuracy was obtained during the early stages of the influenza season (≈ week 50), and 5 months in advance before the end of the influenza season. Furthermore, based on the previous distribution of diagnosed cases, we were able to produce a typical curve which is representative of all the four diagnosed datasets. It is known that for countries and regions where the temperature varies throughout the year, the influenza outbreaks follow this pattern, where the activity reaches its peak during mid-winter [100].

When correlation analysis was applied to compare the actual diagnosed ILIs and the forecasted data, correlation values were all strong. For the 2011/2012 dataset, the Pearson correlation value obtained was 0.933 (p-value < 0.001), for the 2012/2013 dataset a correlation value of 0.910 (p-value < 0.001) was obtained, and for the subsequent two datasets the Pearson correlation values were 0.855 and 0.854 (p-value < 0.001 respectively). For the latest dataset (2015/2016) where the above methodology was tested, the Pearson correlation value is 0.916. These values show that the above model and methodology can be considered strong and reliable. Furthermore, this places the utilised methodology at the top most accurate forecasts when compared to the extant research papers that focused on influenza forecasting [120].

8.3.1 The New Model

Based on the above results, I propose a new model and methodology to predict the seasonal influenza outbreak, which is presented in figure 8.17 below. As already defined in detail, this new model requires the following procedure to obtain the influenza real-time forecasts:

1. By using the influenza datasets for the previous years, run the SEIR model with particle filter algorithm to obtain the posterior parameter values for each diagnosed dataset. For this step, all the historical influenza data points can be considered as known; hence the SEIR model needs to fit the known datasets accurately (as shown in Chapter 4, Figure 4.6).

2. For each individual posterior parameter estimate $(\beta, \varepsilon, \alpha, \tau, \emptyset, \delta)$, this needs to be averaged across all different influenza outbreaks in order to obtain one time series for each posterior parameter. For the above fixed average parameter estimates, the parameter values are to be averaged from the first data point which corresponds to a temperature lower than 14°C. Therefore, $t'=1$ for the above parameters corresponds to the temperature when it is less than 14°C for the first time. All the previous posterior values before this particular data point are ignored.

3. Consider the 'real' start of the influenza season as the first data point when the temperature drops below the 14°C threshold. This particular threshold is based on Malta's datasets and characteristics. Therefore, this needs to be investigated and tested further for other countries.

4. In order to estimate the initial value of the infected individuals ($I(0)$), calculate the average of the two weekly data points (total diagnosed reported cases) at a time when the temperature falls below the 14°C for the first time.

5. In order to estimate $R(0)$, either make use of the methodology defined in chapter 6 (Sensitivity Analysis), or utilise the previous initial values of other influenza datasets, or use previous influenza survey results.

6. Based on steps 4 and 5, estimate $E(0)$ and $S(0)$. $E(0)$ can be assumed equal to $I(0)$, while $S(0)$ is equal to the population size (N) without $E(0)$, $I(0)$ and $R(0)$ (N - $E(0)$ - $I(0)$ - $R(0)$).

7. Run the SEIR model without the particle filtering algorithm to predict the total number of infected individuals until the end of the influenza season. Then, apply the observation model D_t to predict the total number of diagnosed ILI individuals.

8. Through the use of the 'typical shape' of the diagnosed ILI datasets (based on historical data), predict the spread of the remaining time points of the outbreak. Hence, the peak of the diagnosed cases can be forecasted, together with an estimate of the total number of weekly diagnosed ILI cases.

9. Utilise the under-reporting factor rate of the previous year's survey (or other current survey), and estimate the range of the total number of infected individuals in that respective country.

In order to improve the average posterior parameter datasets (for SEIR implementation), the new posterior parameter values for any new seasonal influenza datasets need to be incorporated. Ideally, posterior parameter values are updated on a yearly basis. It is important to keep monitoring all the other related variables on a yearly basis to ensure that any irregularities are captured. Preferably, the seasonal influenza survey needs to be carried out on a yearly basis. This helps to improve the general information of the seasonal influenza, the priors as well as the under-reporting rates.

As discussed in chapter 3, an analysis of different proxies, related to the same outbreak, can improve the understanding of the epidemic. In this chapter, we used all the information available in this thesis to create a model that serves as a good early modelling technique with the predictions being calculated when the influenza epidemic is still at a low starting point (≈ week 50). The above new methodology is an improvement on the methods of real-time forecasting defined in chapter 5.

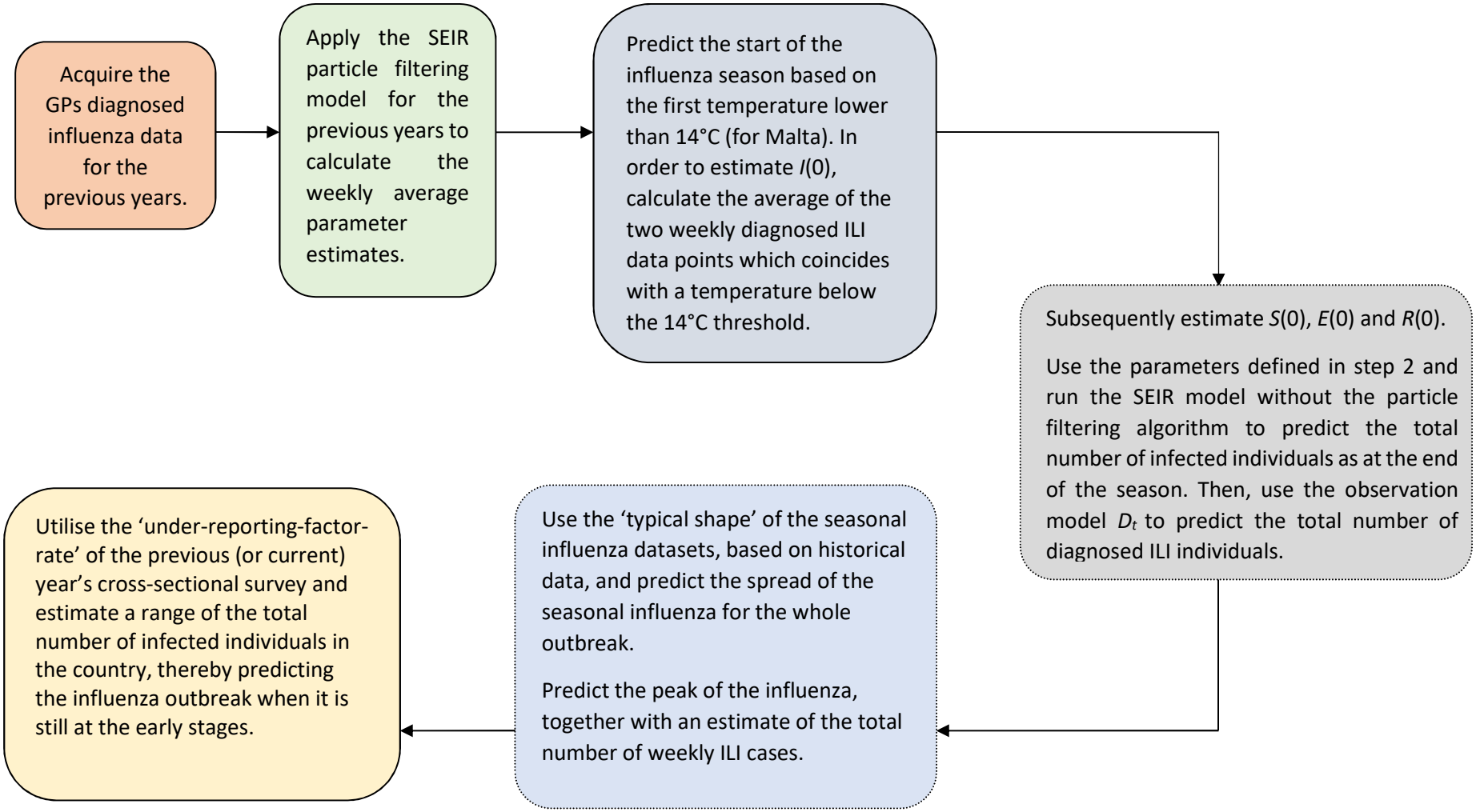


Figure 8.17 - The Prediction Model, incorporating different data sources, historical results, posterior parameters and survey data to obtain an estimate of the number of individuals with seasonal influenza by the end of the season. Furthermore, based on the 'typical shape' of the influenza, the spread of the influenza can also be predicted. This is an early warning modelling technique.

Chapter 9

Conclusions and future work

9.1 Conclusions

The importance of mathematical modelling for the transmission of infectious diseases is becoming more popular across the globe as these serve as an important tool notably to policy makers who desire to control epidemics [170]. After several meetings that I held with health officials in Malta (Appendix A), the need for a reliable mathematical model emerged as a prime objective to help key stakeholders in developing health strategies during the seasonal influenza period. In this thesis, I presented several methods and models aimed at understanding the underlying factors related to the influenza outbreaks. All analysis focused on a principle objective, that is, to predict infectious disease outbreaks based on limited information. At the end of the analysis (Chapter 8), I developed a general framework that incorporates different sources of information to serve as an early warning modelling technique for influenza outbreaks.

In chapter 3, I have shown that for four datasets (consultations, diagnosed, swabbed and positive), collected during the 2009/2010 pandemic period, these have several common features. I have shown that the effective reproduction ratio from different proxies are consistent, although there are some cases where we observe $R_t<1$ from some proxies and $R_t>1$ for others. Even when different sampling rates were considered (daily and weekly), the R_t led to similar results, especially later in the epidemic. However, individual parameter values (infection rate, importation rate and latent period) vary between different proxies. Furthermore, I have shown that there is a general linear relationship between different epidemic proxies and this relationship varies as the epidemic progresses.

In chapter 4, I analysed in detail the relationship between the consultations and diagnosed datasets. When all datasets for different seasons were combined together, a strong linear relationship between consultation and diagnosed variables was observed. This shows that the relationship between the consultation and diagnosed variables is shared for different influenza seasons. Such a finding suggests that, for a new epidemic, this outcome might also hold. This would be an interesting future research to test such a finding for future influenza outbreaks and for other forms of diseases.

Throughout chapter 4, I showed that the consultations dataset can be divided in various groups, thereby establishing a certain level of baseline non-influenza consultations. Furthermore, a portion of the consultations data is related to the diagnosed ILI cases, another portion is related to the false-diagnosed ILI cases, while another portion is related to sub-clinical cases. It was shown that the sub-clinical cases are a substantial part of the consultations data. In fact, these vary between 200 and 14,000 weekly cases for the seasonal influenza datasets, while between 1,000 and 47,000 weekly cases for the pandemic dataset. These cases form part of the uncertainty that exists in epidemiological studies. In fact, such sub-clinical cases might include those individuals who were misdiagnosed (but were actually real influenza cases) (Figure 9.1), those who did not have sufficient influenza symptoms for a diagnosis of ILI (but eventually might have acquired influenza i.e. real influenza case) (Figure 9.1), those who were not reported as a diagnosed ILI by the GPs (but were actually real influenza cases) (Figure 9.1), those who were suspicious of having influenza but never developed the illness, or those considered as non-influenza related cases. Hence, some of the sub-clinical cases might be real influenza cases (Figure 9.1).

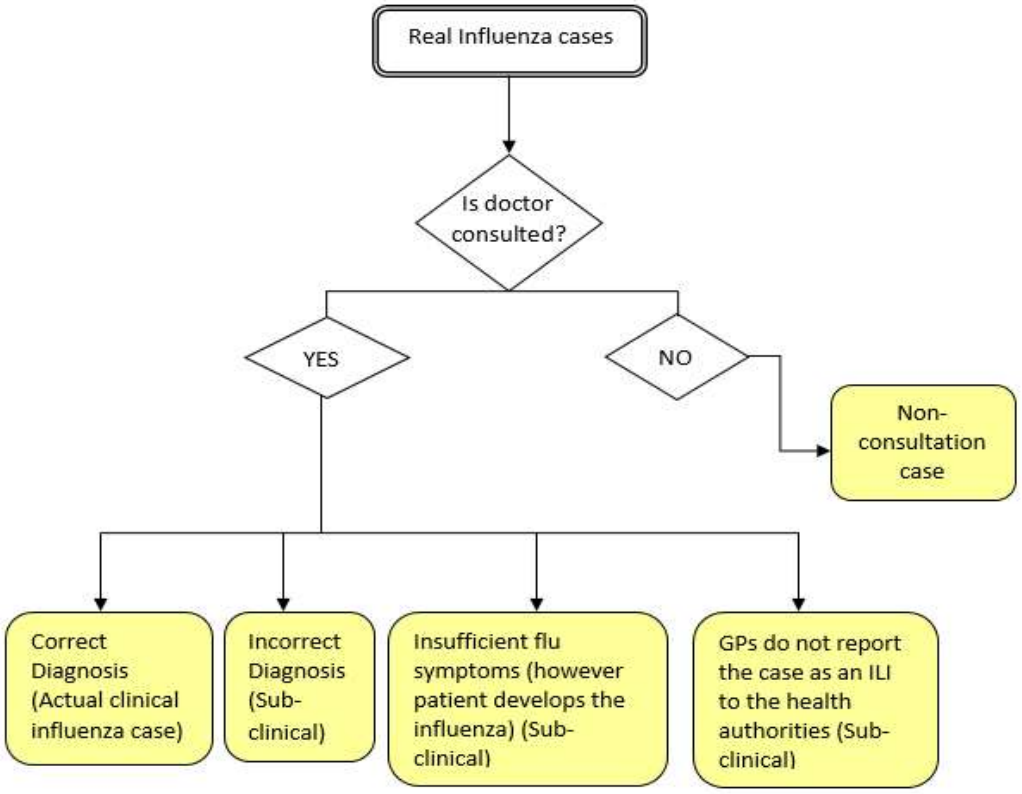


Figure 9.1 – Different pathways related to the real influenza cases. This can be divided in two main groups: individuals consulting the doctors and individuals not consulting their GP. Those consulting their doctor can be divided into four groups: GPs performing correct diagnosis, GPs performing a wrong diagnosis, patients who do not have enough influenza symptoms but actually have the influenza, and GPs not reporting influenza cases to the health authorities.

The datasets presented in this study allow us to analyse these results in further detail. In fact, when the datasets registered lower values of consultations (2013/2014 and 2014/2015), they provided a weak correlation between the diagnosed and the consultation cases. In relation to this, we showed that a ratio, higher than 4% between diagnosed and consultations, provided an R^2 value higher than 0.5, and moderate/strong correlation values. Otherwise, other ratios less than 4% provided a weak R^2 value (< 0.32). Furthermore, in almost all cases, such low ratio values provided weak correlation values between consultations and diagnosed variables. However, there is an additional time-dependent factor which was discussed in detail throughout this thesis. In fact, we showed that the relationship between consultations and diagnosed variables is stronger during the mid-part of the influenza season. This can be attributed to a higher number of consultations and diagnosed ILI cases.

The time-dependent factor was analysed in further detail when the SEIR model was extended to a joint model. The latter model allowed the consultations data to be modelled through the linear relationship between the diagnosed and consultation datasets. In contrast to the general linear regression model, the joint model allowed the parameters of the linear regression model to be updated at every time point. Although for some datasets the general linear regression model provided a good fit for the consultations dataset (dependent variable), the time-dependent linear regression model improved the consultation predictions. Such improvement was shown to be better when weak association existed between the consultations and diagnosed datasets (2013/2014 and 2014/2015 datasets).

The Bayesian framework was applied to all influenza datasets, except for the latest dataset (2015/2016). We showed that the SEIR model accurately fit all four seasonal influenza datasets, hence allowing us to examine the posterior parameter values in further detail. Through the Bayesian framework, we showed that the posterior infection rate is associated with the total number of diagnosed cases throughout the season, where the higher the number of cases throughout the season, the higher the infection rate parameter. Furthermore, some posterior parameter estimates for the diagnosed datasets are consistent across datasets. The latent period, background rate and the reporting rate are broadly consistent across the four seasonal influenza datasets. In contrast, the consultation posterior parameter values vary widely across different years. Furthermore, the

consultation posteriors are not consistent with the diagnosed posteriors. These results conclude that posterior parameters for different proxies need to be treated separately for each different seasonal influenza proxy. However, since the diagnosed ILI cases are a more direct signal of the seasonal influenza (when compared to the consultations dataset), such parameters can be used as a source of information for future influenza outbreaks.

The most important epidemiological parameter is the effective reproduction ratio, R_t. We showed that the R_t values are consistent for different diagnosed ILI datasets. In chapter 4, we showed that although consistency exists in the R_t values for different consultation datasets, these are under the value of 1 during the peak of the influenza. This is in contrast to the diagnosed datasets; hence, the R_t values through the consultation datasets are not providing a signal of an epidemic. For the diagnosed datasets, the R_t values provide a good quality signal when the seasonal influenza actually reaches its first peak value. For example, for the diagnosed 2013/2014 data, the R_t values (Figure 9.2) reached their peak during the month of February, which corresponds to the first data point that represents the peak of the diagnosed ILI cases. Subsequently, the seasonal influenza persisted for a couple of weeks following this first data point, while the R_t values start decreasing soon after the peak of the reproduction ratio. This fact can be observed for all the other three datasets. Furthermore, the real signal is when the R_t approaches or exceeds the value of 2 (ignoring the initial values of R_t), since this corresponds to the first sharpest growth in the number of diagnosed ILI cases. Subsequently, when the influenza reaches the peak, the R_t value declines soon after to a value close or under 1.

As mentioned above, the relationship between consultations and diagnosed datasets was further studied through a joint model. In chapter 5, we showed, that this model improved some real-time forecasts of the consultations dataset, particularly for the higher number of consultation and diagnosed cases (2009/2010 dataset). Although the joint-model technique provided further insight into some improved real-time forecasts, we conclude that the real-time forecasts are not always consistent. In general, the real-time forecasting provides some accurate future predictions; however an earlier signal is desired to try to mitigate the impact of the seasonal influenza on the entire population.

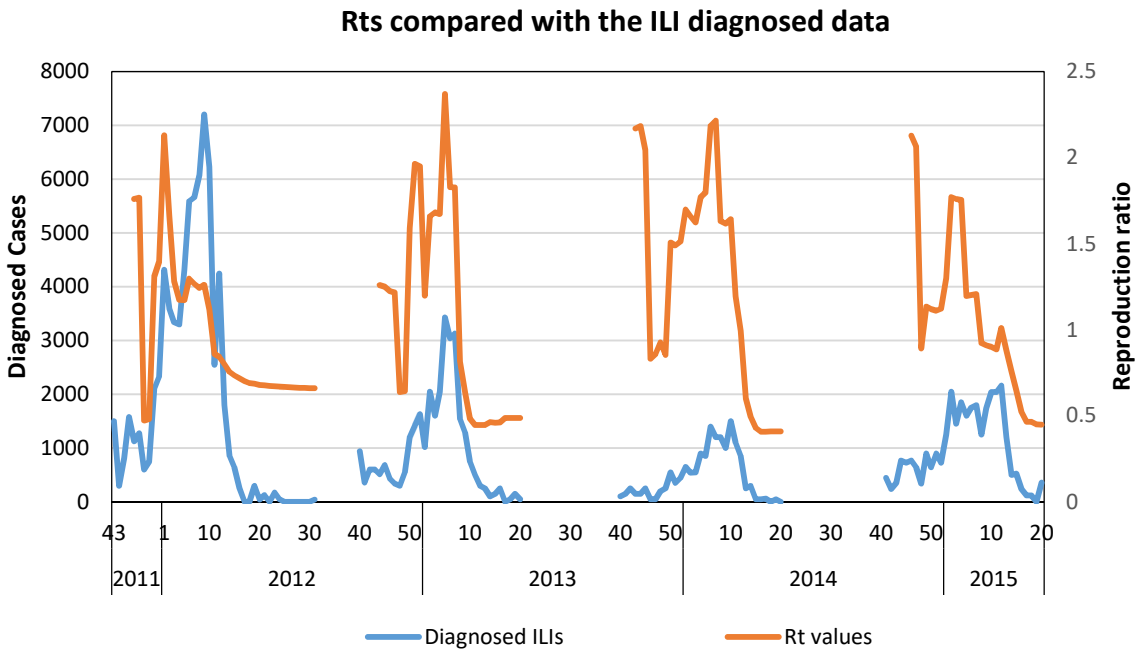


Figure 9.2 – Relationship between the diagnosed ILI cases and the reproduction ratio for the four seasonal influenza datasets. The x-axis represents the week number, while the left y-axis represents the number of diagnosed cases and the right y-axis represents the reproduction number. The blue line corresponds to the GPs diagnosed ILI datasets and the orange line is the effective reproduction ratio obtained from the SEIR models. For this figure, the diagnosed ILI datasets are being taken into consideration as these correspond to a direct and more reliable proxy of the influenza season. Note that the initial 2 to 5 reproduction ratio values (depending on the dataset) were ignored, as described in chapter 6.

As shown in chapters 3 and 4, the reproduction ratio commences with a high peak value, which in reality is probably unrealistic. In chapter 6, I showed that the initial R_t values are sensitive to the choice of the initial values of the infected (I), susceptible (S), exposed (E) and removed (R) individuals. By increasing the values of $R(0)$, the initial R_t values decreased (for both consultations and diagnosed), while for the diagnosed data, the higher the values of $I(0)$ (and $E(0)$), the higher the initial R_t values. On the contrary, for the consultations data, a higher value of $I(0)$ (and $E(0)$) does not influence the initial R_t values. In most cases, the exclusion of the initial two to four points was enough to eliminate most of the R_t inconsistencies as the effective reproduction ratio becomes stable. For high values of $R(0)$, $I(0)$ and $E(0)$, some inconsistencies amongst different R_t plots are more likely to be observed. I also provided a summary of the method to avoid these high initial values of R_t.

The uncertainty in such studies is not limited to the sub-clinical cases, but there are a substantial number of individuals with seasonal influenza symptoms who do not consult

their doctor. Therefore, a high level of missing data exists, and hence this requires special attention in order to understand any underlying factors that might provide further insight to disease outbreaks. To some extent, such missing data can be revealed through the use of surveys. In chapter 7, two scientific cross-sectional surveys carried out amongst the Maltese population, provided further information to understand core factors related to the seasonal influenza in Malta as well as the most common influenza symptoms amongst the Maltese citizens. Furthermore, respondents provided information on the duration of the influenza, hospitalisation due to the influenza, GP consultations and other medical factors.

According to the representative sample of the Maltese population, 30% of the total respondents claimed that they had the seasonal influenza during 2014/2015, while around 85% claimed they had at least one symptom related to the seasonal influenza during the same season. These two percentages are substantially different; hence it is very likely that respondents have different perceptions of the definition of the seasonal influenza. From the survey results, it is clear that a substantial proportion of the Maltese population did not visit their GP due to the seasonal influenza, and opted to self-examine their ILI symptoms which resulted in a 'self-diagnosis'. Based on several assumptions and the survey results (2015 survey), it was shown that the reporting rate might vary between 13.9% and 17.8% when immunity from seasonal influenza is not assumed, while the reporting rate varies between 20.6% and 26.7% when assuming immunity of individuals from seasonal influenza. Furthermore, we estimated that for the 2014/2015 influenza season in Malta, between 120,000 and 150,000 had the seasonal influenza. Thus, this concludes that between 28% and 36% of the Maltese citizens had seasonal influenza during the 2014/2015 period.

I also compared the diagnosed data collected by GPs and the survey results. The monthly occurrences of the seasonal influenza cases, as stated by the survey respondents, were compared against the monthly diagnosed ILI cases (GP data). The distribution of the months for both datasets was shown to be similar, with peaks occurring at the same time points. The same result was obtained when the monthly occurrences of the symptoms related to the seasonal influenza, as stated by the survey respondents, was compared with the GPs diagnosed monthly ILI cases. In general, the results for different questions between the two surveys (which were carried out during different periods) are in

agreement and there is no effect on the time at which surveys are conducted. Hence, such results were considered to be consistent for the two consecutive cross-sectional surveys.

The above information can be used to understand specific features of the seasonal influenza; however, the potential of such information can be maximized if a holistic framework is considered. In chapter 8, I defined an innovative methodology of how to incorporate most of the above information into one single framework. Furthermore, in chapter 8, it was shown that the temperature data is an important factor for understanding further the seasonal influenza in Malta. In relation to this, a moderate negative correlation between the diagnosed ILI datasets and the temperature data was established. The findings show that the lower the temperature, the higher the diagnosed ILI cases. Furthermore, when the temperature drops below the 14°C threshold, this triggers the first substantial increase in the number of diagnosed ILI cases (the 'start' of the seasonal influenza epidemic). In general, the first drop in temperature below the 14°C threshold coincides with a sudden and largest drop (≈3°C drop) in temperature throughout the entire influenza season. In the results presented in chapter 8, the number of diagnosed ILI cases remained consistently high for a period of 13 weeks. This is similar to the number of weekly temperatures (13 weeks) below the 14°C threshold during the seasonal influenza period. Furthermore it was established that, when the temperature drops below 14°C, the number of diagnosed ILI cases at that point predicts the strength of the influenza for the entire season. In fact, it was shown that the higher the number of diagnosed cases (when the temperature drops below 14°C for the first time), the higher the number of the total diagnosed ILI cases for the entire season.

The R_t values were compared to the temperature data to understand how the effective reproduction ratio is dependent on the temperature data. During the period when the temperature is lower than 14°C, the R_t values in general are greater than 1.

Through the new developed framework, an accurate estimate of the number of diagnosed ILI individuals was established for each individual season. In fact, the total number of forecasted diagnosed individuals varied by ±8% when compared with the total reported diagnosed individuals (GP data). This framework was tested for the 2015/2016 (since no posterior parameters related to this dataset were used), and the total number of diagnosed cases till the end of the influenza season was predicted with a precision of 92% when

compared with the actual data. Such an estimate was obtained during the early stages of the influenza season (11^{th} week of the epidemic), and 5 months before the end of the influenza season. A typical shape of the diagnosed ILI cases for 2015/2016 season was established (through the use of previous seasonal influenza datasets). Therefore, the new developed model (Chapter 8) provided improved real-time forecasts when compared to the real-time forecasts of chapter 5. We showed that the new framework is able to forecast the spread of influenza in Malta, its peak and the number of diagnosed cases at a very early stage in the outbreak.

Throughout chapter 8, the temperature data and various other results of the previous chapters were incorporated to optimize all available information. In fact, the posterior parameter values of the previous chapters were incorporated in the new developed model. In relation to this, sensitivity analysis (chapter 6) was used to refine the prior parameters and to improve the SEIR reproduction ratio. The SEIR model (without the particle filter algorithm) together with the observation model D_t was then used to predict the number of diagnosed cases. Subsequently, the survey results were used to estimate the number of individuals that had acquired the seasonal influenza in Malta.

9.2 Implications for practice

Through the above conclusions, one can present several questions about future influenza outbreaks. Since we considered the 2015/2016 influenza season as the test example (in Chapter 8) where we tested the new developed model, we can now provide answers to the following questions based on the survey results (Chapter 7) and the new model (Chapter 8). Additionally, the use of the linear regression model throughout the next paragraphs shall be of value to answer the forthcoming questions. This information could be useful to policy-makers and hospital management.

1. How many people will be diagnosed at each time point?

I used the model described in chapter 8 to provide the number of forecasted weekly data for the 2015/2016 influenza season (Figure 9.3). Data points before week 51 were taken as known; hence forecasts were calculated from this point onwards. The peak is forecasted to be reached during week 6 (January 2016) with approximately 2,014

diagnosed ILI individuals by the GPs. In week 13 (end March 2016), it can be assumed that the seasonal influenza is no longer a national concern for this season, as the number of diagnosed cases are at a sharp decline. I therefore estimate that, by the end of the influenza season, around 27,000 individuals would have been diagnosed with ILI by the Maltese GPs.

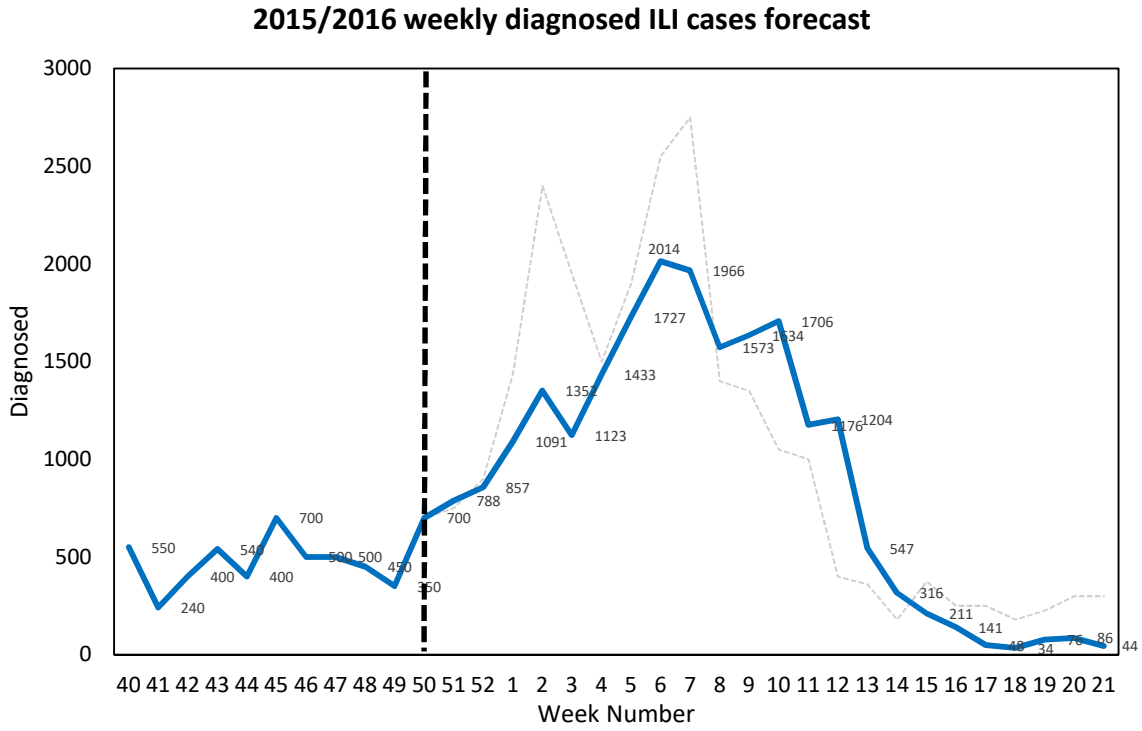


Figure 9.3 – The number of weekly forecasted diagnosed ILI cases during the 2015/2016 season. The vertical dashed line is the last known data point that was taken into consideration for model application. The dotted grey line represents the actual reported diagnosed cases.

2. How many people will visit a doctor per week?

Figure 9.4 shows the number of consultations per week. Based on the data up to week 50, I used the same method as in chapter 4 to establish κ and Δ for the 2015/2016 dataset. Then, by using the forecasted diagnosed ILI cases, the number of weekly consultations can be calculated following week 50. Similar to the above, before week 51, the data points were taken as known. The highest number of consultation cases is being estimated to be reached during week 6, with a total number of 34,395 consultation cases. I therefore estimate that the average number of weekly consultations will be around 24,000 consultations per week. These consultations include non-influenza related cases, influenza-related cases and other cases which can be considered as sub-clinical. Hence,

the sub-clinical cases might include the misdiagnosed ILI cases, individuals with insufficient influenza symptoms to be diagnosed with ILI, GPs not reporting the case as an ILI to health authorities, and non-influenza related cases.

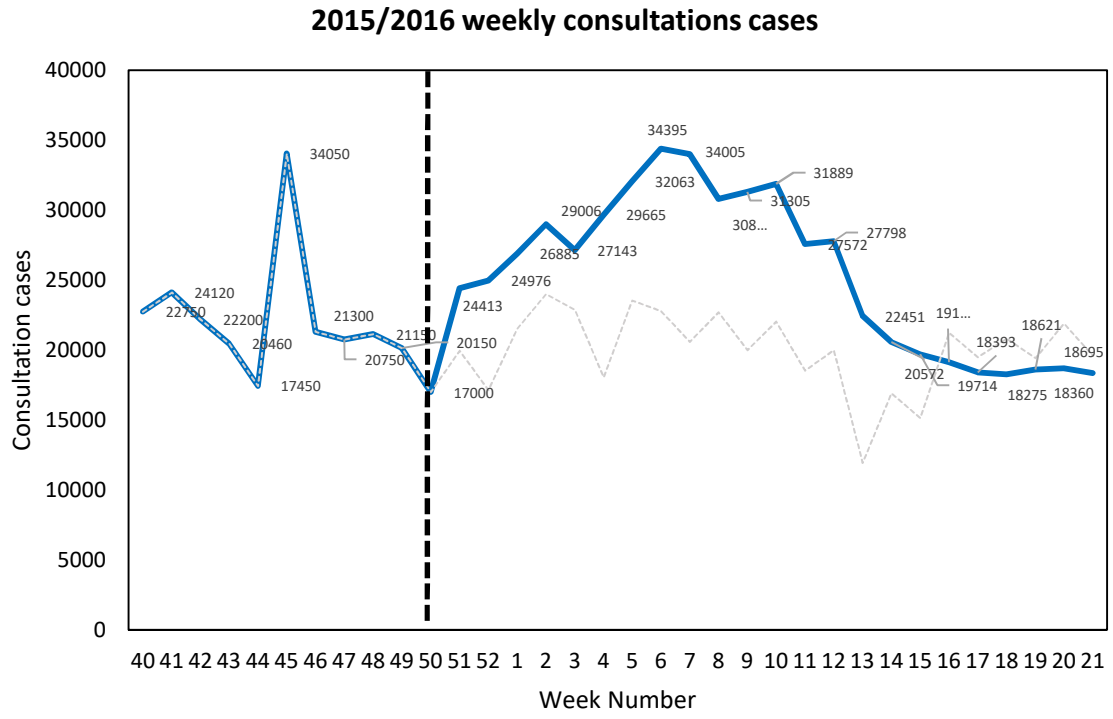


Figure 9.4 – The number of weekly forecasted consultations cases during the 2015/2016 season. The vertical dashed line is the last known data point that was taken into consideration for model application. The dotted grey line represents the actual reported consultation cases.

3. How many people will form part of the sub-clinical cases per week?

Figure 9.5 shows the number of weekly sub-clinical cases. All data points were estimated through the general linear regression model. Similar to the occurrences in the previous questions, the peak of the number of sub-clinical cases (n = 14,382) is forecasted to be reached at week 6.

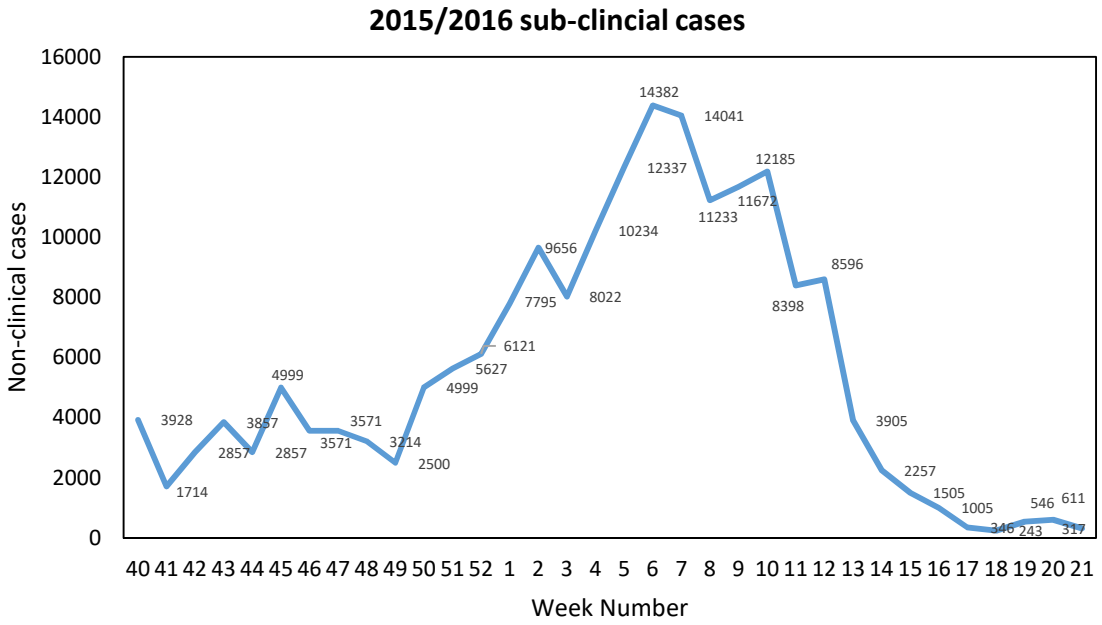


Figure 9.5 – The number of weekly forecasted sub-clinical cases during the 2015/2016 season. Note that unlike figures 9.3 and 9.4, no data is available for comparison.

4. How many people are likely to acquire the seasonal influenza in Malta per week?

Based on the 2014/2015 survey results and the GPs diagnosed data, the reporting rate for the seasonal influenza was between 20.6% and 26.7% (assuming that a person acquires immunity from seasonal influenza after diagnosis and recovers from seasonal influenza). Hence, by taking an average of these two percentages, this gives a reporting rate of 23.7%. Thus, based on this estimate and the forecasted diagnosed cases, the following weekly seasonal influenza cases in Malta can be predicted (Figure 9.6). In total, around 110,000 Maltese individuals are being estimated to acquire the seasonal influenza during the 2014/2015 season, contributing to around 26% of the Maltese population. Furthermore, on average, it is being estimated that there shall be around 3,300 of seasonal influenza cases per week. The peak is predicted to be reached in week 6, and is expected to register around 8,500 seasonal influenza cases during the latter week.

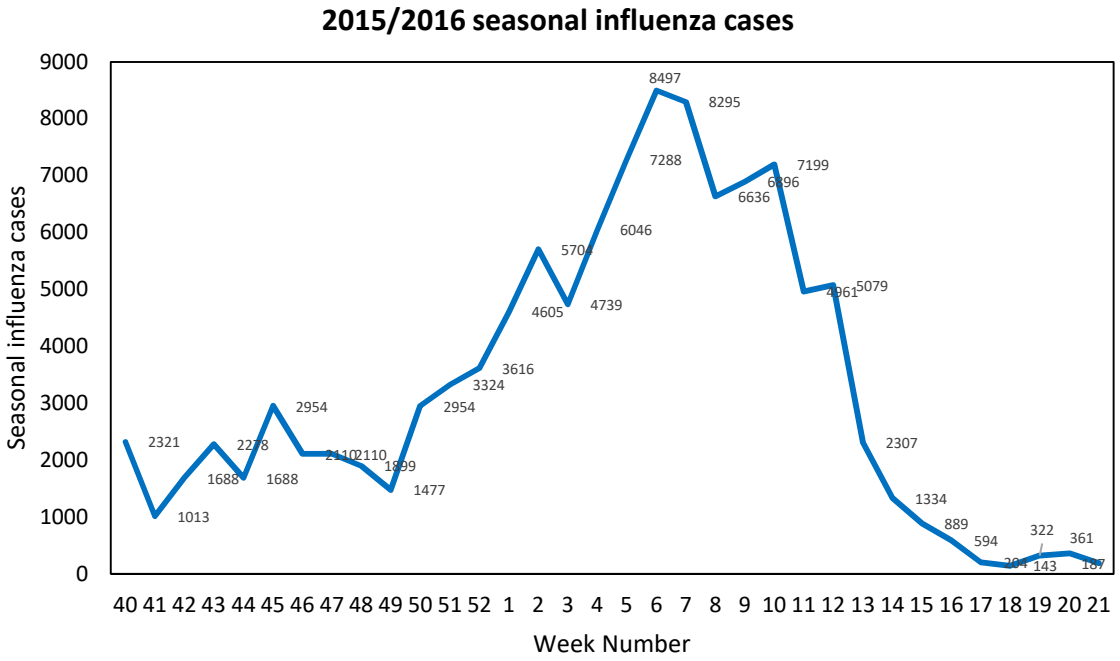


Figure 9.6 – The number of weekly forecasted seasonal influenza cases during the 2015/2016 season. Note that unlike figures 9.3 and 9.4, no data is available for comparison.

5. How many people are likely to visit the hospital per week?

One of the main concerns in the health sector is the demand on the national hospital. According to the 2014/2015 survey, 20% of the seasonal influenza patients visited the hospital due to their influenza. Hence, based on these results and the above estimates of the number of seasonal influenza individuals, it is being predicted that the number of individuals that will visit the hospital due to the seasonal influenza is around 22,500 during the entire season, with an average of 660 individuals per week.

6. How many people are likely to experience the seasonal influenza symptom diarrhoea?

According to the 2014/2015 survey, 18% of the individuals experienced diarrhoea during the influenza season. This is one of the symptoms that can be attributed to the seasonal influenza. In fact, respondents associated this symptom only during the months related to the influenza season. Based on the Maltese population, I therefore estimate that around 77,000 individuals will experience diarrhoea during the 2015/2016 influenza season.

7. How many people are likely to have the seasonal influenza symptom chest pain?

According to the 2014/2015 survey, 2.58% citizens reported chest pain during the influenza season. Hence, based on this estimate, we can say that around 11,000 individuals will experience chest pain during the 2015/2016 influenza season.

We can estimate similar figures for other symptoms. Although the above results are ballpark figures, one can use these as indicative dynamics for future outbreaks.

9.3 Future work

The above results and conclusions unfold a number of opportunities for future research. Although this thesis covered several important factors regarding epidemiological modelling, further work is warranted to understand the extent to which such results can be utilised to forecast other different types of outbreaks in other populations. Furthermore, additional epidemiological work needs to be carried out to incorporate the above results into one holistic Bayesian framework.

The below recommendations can be considered as limitations to this study and suggestions for future work.

1. The 'joint model' developed in chapter 4 was designed through a time-dependent linear regression model. Parameters were updated using a standard technique by updating up to a given time point. For future work, these parameters can be incorporated into the Bayesian framework, by updating these parameters through the use of the particle filter algorithm.

2. The SEIR model can be improved by incorporating other parameters, for instance by capturing the loss of immunity rate and the rate of GPs influenza misdiagnosis.

3. The priors of the SEIR model require further exploration. Initial parameter values have substantial impact on the predictions of outbreaks. The survey results can be used to construct improved informative priors to the SEIR model.

4. Future work can be carried out on the use of other particle filtering techniques to examine whether the posterior parameters can be improved.

5. Running a series of cross-sectional surveys during various stages of the seasonal influenza outbreak might aid to further understand people's perceptions of influenza, and probe deeper into whether the survey results are time-dependent.

6. Further research is required to analyse the seasonal influenza datasets against other variables. Such datasets could include serological data and hospitalisations data due to influenza.

7. All of the aforementioned suggested work requires further exploration in the context of diverse populations, countries and cultural backgrounds, other temporal scales and diverse epidemics.

8. Throughout this study, the relationship between the consultation and diagnosed cases was found collectively for different seasons. This relationship suggests that, for a new epidemic, this finding might also hold. Future research could compare such datasets for other influenza seasons and other different types of epidemics.

9. Survey findings can be tested further using other observed datasets to examine their validity in the context of epidemiological studies.

10. Further testing of the new model defined in chapter 8 is warranted in order to test its application in other countries.

9.4 Final conclusions

This thesis has presented, for the first time, a combination of novel data sources to predict influenza outbreaks. The findings were generated through the rigorous application of epidemiological modelling. My study quantifies the national impact of the influenza and underlines the power of national analysis bolstered by mathematical modelling, and the impact of several factors in predicting the outbreak. My findings also clarify other direct or indirect aspects related to the dynamics of seasonal influenza. Through the application

of nationwide cross-sectional surveys, the under-reporting rate of the seasonal influenza was innovatively established and other underlying factors related to the seasonal influenza were explored. Moreover, I showed that the temperature data triggers the real start of the influenza epidemic. Based on the thesis findings, a novel epidemiological model and framework were established, providing accurate real-time forecasts with a clear early warning signal to the influenza outbreak. Thus, although initially we were presented with limited information to predict the outbreak, throughout the thesis we established fundamental factors to accurately predict the epidemic. The above information additionally extends and adds to the existing understanding of the seasonal influenza epidemic internationally, and is extensively innovative in the Maltese context. It is hoped that the findings presented in this thesis will be useful to policy-makers and health authorities to plan better public health strategies and interventions in order to control epidemics.

References

[1] WHO, World Health Organization, Avian influenza: assessing the pandemic threat, 2005 URL: http://www.who.int/iris/handle/10665/68985.

[2] R. Anderson, R. May R, Infectious Diseases of Humans, first ed., Oxford University Press, Oxford, 1991.

[3] J.A.P. Heesterbeek, M.G. Roberts, How mathematical epidemiology became a field of biology: a commentary on Anderson and May (1981) 'The population dynamics of microparasites and their invertebrate hosts', Phil. Trans. R. Soc. Lond. B, **370**(1666) (2015) 20140307 DOI: 10.1098/rstb.2014.0307.

[4] P. Yan, H. Chen, D. Zeng, Syndromic Surveillance Systems, Annual Review of Information Science and Technology, **42** (2008) 425-495.

[5] J.S. Nguyen-Van-Tama, A.W. Hampson, The epidemiology and clinical impact of pandemic influenza, Vaccine, **21** (2003) 1762–1768.

[6] T. Jefferson, C. Di Pietrantonj, L.A. Al-Ansary, E. Ferroni, S. Thorning, R.E. Thomas, Vaccines for preventing influenza in the elderly, Cochrane Database of Systematic Reviews, **2** (2010) CD004876. DOI: 10.1002/14651858.CD004876.pub3.

[7] C.R. Simpson, N. Lone, J. McMenamin, R. Gunson, C. Robertson, L.D. Ritchie, A. Sheikh, Early estimation of pandemic influenza Antiviral and Vaccine Effectiveness (EAVE): use of a unique community and laboratory national data-linked cohort study, Health Technology Assessment, **19**(79) (2015) DOI: 10.3310/hta19790.

[8] N.I. Lone, C. Simpson, K. Kavanagh, C. Robertson, J. McMenamin, L. Ritchie, A. Sheikh, Seasonal Influenza Vaccine Effectiveness in the community (SIVE): protocol for a cohort study exploiting a unique national linked data set, BMJ Open, **2** (2012) e001019. DOI: 10.1136/bmjopen-2012-001019.

[9] P. Cassar, Medical History of Malta, Welcome Historical Medical Library, London, 1965.

[10] J.F. Bishop, M.P. Murnane, R. Owen, Australia's winter with the 2009 pandemic influenza A (H1N1) virus, The New England Journal of Medicine, **361** (2009) 2591–2594. DOI: 10.1056/NEJMp0910445.

[11] J.K. Louie, M. Acosta, K. Winter, C. Jean, S. Gavali, R. Schechter, D. Vugia, K. Harriman, B. Matyas, C.A. Glaser, M.C. Samuel, J. Rosenberg, J. Talarico, D. Hatch, California Pandemic (H1N1) Working Group, Factors associated with death or hospitalization due to pandemic 2009 influenza A(H1N1) infection in California. JAMA **302**(17) (2009) 1896–1902. DOI: 10.1001/jama.2009.1583.

[12] M.I. Meltzer, N.J. Cox, K. Fukuda, The economic impact of pandemic influenza in the United States: priorities for intervention, Emerging Infectious Diseases, **5**(5) (1999) 659–671.

[13] L. Temime, G. Hejblum, M. Setbon, A.J. Valleron, The rising impact of mathematical modelling in epidemiology: antibiotic resistance research as a case study, Epidemiology and Infection, **136**(3) (2008) 289-298 DOI: 10.1017/S0950268807009442.

[14] M.J. Keeling, L. Danon, Mathematical modelling of infectious diseases, British Medical Bulletin, **92**(1) (2009) 33-42 DOI: 10.1093/bmb/ldp038.

[15] M. Woolhouse, How to make predictions about future infectious disease risks. Philos Trans R Soc Lond B Biol Sci., **366**(1573) (2011) 2045–2054 DOI: 10.1098/rstb.2010.0387.

[16] National Statistics Office, Malta (NSO), 'Demographic Review 2013', 2015, URL: https://nso.gov.mt/en/publicatons/Publications_by_Unit/Documents/C3_Population_and_Tourism_Statistics/Demographic_Review_2013.pdf.

[17] National Statistics Office, Labour force Survey Q2, Malta, 2015 URL: www.nso.gov.mt

[18] C.S. Ventura, Past Influenza Pandemics and their Effect in Malta, Malta Medical Journey, **17**(3) (2005) 16-19.

[19] C.I. Siettos, L. Russo, Mathematical modeling of infectious disease dynamics, Virulence, **4**(4) (2013) 295–306 DOI: 10.4161/viru.24041.

[20] I. Nasell, Stochastic models of some endemic infections, Mathematical Biosciences, **179** (2002) 1–19.

[21] L.J.S. Allen, A.M. Burgin, Comparison of deterministic and stochastic SIS and SIR models in discrete time, Mathematical Biosciences, **163** (2000) 1-33.

[22] A. Lahrouz, L. Omari, A. Settati, A. Belmaati, Comparison of deterministic and stochastic SIRS epidemic model with saturating incidence and immigration, Arabian Journal of Mathematics, **4**(2) (2014) 101-116 DOI 10.1007/s40065-014-0119-0.

[23] M. Imran, M. Hassan, M. Dur-E-Ahmad, A. Khan, A comparison of a deterministic and stochastic model for Hepatitis C with an isolation stage, Journal of Biological Dynamics, **7**(1) (2013) 276-301, DOI: 10.1080/17513758.2013.859856.

[24] T. Britton, Stochastic epidemic models: a survey. Stockholm University, 2009 URL: https://arxiv.org/pdf/0910.4443.pdf.

[25] D. Clancy, P.D. O'Neill, Bayesian estimation of the basic reproduction number in stochastic epidemic models, International Society for Bayesian Analysis, **3** (2008) 737-758.

[26] J.B.S. Ong, M.I.C. Chen, A.R. Cook, H.C. Lee, V.J. Lee, R.T.P. Lin, P.A. Tambyah, L.G. Goh, 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**(4) (2010) DOI: 10.1371/journal.pone.0010036.

[27] G. Streftaris, G.J. Gibson, Bayesian inference for stochastic epidemics in closed populations, Computational Statistics and Data Analysis, Statistical Modelling, **4**(1) 2004 63–75.

[28] P.D. O'Neill, N.G. Becker, Inference for an epidemic when susceptibility varies, Biostatistics, **2**(1) (2001) 99-108.

[29] T.J. McKinley, J.V. Ross, R. Deardon, A.R. Cook, Simulation-based Bayesian inference for epidemic models, Computational Statistics and Data Analysis, **71**(SI) (2014) 434–447.

[30] M.S.Y. Lau, G. Marion, G. Streftaris, G.J. Gibson1, New model diagnostics for spatiotemporal systems in epidemiology and ecology, Journal of the Royal Society Interface, **11**(93) (2014) DOI: 10.1098/rsif.2013.1093.

[31] G.J. Gibson, E. Renshaw, Estimating parameters in stochastic compartmental models using Markov chain methods, IMA Journal of Mathematics Applied in Medicine & Biology, **15**(1) (1998) 19-40.

[32] C. Brun, A.R. Cook, J.S.H. Lee, S.A. Wich, L.P. Koh, L.R. Carrasco, Analysis of deforestation and protected area effectiveness in Indonesia: A comparison of Bayesian spatial models, Global Environmental Change, **31** (2015) 285–295.

[33] F.M. Neri, A.R. Cook, G. J. Gibson, T.R. Gottwald, C.A. Gilligan, Bayesian Analysis for Inference of an Emerging Epidemic: Citrus Canker in Urban Landscapes, PLoS Computational Biology, **10**(4) (2014) DOI: 10.1371/journal.pcbi.1003587.

[34] F. Xiang, P. Neal, Efficient MCMC for temporal epidemics via parameter reduction, Computational Statistics and Data Analysis, **80** (2014) 240–250.

[35] M. Parry, G. J. Gibson, S. Parnell, T. R. Gottwald, M. S. Irey, T. C. Gast, C.A. Gilligan, Bayesian inference for an emerging arboreal epidemic in the presence of control, Proceeding of the National Academy of Science of the United States of America, **111**(17) 2014 6258-6262.

[36] A. Skvortsov, B. Ristic, Monitoring and prediction of an epidemic outbreak using syndromic observations, Mathematical Biosciences, **240**(1) (2012) 12–19.

[37] D.M. Sheinson, J. Niemi, W. Meiring, Comparison of the performance of particle filter algorithms applied to tracking of a disease epidemic, Mathematical Biosciences, **255** (2014) 21–32.

[38] W. Yang, B.J. Cowling, E.H.Y. Lau, J. Shaman, Forecasting Influenza Epidemics in Hong Kong, PLoS Computational Biology, **11**(7) (2015) DOI: 10.1371/journal.pcbi.1004383.

[39] C. Andrieu, A. Doucet, R. Holenstein, Particle Markov chain Monte Carlo methods, Journal of the Royal Statistical Society. Series B (Statistical Methodology), **72**(3) (2010) 269-342.

[40] W. Yang, A. Karspeck, J. Shaman, Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics, PLoS Computational Biology, **10**(4) (2014) DOI:10.1371/journal.pcbi.1003583.

[41] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking, IEEE Transactions on signal processing, **50**(2) 2002 174-188.

[42] E.L. Ionides, A. Bhadra, Y. Atchade, A. King, Iterated filtering, Annals of Statistics, **39**(3) (2011) 1776-1802.

[43] G. Evensen, Sequential data assimilation. Data Assimilation: The ensemble Kalman filter, 2nd edition, New York: Springer, (2008) 27–46.

[44] A.R. Karspeck, J.L. Anderson, Experimental Implementation of an Ensemble Adjustment Filter for an Intermediate ENSO Model, Journal of Climate, **20**(18) (2007) 4638–4658.

[45] J.L. Anderson, A Non-Gaussian Ensemble Filter Update for Data Assimilation, Monthly Weather Review, **138**(11) (2010) 4186-4198.

[46] A. Doucet, S. Godsill, C. Andrieu, On sequential Monte Carlo sampling methods for Bayesian filtering, Statistics and Computing, **10**(3) (2000) 197–208.

[47] J.P. Chretien, D. George, J. Shaman, R.A. Chitale, F.E. McKenzie, Influenza Forecasting in Human Populations: A Scoping Review, PLoS ONE, **9**(4) (2014) DOI: 10.1371/journal.pone.0094130.

[48] G.J. Gibson, E. Renshaw, Likelihood estimation for stochastic compartmental models using Markov chain methods, Statistics and Computing, **11**(4) (2001) 347–358.

[49] M.M. Saito, S. Imoto, R. Yamaguchi, H. Sato, H. Nakada, M. Kami, S. Miyano, T. Higuchi, Extension and verification of the SEIR model on the 2009 influenza A (H1N1) pandemic in Japan, Mathematical Biosciences, **246**(1) (2013) 47–54.

[50] A. Ibeas, M. de la Sen, S. Alonso-Quesada, I. Zamani, Stability analysis and observer design for discrete-time SEIR epidemic models, Advances in Difference Equations, **122** (2015) DOI: 10.1186/s13662-015-0459-x.

[51] L. Liu, J.Wang, X. Liu, Global stability of an SEIR epidemic model with age-dependent latency and relapse, Nonlinear Analysis: Real World Applications, **24** (2015) 18–35.

[52] J.R. Artalejo, A. Economou, M.J. Lopez-Herrero, The stochastic SEIR model before extinction: Computational approaches, Applied Mathematics and Computation, **265** (2015) 1026-1043.

[53] E.V. Grigorieva, E.N. Khailov, Optimal Intervention Strategies for a SEIR Control Model of Ebola Epidemics, Mathematics, **3**(4) (2015) 961-983.

[54] H. Maurer, M. D. De Pinho, Optimal Control of Epidemiological SEIR models with L1-Objectives and Control-State Constraints, Pacific Journal of Optimization, **12**(2) (2016) 415-436.

[55] P. Lai, C.B. Chow, H.T. Wong, K.H. Kwong, Y.W. Kwan, S.H. Liu, W.K. Tong, W.K. Cheung, W.L. Wong, An early warning system for detecting H1N1 disease outbreak – a spatio-temporal approach, International Journal of Geographical Information Science, **29**(7) (2015) 1251–1268.

[56] K.C. Chong, X. Wang, S. Liu, J. Cai, X. Su, B.C. Zee, G. Tam, M.H. Wang, E. Chen, Interpreting the transmissibility of the avian influenza A(H7N9) infection from 2013 to 2015 in Zhejiang Province, China, Epidemiology and Infection, **144**(8) (2016) 1584–1591.

[57] G. Chowell, H. Nishiura, L.M.A. Bettencourt, Comparative estimation of the reproduction number for pandemic influenza from daily case notification data, J. R. Soc. Interface, **4** (2006) 155-166.

[58] G. Chowell, C. Viboud, C.V. Munayco, J. Gomez, L. Simonsen, M.A. Miller, J. Tamerius, V. Fiestas, E.S. Halsey, C.A. Laguna-Torres, Spatial and Temporal Characteristics of the 2009 A/H1N1 Influenza Pandemic in Peru, PLoS ONE **6**(6) (2011). DOI: 10.1371/journal.pone.0021287.

[59] J.T. Griffin, T. Garske, A.C. Ghani, Joint estimation of the basic reproduction number and generation time parameters for infectious disease outbreaks, Biostatistics, **12**(2) (2011) 303-312.

[60] P.Y. Boëlle, P. Bernillon, J.C. Desenclos, A preliminary estimation of the reproduction ratio for new influenza A(H1N1) from the outbreak in Mexico, March–April 2009, Euro surveillance, **14**(19) (2009) pii:19205.

[61] C. Fraser, C.A. Donnelly, S. Cauchemez, W.P. Hanage, M.D. Van Kerkhove, T.D. Hollingsworth, J. Griffin, R.F. Baggaley, H.E. Jenkins, E. J. Lyons, T. Jombart, W.R. Hinsley, N.C. Grassly, F. Balloux, A.C. Ghani, N.M. Ferguson, A. Rambaut, O.G. Pybus,

H. Lopez-Gatell, C.M. Alpuche-Aranda, I.B. Chapela, E.P. Zavala, D.M.E. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings, Science, **324** (2009) 1557-1561.

[62] Y. Hsieh, K. Cheng, T. Wu, T. Liz, C. Cheng, J. Chen, M. Lin, Transmissibility and temporal changes of 2009 pH1N1 pandemic during summer and fall/winter waves, BMC Infectious Diseases, **11**:332 (2011) DOI: 10.1186/1471-2334-11-332.

[63] D. Clancy, P.D. O'Neill, Bayesian estimation of the basic reproduction number in stochastic epidemic models, International Society for Bayesian Analysis, **3** (2008) 737-758.

[64] L.F. White, J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, M. Pagano, Estimation of the Reproductive Number and the Serial Interval in Early Phase of the 2009 Influenza the Current Influenza A/H1N1 Pandemic in the USA, Influenza and Other Respiratory Viruses, **3**(6) (2009) 267-276.

[65] G. Katriel, R. Yaari, A. Huppert, U. Roll, L. Stone, Modelling the initial phase of an epidemic using incidence and infection network data: 2009 H1N1 pandemic in Israel as a case study, Journal of the Royal Society Interface, **8**(59) (2011) 856-867.

[66] M.J. Smith, S. Telfer, E.R. Kallio, S. Burthe, A.R. Cook, X. Lambin, M. Begon, Host-pathogen time series data in wildlife support a transmission function between density and frequency dependence, Proceedings of the National Academy of Sciences of the United States of America, **106**(19) (2009) 7905-7909.

[67] T. Britton, Epidemics in heterogeneous communities: estimation of R-o and secure vaccination coverage, J.R. Statist. Soc. B-Stat. Methodol., **63**(Part 4) (2001) 705-715 DOI: 10.1111/1467-9868.00307.

[68] I.C. Ster, N.M. Ferguson, Transmission parameters of the 2001 foot and mouth epidemic in Great Britain, PloS one, **2**(6) (2007) p.e502.

[69] Y. Hsieh, H. Huang, Y. Lan, On Temporal Patterns and Circulation of Influenza Virus Strains in Taiwan, 2008-2014: Implications of 2009 pH1N1 Pandemic, PLoS One, **11**(5) (2016) DOI: 10.1371/journal.pone.0154695.

[70] F. Yang, L. Yuan, X.Tan, C. Huang, J. Feng, Bayesian estimation of the effective reproduction number for pandemic influenza A H1N1 in Guangdong Province, China, Annals of Epidemiology, **23**(6) (2013) 301-306.

[71] V.H. Borja-Aburto, G. Chowell, C. Viboud, L. Simonsen, M.A. Miller, C. Grajales-Muniz, C.R. Gonzalez-Bonilla, J.A. Diaz-Quinonez, S. Echevarria-Zuno, Epidemiological characterization of a fourth wave of pandemic A/H1N1 influenza in Mexico, winter 2011–2012: age shift and severity, Arch Med Res, **43**(7) (2012) 563–570.

[72] S. Tang, Y. Xiao, Y. Yang, Y. Zhou, J. Wu, Z. Ma, Community-based measures for mitigating the 2009 H1N1 pandemic in China, PLoS One, **5**(6) (2010) DOI: 10.1371/journal.pone.0010911.

[73] O. Barnea, R. Yaari, G. Katriel, L. Stone, Modelling seasonal influenza in Israel, Mathematical Bioscience and Engineering, **8**(2) (2011) 561-573.

[74] A. Huppert, O. Barnea, G. Katriel, R. Yaari, U. Roll, L. Stone, Modeling and statistical analysis of the spatio-temporal patterns of seasonal influenza in Israel, PLoS One, **7**(10) (2012) DOI: 10.1371/journal.pone.0045107.

[75] T. Smieszek, M. Balmer, J. Hattendorf, K.W. Axhausen, J. Zinsstag, R.W. Scholz, Reconstructing the 2003/2004 H3N2 influenza epidemic in Switzerland with a spatially explicit, individual-based model, BMC Infectious Diseases, **11**(115) (2011) DOI: 10.1186/1471-2334-11-115.

[76] A. Lunelli, C. Rizzo, S. Puzelli, A. Bella, E. Montomoli, M.C. Rot, I. Donatelli, A. Pugliese, Understanding the dynamics of seasonal influenza in Italy: incidence, transmissibility and population susceptibility in a 9-year period, Influenza Other Respiratory Viruses, **7**(2) (2013) 286-295.

[77] T. Britton, N.G. Becker, Estimating the immunity coverage required to prevent epidemics in a community of households, Biostatistics (Oxford, England), **1**(4) (2000) 389-402.

[78] D.Y. Chao, K.F. Cheng, T.C. Li, T.N. Wu, C.Y. Chen, C.A. Tsai, J.H. Chen, H.T. Chiu, J.J. Lu, M.C. Su, Y.H. Liao, W.C. Chan, Y.H. Hsieh, Serological evidence of subclinical transmission of the 2009 pandemic H1N1 influenza virus outside of Mexico, PLoS One, **6**(2) 2011 DOI: 10.1371/journal.pone.0014555.

[79] V. Marmara, A. Cook, A. Kleczkowski, Estimation of force of infection based on different epidemiological proxies: 2009/2010 Influenza epidemic in Malta, Epidemics, **9** (2014) 52-61.

[80] M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, L Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature, BMC Infectious Diseases, **14**(480) (2014) DOI: 10.1186/1471-2334-14-480.

[81] N.M. Ferguson, D.A. Cummings, C. Fraser, J.C. Cajka, P.C. Cooley, D.S. Burke DS, Strategies for mitigating an influenza pandemic, Nature, **442**(7101) (2006) 448–452.

[82] WHO, World Health Organization, Influenza (Seasonal), World Wide Web electronic, 2014, URL: http://www.who.int/mediacentre/factsheets/fs211/en/.

[83] NHS, National Health Service (UK), Symptoms of flu, 2015, URL: http://www.nhs.uk/Conditions/Flu/Pages/Symptoms.aspx.

[84] CDC, Centers for Disease Control and Prevention (USA), Key Facts about Influenza (Flu) & Flu Vaccine, 2015, URL: http://www.cdc.gov/flu/keyfacts.htm#whatis.

[85] CDPH, California Department for Public Health (USA), H1N1 Influenza, 2009, URL: https://www.cdph.ca.gov/HealthInfo/h1n1flufaqs/Pages/H1N1fluFAQs-01-GenInfo.aspx#01.06.

[86] E.Y. Chan, C.K. Cheng, G.C. Tam, Z. Huang, P.Y. Lee, Willingness of future A/H7N9 influenza vaccine uptake: A cross-sectional study of Hong Kong community, Vaccine, **33** (2015) 4737–4740.

[87] P. Mangtani, E. Breeze, S. Stirling, S. Hanciles, S. Kovats, A. Fletcher, Cross-sectional survey of older peoples' views related to influenza vaccine uptake, BMC Public Health, **6**:249 (2006) DOI: 10.1186/1471-2458-6-2492011.

[88] A. Pfeil, M. Mutsch, C. Hatz, T.D. Szucs, A cross-sectional survey to evaluate knowledge, attitudes and practices (KAP) regarding seasonal influenza vaccination among European travellers to resource-limited destinations, BMC Public Health, **10**:402 (2010) DOI: 10.1186/1471-2458-10-402

[89] G.E. Khoury, P. Salameh, Influenza Vaccination: A Cross-Sectional Survey of Knowledge, Attitude and Practices among the Lebanese Adult Population, International Journal of Environment Research and Public Health, **12** (2015) 15486-15497.

[90] J. Zhang, A.E. While, I.J. Norman, Nurses' knowledge and risk perception towards seasonal influenza and vaccination and their vaccination behaviours: A cross-sectional survey, International Journal of Nursing Studies, **48** (2011) 1281–1289.

[91] J. Mereckiene, S. Cotter, A. Nicoll, P. Lopalco, T. Noori, J.T. Weber, F. D'Ancona, D. Levy-Bruhl, L. Dematte, C. Giambi, P, Valentiner-Branth, I. Stankiewicz, E. Appelgren, D. O'Flanagan, the VENICE project gatekeepers group, Seasonal influenza immunisation in Europe. Overview of recommendations and vaccination coverage for three seasons: pre-pandemic (2008/09), pandemic (2009/10) and post-pandemic (2010/11), Euro Surveillance, **19**(16) (2014) URL: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20780.

[92] J. Yang, M. Jit, K.S. Leung, Y. Zheng, L. Feng, L. Wang1, E.H.Y. Lau, J.T. WU, H. Yu, The economic burden of influenza-associated outpatient visits and hospitalizations in China: a retrospective survey, Infectious Diseases of Poverty, **4**:44 (2015) DOI: 10.1186/s40249-015-0077-6.

[93] B.Y. Lee, K. Bacon, J.M. Donohue, A.E. Wiringa, R.R. Bailey, R.K. Zimmerman, From the Patient Perspective: the Economic Value of Seasonal and H1N1 Influenza Vaccination, Vaccine, **29**(11) (2011) 2149–2158.

[94] S. Corson, C. Robertson, A. Reynolds, J. McMenamin, Modelling the population effectiveness of the national seasonal influenza vaccination programme in Scotland: the impact of targeting all individuals aged 65 years and over, Vaccine, **00** (2015) 1–12.

[95] G. Chowell, S. Towers, C. Viboud, R. Fuentes, V. Sotomayor, L. Simonsen, M.A. Miller, M. Lima, C. Villarroel, M. Chiu, J.E. Villarroel, A. Olea, The influence of climatic conditions on the transmission dynamics of the 2009 A/H1N1influenza pandemic in Chile, BMC Infectious Diseases, **12**:298 (2012) URL: http://www.biomedcentral.com/1471-2334/12/298.

[96] P. Hoagland, D. Jin, L.Y. Polansky, B. Kirkpatrick, G. Kirkpatrick, L.E. Fleming, A. Reich, S.M. Watkins, S.G. Ullmann, L.C. Backer, The Costs of Respiratory Illnesses Arising from Florida Gulf Coast Karenia brevis Blooms, Environmental Health Perspectives, Environmental Health Perspectives, **117**(8) (2009) 1239-1243.

[97] C. K. Irwin, K. J. Yoon, C. Wang, S. J. Hoff, J. J. Zimmerman, T. Denagamage, and A. M. O'Connor, Using the Systematic Review Methodology To Evaluate Factors That Influence the Persistence of Influenza Virus in Environmental Matrices, Applied and Environmental Microbiology, **77**(3) (2011) 1049–1060.

[98] N. Tuncer, M. Martcheva, Modeling seasonality in avian influenza H5N1, Journal of Biological Systems, **21**(4) (2013) DOI: 10.1142/S0218339013400044.

[99] L. Hogerwerf, R.G. Wallace, D. Ottaviani, J. Slingenbergh, D. Prosser, L. Bergmann, M. Gilbert, Persistence of Highly Pathogenic Avian Influenza H5N1 Virus Defined by Agro-Ecological Niche, EcoHealth, **7**(2) (2010) 213-225.

[100] J. Steel, P. Palese, A.C. Lowen, Transmission of a 2009 Pandemic Influenza Virus Shows a Sensitivity to Temperature and Humidity Similar to That of an H3N2 Seasonal Strain, Journal of Viriology, **85**(3) (2011) 1400–1402.

[101] K.S. Li, Y. Guan, J. Wang, G.J. Smith, K.M. Xu, L. Duan, A.P. Rahardjo, P. Puthavathana, C. Buranathai, T.D. Nguyen, A.T. Estoepangestie, A. Chaisingh, P. Auewarakul, H.T. Long, N.T. Hanh, R.J. Webby, L.L. Poon, H. Chen, K.F. Shortridge, K.Y. Yuen, R.G. Webster, J.S. Peiris, Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia, Nature, **8**(430) (2004) 209-213.

[102] A.C. Lowen, S. Mubareka, J. Steel, P. Palese, Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature, Journal of Viriology, **3**(10) (2007) 1470-1476.

[103] C. Wu, J. Lu, M.H. Wang, X. Lv, Y. Chen, H. Kung, B. Zee, X. Cheng, M. He, Cross Sectional Survey of Influenza Antibodies before and during the 2009 Pandemic in Shenzhen, China, PLoS ONE, **8**(1) (2013) DOI: 10.1371/journal.pone.0053847.

[104] X. Tan, L.Yuan, J. Zhou, Y. Zheng, F. Yang, Modeling the initial transmission dynamics of influenza A H1N1 in Guangdong Province, China, International Journal of Infectious Diseases, **17** (2013) e479–e484.

[105] V.J. Lee, M.I. Chen, J. Yap, J. Ong, W. Lim, R.T.P. Lin, I. Barr, J.B.S. Ong, T.M. Mak, L.G. Goh, Y.S. Leo, P.M. Kelly, A.R. Cook, Comparability of Different Methods for Estimating Influenza Infection Rates Over a Single Epidemic Wave, American Journal of Epidemiology, **174**(4) (2011) 468–478.

[106] F. Anday, P. Crepey, A. Kieffer, N. Salez, A.A. Abdo, F. Carrat, A. Flahault, X. Lamballerie, Determinants of individuals' risks to 2009 pandemic influenza virus infection at household level amongst Djibouti city residents - A CoPanFlu cross-sectional study, Virology Journal, **11**:13 (2014) URL: http://www.virologyj.com/content/11/1/13.

[107] S.E. Soh, A.R Cook, M.I.C. Chen, V.J. Lee, J.L. Cutter, V.T.K. Chow, N.W.S. Tee, R.T.P. Lin, W. Lim, I.G. Barr, C. Lin, M.C. Phoon, L.W. Ang, S.K. Sethi, C.Y. Chong, L.G. Goh, D.L.M. Goh, P.A. Tambyah, K.C. Thoon, Y.S. Leo, S.M. Saw, Teacher led school-based surveillance can allow accurate tracking of emerging infectious diseases - evidence from serial cross-sectional surveys of febrile respiratory illness during the H1N1

2009 influenza pandemic in Singapore, BMC Infectious Diseases, **12**:336 (2012) URL: http://www.biomedcentral.com/1471-2334/12/336.

[108] P.A. Oria, G. Arunga, E. Lebo, J.M. Wong, G. Emukule, P. Muthoka, N. Otieno, D. Mutonga, R.F. Breiman, M.A Katz, Assessing parents' knowledge and attitudes towards seasonal influenza vaccination of children before and after a seasonal influenza vaccination effectiveness study in low-income urban and rural Kenya, 2010–2011, BMC Public Health, **13**:391 (2013) URL: http://www.biomedcentral.com/1471-2458/13/391.

[109] UK flusurvey, London School of Hygiene and Tropical Medicine and Public Health England, 2016 URL: https://flusurvey.org.uk/en/.

[110] A. Camacho, K. Eames, A. Adler, S. Funk, J. Edmunds, Estimate of the quality of life effect of seasonal influenza infection in the UK with the internet-based Flusurvey cohort: an observational cohort study, Lancet, **383**(3) (2013) 8-8.

[111] A.J. Adler, K.T.D. Eames, S. Funk, W.J. Edmunds, Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey, BMC Infectious Diseases, **14**:232 (2014) URL: http://www.biomedcentral.com/1471-2334/14/232.

[112] Gripenet, Department of Epidemiology, National Institute of Health Dr. Ricardo Jorge, Lisbon Portugal, 2016 URL: http://www.gripenet.pt/pt/.

[113] GrippeNet, Institute Pierre Louis, France, 2016 URL: https://www.grippenet.fr/fr/.

[114] Gripenet, Complex Systems and Networks Lab, Institute for Biocomputation and Physics of Complex Systems (BIFI), Campus Rio Ebro, University of Zaragoza, Spain, 2016 URL: https://www.gripenet.es/es/.

[115] S. Galea, M. Tracy, Participation Rates in Epidemiologic Studies, Ann Epidemiology, **17** (2007) 643–653.

[116] A. Jutel, M.G. Baker, J. Stanley, Q.S. Huang, D. Bandaranayake, Self-diagnosis of influenza during a pandemic: a cross-sectional survey, BMJ Open, **1** (2011) DOI: 10.1136/bmjopen-2011-000234

[117] S.J. Rolnick, E.D. Parker, J.D. Nordin, B.D. Hedbloma, F. Wei, T. Kerby, J.M. Jackson, A.L Crain, G. Euled, Self-report compared to electronic medical record across eight adult vaccines: Do results vary by demographic factors?, Vaccine, **31**(37) (2013) 3928–3935.

[118] P.J. Birrell, G. Ketsetzis, N.J. Gay, B.S. Cooper, A.M. Presanis, R.J. Harris, A. Charlett, X. Zhang, P.J. White, R.G. Pebody, D.D Angelis, Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London, PNAS, **108**(45) (2008) 18238-18243.

[119] J.L. Malone, M. Madjid, S.W. Casscells, Telephone Survey to Assess Influenza-like Illness, United States, 2006, Emerging Infectious Diseases, **14**(1) (2008) URL: www.cdc.gov/eid.

[120] E.O. Nsoesie, J.S. Brownstein, N. Ramakrishnan, M.V. Marathe, A systematic review of studies on forecasting the dynamics of influenza outbreaks, Influenza and other respiratory viruses, **8**(3) (2014) 309-316.

[121] C. Viboud, P. Boëlle, F. Carrat, A. Valleron, A. Flahault, Prediction of the Spread of Influenza Epidemics by the Method of Analogues, American Journal of Epidemiology, **158**(10) (2003) 996-1006.

[122] A. Aguirre, E. Gonzalez, The feasibility of forecasting influenza epidemics in Cuba, Mem. Inst Oswaldo Cruz, **87**(3) (1992) 429-432.

[123] X. Jiang , G. Wallstrom, G.F. Cooper, M.M. Wagner, Bayesian prediction of an epidemic curve, Journal of Biomedical Informatics, **42**(1) (2009) 90–99.

[124] R.P. Soebiyanto1, F. Adimi, R.K. Kiang, Modeling and Predicting Seasonal Influenza Transmission in Warm Regions Using Climatological Parameters, PLoS One, **5**(3) (2010) DOI: 10.1371/journal.pone.0009450.

[125] P.M. Polgreen, F.D. Nelson, G.R. Neumann, Use of prediction markets to forecast infectious disease activity, Clinical Infectious Diseases, **44**(2) (2007) 272-279.

[126] E. Nsoesie, M. Mararthe, J. Brownstein, Forecasting Peaks of Seasonal Influenza Epidemics, PLoS Currents, **1** (2013) DOI: 10.1371/currents.outbreaks.bb1e879a23137022ea79a8c508b030bc

[127] S. Towers, Z. Feng, Pandemic H1N1 influenza: predicting the course of a pandemic and assessing the efficacy of the planned vaccination programme in the United States, Eurosurveillance, **14**(42) (2009) 6-8.

[128] J. Shaman, A. Karspeck, Forecasting seasonal outbreaks of influenza, PNAS, **109**(50) (2012) 20425-20430.

[129] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, M. Lipsitch, Real-time influenza forecasts during the 2012–2013 season, Nature Communications, **4** (2013) DOI: 10.1038/ncomms3837.

[130] A. Kleczkowski, C.A. Gilligan, Parameter estimation and prediction for the course of a single epidemic outbreak of a plant disease, Journal of the Royal Society Interface, **4**(16) (2007) 865-877.

[131] E.O. Nsoesie, R.J. Beckman, S. Shashaani, K.S. Nagaraj, M.V. Marathe, A Simulation Optimization Approach to Epidemic Forecasting, PLoS ONE, **8**(6) (2013) DOI: 0.1371/journal.pone.0067164.

[132] E. Andersson, S.K. Hlmann-Berenzon, A. Linde, L. Schioler, S. Rubinova, M. Frisen, Predictions by early indicators of the time and height of the peaks of yearly influenza outbreaks in Sweden, Scandinavian Journal of Public Health, **36** (2008) 475–482.

[133] R Development Core Team (2010) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, (2010) URL: http://www.R-project.org, ISBN 3-900051-07-0.

[134] C. Savona-Ventura, Civil Hospitals in Malta, Archived from the original on 26 October 2009, Retrieved: 27 July 2016, URL: http://www.webcitation.org/query?url=http://www.geocities.com/hotsprings/2615/medhist/hospital2.htm&date=2009-10-26+00:12:01.

[135] P. Coustsoukis, The World Health Organization's raking of the world's health systems, Retrieved: 27 July 2016, URL: http://www.photius.com/rankings/healthranks.html.

[136] St. Philip's Hospital, St. Philip's Hospital – A modern 75-bed hospital quipped with the latest medical technology – Malta, Retrieved: 27 July 2016, URL: www.stphilips.com.mt.

[137] Ministry for Health, The Health Care System in Malta, Retrieved: 27 July 2016, URL: https://web.archive.org/web/20070711220500/http:/www.sahha.gov.mt/pages.aspx?page=156.

[138] Allo' Expat Malta, Healthcare in Malta, Retrieved: 27 July 2016, URL: http://www.alloexpat.com/moving_to_malta_forum/healthcare-in-malta-t162.html.

[139] Ministry for Health, Directorate for Health Information and Research, Retrieved: 27 July 2016, URL: https://health.gov.mt/en/dhir/Pages/Introduction.aspx.

[140] S. Kumar, G.S. Preetha, Health Promotion: An Effective Tool for Global Health, Indian Journal of Community Medicine, **37**(1) (2012) 5-12. DOI: 10.4103/0970-0218.94009.

[141] Ministry for Health - Primary Child & Youth Health & Immunisation Unit, Vaccines, Retrieved: 27 July 2016, URL: https://health.gov.mt/en/phc/pchyhi/Pages/Vaccines.aspx.

[142] Creative Research Systems, 'The Survey System', Retrieved: 10 September 2015, URL: http://www.surveysystem.com/sscalc.htm.

[143] Malta International Airport, Weather services, Retrieved: 15 July 2016, URL: https://www.maltairport.com/weather/weather-services/.

[144] V.M. Trenkel, D.A. Elston, S.T. Buckland, Fitting population dynamics models to count and cull data using sequential importance sampling, Journal of the American Statistical Association, **95** (2000) 363–74.

[145] Y. Hu, P. Baraldi, F. Di Maio, E. Zio, A particle filtering and kernel smoothing-based approach for new design component prognostics, Reliability Engineering and System Safety, **134** (2015) 19–31.

[146] IBM SPSS, IBM SPSS Software, United States, 2016, URL: http://www.ibm.com/analytics/us/en/technology/spss/spss.html.

[147] S. Froda, H. Leduc, Estimating the basic reproduction number from surveillance data on past epidemics, Mathematical Biosciences, **256** (2014) 89–101.

[148] Q. Lin, Z. Lin, A.P.Y. Chiu, D. He, Seasonality of Influenza A(H7N9) Virus in China—Fitting Simple Epidemic Models to Human Cases, PLoS One, **11**(3) (2016) DOI: 10.1371/journal.pone.0151333.

[149] R. Pebody, F. Warburton, J. Ellis, N. Andrews, A. Potts, S. Cottrell, J. Johnston, A. Reynolds, R. Gunson, C. Thompson, M. Galiano, C. Robertson, D. Mullett, N. Gallagher, M. Sinnathamby, I. Yonova, C. Moore, J. McMenamin, S. de Lusignan, M. Zambon, Effectiveness of seasonal influenza vaccine in preventing laboratory-confirmed

influenza in primary care in the United Kingdom: 2015/16 mid-season results, Euro Surveill., **21**(13) 2016 DOI: http://dx.doi.org/10.2807/1560-7917.ES.2016.21.13.30179.

[150] K. Mizumoto, H. Nishiura, T. Yamamoto, Effectiveness of antiviral prophylaxis coupled with contact tracing in reducing the transmission of the influenza A (H1N1-2009): a systematic review, Theoretical Biology and Medical Modelling, **10**(4) (2013) DOI:10.1186/1742-4682-10-4.

[151] B. Pourbohloul, A. Ahued, B. Davoudi, R. Meza, L.A. Meyers, D.M. Skowronski, I. Villasenor, F. Galvan, P. Cravioto, D.J.D. Earn, J. Dushoff, D. Fisman, W.J. Edmunds, N. Hupert, S.V. Scarpino, J. Trujillo, M. Lutzow, J. Morales, A. Contreras, C. Chavez, D.M. Patrick, R.C. Brunham, Initial human transmission dynamics of the pandemic (H1N1) 2009 virus in North America, Influenza and Other Respiratory Viruses **3**(5), 215–222.

[152] J. Davila-Torres, G. Chowell, V.H. Borja-Aburto, C. Viboud, C. Grajalez-Muniz, M.A. Millerb, Intense Seasonal A/H1N1 Influenza in Mexico, Winter 2013-2014, BMC Infectious Diseases, Archives of Medical Research, **46**(1) (2015) 63-70.

[153] WHO, Pandemic (H1N1) 2009 - update 100, 2010 URL http://www.who.int/csr/don/2010_05_14/en/index.html.

[154] A. Ishak , D. Tee, I. Nawmar, L.K. Pang, N. Ruslan, N. Che Mansor, L. Gam, H1N1 Influenza: A Viral Infection, Webmed Central Infectious Diseases, **2**(12) (2011) WMC002736.

[155] C. Reed, F.J. Angulo, D.L. Swerdlow, M. Lipsitch, M.I. Meltzer, D. Jernigan, and L. Finelli, Estimates of the Prevalence of Pandemic (H1N1) 2009, United States, April–July 2009, Nature Communication, **4** (2013) 2004-2007.

[156] NHS England, The national flu immunisation programme 2014/15, World Wide Web electronic, 2014 URL https://www.england.nhs.uk/south/wp-content/uploads/sites/6/2014/08/phe-nat-flu-prog-slides-14-15.pdf.

[157] L.Yang, K.H. Chan, L.K.P Suen, K.P. Chan, X. Wayng, P. Cao, D. He, J.S.M. Peiris, C.M. Wong, Age-specific epidemic wave of influenza and respiratory syncytial virus in a subtropical city, Scientific Reports, **5** (10390) (2015) DOI: 10.1038/srep10390.

[158] PAGB, Proprietary Association of Great Britain, World Wide Web electronic, 2016 URL: http://www.pagb.co.uk/publications/pdfs/Coldandfluleaflet.pdf, 2014

[159] WebMD, Cold, Flu, & Cough Health Center, World Wide Web electronic, 2016 URL: http://www.webmd.com/cold-and-flu/flu-guide/advanced-reading-types-of-flu-viruses#4.

[160] Malta International Airport, Last winter, driest winter on record, World Wide Web electronic, 2016 URL: https://www.maltairport.com/last-winter-driest-winter-record.

[161] J.M. Nagata, I. Hernández-Ramos, A.S. Kurup, D. Albrecht, C. Vivas-Torrealba, C. Franco-Paredes, Social determinants of health and seasonal influenza vaccination in adults ≥65 years: a systematic review of qualitative and quantitative data, BMC Public Health, **13**(388) (2013) DOI: 10.1186/1471-2458-13-388.

[162] S. Shahrabani, U. Benzion, How Experience Shapes Health Beliefs: The Case of Influenza Vaccination, Health Education & Behavior, **39**(5) (2012) 612–619.

[163] A.M.S. Wu, J.T.F. Lau, Y.L. Ma, M.M.C. Lau, Prevalence and associated factors of seasonal influenza vaccination among 24- to 59-month-old children in Hong Kong, Vaccine, **33**(30) (2015) 3556-3561.

[164] National Statistics Office, Malta (NSO), 'Malta in Figures 2014', 2015 URL: http://nso.gov.mt/en/publicatons/Pages/Publications-by-Date.aspx

[165] K.S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, S.Y. Del Valle, Forecasting the 2013-2014 Influenza Season using Wikipedia, Plos Computational Biology, **11**(5) (2015) DOI: 10.1371/journal.pcbi.1004239.

[166] M. Shubin, A. Lebedev, O. Lyytikäinen, K. Auranen, Revealing the True Incidence of Pandemic A (H1N1)pdm09 Influenza in Finland during the First Two Seasons—An Analysis Based on a Dynamic Transmission Model, Plos Computational Biology, **11**(3) (2016) DOI: 10.1371/journal.pcbi.1004803.

[167] J. Mossong, N. Hens, M. Jit , P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G.S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, W.J. Edmunds, Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases, PLOS Medicine, **5**(3) (2008) 381-391.

[168] S. Blaizot, B. Riche, D. Maman, I. Mukui, B. Kirubi, J.F. Etard, R. Ecochard, Estimation and Short-Term Prediction of the Course of the HIV Epidemic Using Demographic and Health Survey Methodology-Like Data, PLOS ONE, **10**(6) (2015) DOI: 10.371/journal.pone.0130387.

[169] Q. Lin, Z. Lin, A.P.Y. Chiu, D. He, Seasonality of Influenza A(H7N9) Virus in China—Fitting Simple Epidemic Models to Human Cases, PLoS One, **11**(3) (2016) DOI: 10.1371/journal.pone.0151333.

[170] P.D. O'Neill, Introduction and snapshot review: Relating infectious disease transmission models to data, Statistics in medicine, **29**(20) (2010) 2069-2077.

Appendix A

Minutes of Meetings held in Malta with Health Officials

Meeting with the CEO of Primary Health Care – Dr. Renzo Degabriele

Members Present: Dr. Renzo Degabriele 'RD' (CEO), Vincent Marmarà 'VM' (PhD Student)

Date: 7th October 2015

Time: 10am

Office: Primary Health Care Directorate (PHCD)

During the meeting, VM gave an overview of his published research paper and the current analysis being carried out about Malta regarding the seasonal influenza. After going through the salient findings of the research, a number of points were raised, mainly that:

1. PHCD are interested in having an early warning technique about the level of aggressiveness of any disease, hence the earlier the information is available, the better it will be for key stakeholders in the planning of strategies, which will include all the required logistics.

2. RD is very interested in this research as it helps him to plan strategically.

3. Such information can help RD to plan with regards to human resources, interventions, annual leave and sick leave of the employees and other administrative matters.

4. Such information helps the department when planning new services within the directorate; hence, by knowing the level of extent of seasonal influenza, the key stakeholders can plan more appropriately.

5. Such information helps when developing several health promotion campaigns. Hence, they can adjust the scale of national campaigns accordingly.

6. If a high number of infected individuals are predicted, then they can issue nation-wide warnings, initiate earlier campaigns, increase the level of hygienic initiatives in schools and in other public spaces.

7. Furthermore, health officials will be able to strategically and adequately plan the number of medical staff required during this period.

8. We then discussed briefly the required amount of medicines when such outbreaks occur. Hence, I questioned: how and who will decide on the number of medicines required in such cases? However, the appropriate person to answer such a question is the Head of Health Information and Research. VM will be setting a meeting with this department (Dr. Neville Calleja, Head) next week and will try to obtain further information about this interesting issue.

Meeting with the Head of Health Information and Research – Dr. Neville Calleja

Members Present: Dr. Neville Calleja 'NC' (Head), Vincent Marmarà 'VM' (PhD Student)

Date: 14th October 2015

Time: 10am

Office: Department of Health Information and Research (DHIR)

The following are the salient points of this meeting:

1. NC: 'During the 2014/2015 Seasonal Influenza, Malta was close to having a mini-epidemic since the rates were higher than usual.'

2. NC: January is the point of major influenza increase as children start school again and hence the spread of the viruses increases.

3. NC's suggestion: To look at the temperatures during the seasonal influenza period. He said, "Usually, a sharp drop in temperatures triggers the acceleration of influenza." Hence, VM will try to acquire the temperature dataset.

4. For the department, predicting the spread of influenza and creating an early model warning technique is very useful. This is the kind of information that is mostly needed.

5. VM went through the analysis that we are currently performing regarding the understanding of the relationship between Consultations and Diagnosed individuals. For the department, any early signal that the consultations/diagnosed ILIs can provide is very helpful and useful for strategic planning.

6. DHIR is concerned about the under-reporting rate by GPs and in his opinion, this needs to be divided into 3 groups, mainly: a) people not reporting their influenza, b) GPs not reporting their cases and c) GPs not diagnosing correctly.

7. NC: It is interesting to look at the relationship between different years of Seasonal Influenza.

8. Medicines are ordered in January, basing the number of required medicines on the consumption during the previous November - December vaccination campaign. They are usually delivered with the vaccine in October of the same year.

9. Regarding the symptoms of the influenza, (with reference to the Under-reporting survey) and due to the fact that around 85% of the Maltese population indicated that they had any of the mentioned symptoms, NC said that those are symptoms related to a possible Seasonal Influenza case.

10. Suggestion: NC said that it would be interesting to look at the Influenza-related admissions at Malta's state hospital. NC directed VM to another department i.e. to set a meeting with the Consultant of Public Health Medicine to try to obtain the admissions data.

Meeting with the Health Promotion Department - Infectious Disease Prevention and Control Unit, Health Promotion and Disease Prevention Directorate – Dr. Tanya Mellilo & Dr. Jackie Mellilo

Members Present: Dr. Tanya Mellilo 'TM' (Officer), Dr. Jackie Mellilo 'JM' (Officer), Vincent Marmarà 'VM' (PhD Student)

Date: 26th October 2015

Time: 9am

Office: Malta Health Promotion Department (MHPD)

The following are the salient points of this meeting:

1. MHPD are very interested to obtain an actual estimate of the number of people who acquire Seasonal Influenza per year, ideally also by age (***We can obtain this estimate from the Survey***). The public health professionals will highly appreciate such data in order to plan and improve their strategy.

2. Prior to the 'Seasonal influenza' data collection (end of September), it is believed that there are only limited numbers of Seasonal Influenza cases.

3. As a Health Promotion department, public health officials focus mainly on the impact of the Seasonal Influenza, especially due to its high financial impact on the health sector.

4. The costs of the Seasonal Influenza are very high due to the high demands on doctors, hospitals, staff, vaccines and marketing, especially amongst elderly people.

5. The impact of influenza is very important for the Health Promotion department. Hence, they are interested to receive more insight about the Seasonal Influenza in Malta.

6. Better information and early warning techniques will help them to design an improved policy and to adjust the local needs according to the demands.

7. Their main issue on H1N1 is that there was a peak during the summer period (this goes against the norm to have the influenza peak during the summer period).

8. MHPD suggested a comparison of the seasonal influenza data with the temperatures in Malta. MHPD's hypothesis is that the lowest temperature may lead to an influenza outbreak (ACTION: Currently VM is trying to obtain more temperature data).

9. A very interesting discussion on the published research paper took place as well and further points were highlighted on the current analysis. Currently, the health promotion department is discussing internally when to administer the vaccine to the general population in Malta. The immunity to the vaccine remains effective for 6 months. After

discussing the PhD's datasets together, the department decided to administer the seasonal influenza vaccine during the end of November instead of early November. Basically, the influenza's peak is occurring during February-April, hence it is more useful to maximize the strength of the vaccine by postponing the administration of the vaccine by 3 weeks, rather than administering it during early November (low influenza month).

10. During the meeting, TM said that seasonal influenza occurs due to the circulating influenza sub-types that are circulating at the time. Usually the A and B influenza type virus are two components that determine the characteristics of the vaccine. Hence, an individual will not acquire the influenza (Type A and B) if the vaccine is administered. If a person did not accept to receive the vaccine, one can acquire the influenza caused by the A virus and if unlucky, could also acquire the influenza caused by the B type in the same season. Usually, a person acquires the influenza once in a season due to one of the circulating viruses that is the most dominant. TM insisted that a person becomes immune to the seasonal influenza virus once exposed to it, but could still acquire it from another seasonal virus which is in circulation. TM concluded by saying: "that is why the vaccine is made up of 3 circulating viruses - the ones they think will be circulating during the winter season".

11. Further actions: the department will approve and forward the seasonal influenza 2015/2016 data to VM on a monthly basis to conduct further research work.

12. They are trying to obtain the data of the public local clinic consultations during the seasonal influenza period. As soon as they capture this data, they will try to approve and forward this data to VM.

Meeting with the Minister responsible for the Health Sector in Malta – Hon. Mr. Chris Fearne.

Members Present: Hon. Mr. Chris Fearne (CF), Vincent Marmarà 'VM' (PhD Student)

Date: 19th November 2015

Time: 1pm

Office: Health Ministry, Malta

This is an update after the meeting with the Minister responsible for the health sector in Malta:

1. CF is very interested in this research but is mainly interested in a predictive model that predicts the number of admissions at the General hospital. Such numbers can help to predict the healthcare demands in Malta.

2. He believes that temperature data has an important role in the epidemic.

3. From a political perspective, Hon. Chris Fearne is responsible to ensure that there are adequate beds available in the hospital. Hence, any work related to this area is of great interest to him.

4. Hon. Chris Fearne will forward VM's research paper to a health advisor within the hospital and will facilitate a meeting for VM with the health advisor in order to carry out further discussions on the topic.

Meeting with the Minister's Consultant responsible for the Health Sector in Malta – Mr. Mike Farrugia

Members Present: Mr. Mike Farrugia (Consultant) (MF), Vincent Marmarà 'VM' (PhD Student)

Date: 4th December 2015

Time: 10:30am

Office: Mater Dei General Hospital

The following are the major points of interest for the Consultant:

1. MF: "How to keep out patients from coming to hospital due to Seasonal Influenza?"

2. MF: "To what extent can we help people not to acquire influenza?"

3. MF is very interested in predicting the demand on the hospital beds

4. MF: "Temperature data: is this a major predictor?"

Appendix B

The research instrument

Under reporting Seasonal Flu Survey (English Version)

Gender _____ Locality _____ Age _____ Married _____

Status: Employee □ Student □ Housewife □ Unemployed □ Pensioner □

What is your job?_____ Level of Education reached _____

Number of individuals at your household (including you): _____ What is their age? _____

1. What is your main means of transport?
   Walking ___ Bike ___ Motorbike ___ Car ___ Public transport ___ Other ___
2. Did you receive a flu vaccine this winter/autumn season? (2014-2015)
   Yes ___ No ___ I don't know ___
3. If 'No', why?
   _____
4. How many times did you visit your GP (doctor) during this past year? _____
5. Do you take regular medication for any medical conditions such as asthma, diabetes, heart disorders, kidney disorder or other? Yes ___ No ___ I don't know ___
6. Do you smoke? Yes ___ No ___
7. If 'Yes', how many cigarettes per day? _____
8. Have you had any of the following symptoms during the past year?
   Fever ___ runny or blocked nose ___ Sneezing ___ Sore throat ___ Cough ___
   Shortness of breath ___ Headache ___ Muscle/joint pain ___ Chest pain ____
   Feeling tired or exhausted ___ Loss of appetite ___ Watery eyes ___ Nausea ___
   Vomiting ___ Diarrhoea ___ Stomach ache ___ Other symptoms ___ Nothing ___
9. When did your symptoms appear for the above during the past year? _____
10. Were you restricted to staying at home? Yes ___ No ___ I don't know ___
11. Approximately, in days, how long was the duration for the above symptoms? _____
12. Since 'August 2014' did you have the seasonal influenza? Yes ___ No ___ I don't know ___
13. If 'Yes', when did you have the seasonal influenza? _____
14. Approximately, in days, how long was the duration of influence? _____
15. What were the symptoms?
    _____
16. Did you have high temperature? Yes ___ No ___ I don't know ___
17. Because of your seasonal influenza fever, did you visit a doctor?
    Yes ___ No ___ I don't know ___
18. Did you take any medication due to this influenza? Yes ___ No ___ I don't know ___
19. Were you hospitalized due to your influenza? Yes ___ No ___
20. If 'Yes' for how many nights? _____
21. Did any of your household members had the seasonal influenza?
    Yes ___ No ___ I don't know ___
22. If 'Yes', how many members? _____
23. And what is their age? _____

Under reporting Seasonal Influenza Survey (Maltese Version)

Sess _____ Lokalita _____ Eta _____ Mizzewweg _____
Stat: Haddiem □ Student □ Mara tad-dar □ Bla xoghol □ Pensjonant □
X'inhu l-job tieghek?_____ Livell ta' Edukazzjoni li wasalt: _____
Inkluz inti, kemm toqghodu nies id-dar? _____ X'inhi l-eta taghhom? _____

1. X'inhu l-mezz principali tat-trasport tieghek?
   Nimxi ___ Rota ___ Mutur ___ Karrozza ___ Tal-linja ___ Ohrajn ___
2. Hadtu l-vacin tal-influenza f'din l-ahhar sena minn Awissu tal-2014? Iva __ Le __ Ma nafx _
3. Jekk 'Le' ghalxiex? _____
4. Kemm il-darba zort it-tabib tal-familja f'din l-ahhar sena? _____
5. Tiehu medikazzjoni regolari minhabba diversi mard u kunduzzjonijiet kronici bhal asthma, diabetes, mard tal-qalb, problem fil-kliewi u ohrajn? Iva ___ Le ___ Ma nafx __
6. Inti tpejjep? Iva ___ Le ___
7. Jekk 'Iva', kemm tpejjep sigaretti kulljum? _____
8. Kellek xi sintomi min dawn li gejjien f'din l-ahhar sena, minn Awissu tal-2014?
   Deni ___ imnieher ibblukat ___ Hafna ghatis ___ Ugiegh fil-grizmejn ___ Sola ___
   Qtuh ta' nifs ___ Ugieh ta' ras ___ Ugiegh fil-joints jew muskoli ___ Ugiegh f'sidrek ____
   Ghajja kbira u bla sahha ___ Nuqqas t'aptit ___ Ghajnejk jdemmghu ___ Dardir u tqallieh ___ Remettar ___ Diarrhoea ___ Ugiegh fl-istonku ___ Sintomi ohra ___ Xejn __
9. F'liema xhur kellek dawn is-sintomi f'din l-ahhar sena, minn Lulju tal-2014? _____ (xi xhur partikolari)
10. Minhabba dawn is-sintomi kellek toqghod id-dar? Iva ___ Le ___ Ma nafx ___
11. Bejn wiehed u iehor, fi granet, kemm damu dawn is-sintomi? _____
12. Minn 'Awissu tal-2014' sal-lum kellek influwenza (seasonal influenza)? **Iva** __ Le__ Ma nafx__
13. Jekk '**Iva**', f'liema xhur kellek l-influwenza? _____ (xi xhur partikolari)
14. Jekk '**Iva**', bejn wiehed u iehor, fi granet, kemm damet l-influwenza? _____
15. Jekk '**Iva**', x'kienu is-sintomi tal-influwenza? _____
16. Jekk '**Iva**', kellek deni matul dan iz-zmien tal-influwenza? Iva ___ Le ___ Ma nafx ___
17. Jekk '**Iva**', minhabba l-influwenza zort it-tabib? Iva ___ Le ___ Ma nafx ___
18. Jekk '**Iva**', hadt xi medicina minhabba l-influwenza? Iva ___ Le ___ Ma nafx ___
19. Jekk '**Iva**', kellek tidhol l-isptar minhabba l-influwenza? **Iva** ___ Le ___ Ma nafx ___
20. Jekk '**Iva**' ghal kemm iljieli? _____
21. Uhud mill-membri li qeghdin fid-dar mieghek kellhom din l-influwenza? **IVA** __ Le __ Ma nafx_
22. Jekk '**IVA**', kemm membri? _____
23. Jekk '**IVA**', kemm ghandhom zmien dawn il-membri? _____

Appendix C

The SEIR model together with the Particle Filter Algorithm code

The following code is a modification of Professor Alex Cook's code and I used this code with the author's permission.

```
i=Hist$t

#Print mean of parameters
x=Hist$beta[i,];print(paste("beta : mean =",mean(x),"sd =",sd(x)))
x=Hist$epsilon[i,];print(paste("epsilon : mean =",mean(x),"sd =",sd(x)))
x=Hist$lambda[i,];print(paste("lambda : mean =",mean(x),"sd =",sd(x)))
x=Hist$gamma[i,];print(paste("gamma : mean =",mean(x),"sd =",sd(x)))
x=Hist$falseflu[i,];print(paste("phi : mean =",mean(x),"sd =",sd(x)))

x=Hist$delta1[i,];print(paste("delta1 : mean =",mean(x),"sd =",sd(x)))

#Print 95% CI for number of removed individuals
x=Hist$R[i,]/414000;print(paste("R(inf):mean =",mean(x),"sd =",sd(x),"CI
=",quantile(x,0.025),",",quantile(x,0.975)))
x=rep(0,20000)
#Print 95% CI for consultation rates
for(k in 24:24)
{
 d=Hist$delta1
 x=x+.17*Hist$I[k,]*d
}
print(paste(k,mean(x),quantile(x,0.025),quantile(x,0.975)))

#Print mean and variance of number of infections
pr0=function(Hist,n=1000)
{
 i=round(Hist$t/1)
 b=Hist$beta[i,]
 g=Hist$gamma[i,]
 prec=1-exp(-1/g)
 INFS=c()
 for(j in 1:length(b))
  {
   dinf=1+rgeom(n,prec)
   rinf=b*dinf
   infs=rpois(n,rinf)
   INFS[j]=mean(infs)
  }
  print(paste("Mean",round(mean(INFS),3),"Standard deviation",round(sd(INFS),3)))
}
pr0(Hist,n=1000)
prt=function(i,Hist,n=1000)
{
 b=Hist$beta[i,]
 g=Hist$gamma[i,]
 S=Hist$S[i,]
 prec=1-exp(-1/g)
```

```
 INFS=c()
 for(j in 1:length(b))
  {
   dinf=1+rgeom(n,prec)
   rinf=b*dinf
   infs=rpois(n,rinf)
   INFS[j]=mean(infs)
  }
 INFS=INFS*S/Hist$N#to get Rt distn
 INFS
}

#Initialize matrices to store statistics
ma=list()
ma$beta=matrix(0,MAXDAYSTOREAD,3)
ma$epsilon=matrix(0,MAXDAYSTOREAD,3)
ma$lambda=matrix(0,MAXDAYSTOREAD,3)
ma$gamma=matrix(0,MAXDAYSTOREAD,3)
ma$falseflu=matrix(0,MAXDAYSTOREAD,3)
ma$delta1=matrix(0,MAXDAYSTOREAD,3)
ma$Rt=matrix(0,MAXDAYSTOREAD,3)

#Store and print 95% CI for parameters
for(i in 1:MAXDAYSTOREAD)
{
 x=Hist$beta[i,];ma$beta[i,2]=mean(x);ma$beta[i,c(1,3)]=quantile(x,c(0.025,0.975))
 x=Hist$epsilon[i,];ma$epsilon[i,2]=mean(x);ma$epsilon[i,c(1,3)]=quantile(x,c(0.025,0.975))
 x=Hist$lambda[i,];ma$lambda[i,2]=mean(x);ma$lambda[i,c(1,3)]=quantile(x,c(0.025,0.975))
 x=Hist$gamma[i,];ma$gamma[i,2]=mean(x);ma$gamma[i,c(1,3)]=quantile(x,c(0.025,0.975))
 x=Hist$falseflu[i,];ma$falseflu[i,2]=mean(x);ma$falseflu[i,c(1,3)]=quantile(x,c(0.025,0.975))
 x=Hist$delta1[i,];ma$delta1[i,2]=mean(x);ma$delta1[i,c(1,3)]=quantile(x,c(0.025,0.975))
}

for(i in 1:MAXDAYSTOREAD)
{
 print(paste(i,"of",MAXDAYSTOREAD))
 x=prt(i,Hist,n=1000);ma$Rt[i,2]=mean(x);ma$Rt[i,c(1,3)]=quantile(x,c(0.025,0.975))
}

#Plot the number of ILIs reported daily per private doctor
x=1:MAXDAYSTOREAD
todaysreports=dataset$ILI_priv[x]
plot(x,todaysreports,type='l',col=8)

plotmats=function(y,ylm,yla,prio="",RT=FALSE)
{
 x=1:MAXDAYSTOREAD
 xlm=range(x)

plot(x,todaysreports*0.75*ylm[2]/max(todaysreports),type='l',col=grey(0.85),ylab=yla,ylim=yl
m,xlim=xlm,yaxt='n',xaxt='n',xlab='')
```

```
 if(ylm[2]==12)axis(2,at=c(0,5,10),las=1)
 if(ylm[2]==10)axis(2,at=c(0,2,4,6,8,10),las=1)
 if(ylm[2]==1)axis(2,at=c(0,0.2,0.4,0.6,0.8,1),las=1)
 if(ylm[2]==500)axis(2,at=c(0,100,200,300,400,500),las=1)

 mxt=MAXDAYSTOREAD
 mo_lab=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
 for(m in
7:9){mos=dataset$date.m[1:mxt];q=mos==m;if(sum(q)>0){xlo=mean((1:mxt)[q==TRUE]);axis(1
,at=xlo,labels=mo_lab[m],line=-0.5,tick=FALSE)}}

 if(RT)lines(xlm,c(1,1),col=2)
 ltys=c(2,1,2)
 for(k in 1:3)lines(x,y[,k],lty=ltys[k])
 text(xlm[2],ylm[2],prio,adj=c(1,1.5))
}

plotmats(ma$Rt,c(0,10),"Rt",prio="(a)",RT=TRUE)




Nbounds=20
probs=seq(0.025,0.975,length.out=(Nbounds))


##ILIs per GP
CI=matrix(0,predictuntil,Nbounds)
sa = sample(1:Hist$n_particles,20000,replace=TRUE)
Ancestors=sa
errorsample=0
for(t in predictuntil:1)
{
 ns=Hist$I[t,Ancestors]
 p1 = Hist$delta1[t,]

 p2 = 1 #propn going to private
 p3 = 1/300 #propn cases from private practice making it into data
 ps = p1*p2*p3
 xs = ns*ps#rbinom(length(ns),ns,ps)

  h=0.9;xbar=mean(xs);sigma=var(xs);x_new=rnorm(length(xs),0,sqrt(sigma))*sqrt(1-h*h)
  x_new=x_new+xbar+h*(xs-xbar);x_new=pmax(x_new,rep(0,length(x_new)))
  x_new=x_new+Hist$falseflu[t,Ancestors]*p1
  if(t>DAYSTOREAD & t<=(DAYSTOREAD+52))
  {
   target=(dataset$ILI_priv/dataset$ndr_priv)[t]
   diffs=abs(x_new-target)
   errorsample=errorsample+mean(diffs)
   #print(mean(diffs))
  }
```

```
 CI[t,]=quantile(x_new,probs,na.rm=TRUE)
 if(t<predictuntil)Ancestors=Hist$parent[t,Ancestors]
}
write.table(round(CI,4),paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code","/output/v1_",DAYSTOREAD,".txt",sep=""),col.names=FALSE,row.names=FALSE)
write.table(errorsample,paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code","/output/error_",DAYSTOREAD,".txt",sep=""),col.names=FALSE,row.names=FALSE)


##Total ILIs if every day were Monday
CI=matrix(0,predictuntil,Nbounds)
sa = sample(1:Hist$n_particles,20000,replace=TRUE)
Ancestors=sa
for(t in predictuntil:1)
{
 ns=Hist$I[t,Ancestors]
 ps = Hist$delta1[t,]
 xs = rbinom(length(ns),ns,ps)
 #xs[is.na(xs)]=0

h=0.9;xbar=mean(xs,na.rm=TRUE);sigma=var(xs,na.rm=TRUE);x_new=rnorm(length(xs),0,sqrt(
sigma))*sqrt(1-h*h)
   x_new=x_new+xbar+h*(xs-xbar)
   x_new=pmax(x_new,rep(0,length(x_new)))
 CI[t,]=quantile(x_new,probs)
 if(t<predictuntil)Ancestors=Hist$parent[t,Ancestors]
}
write.table(round(CI,2),paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code","/output/v2_",DAYSTOREAD,".txt",sep=""),col.names=FALSE,row.names=FALSE)


##Cumulative total
CASES=matrix(0,predictuntil,n_particles)
CI=matrix(0,predictuntil,Nbounds)
sa = sample(1:Hist$n_particles,20000,replace=TRUE)
Ancestors=sa
for(t in predictuntil:1)
{
 ns=Hist$I[t,Ancestors]+Hist$R[t,Ancestors]
 ns=Hist$I[t,]+Hist$R[t,]
 #ns[is.na(ns)]=0
 h=0.9;xbar=mean(ns);sigma=var(ns);x_new=rnorm(length(ns),0,sqrt(sigma))*sqrt(1-h*h)
  x_new=x_new+xbar+h*(ns-xbar)
  x_new=pmax(x_new,rep(0,length(x_new)))
 CI[t,]=quantile(x_new,probs)
 #CI[t,]=quantile(ns,probs,na.rm=TRUE)
 CASES[t,]=ns
 if(t<predictuntil)Ancestors=Hist$parent[t,Ancestors]
}

for(t in predictuntil:1)
```

```
{
 ns=CASES[t,]
 h=0.9;xbar=mean(ns);sigma=var(ns);x_new=rnorm(length(ns),0,sqrt(sigma))*sqrt(1-h*h)
  x_new=x_new+xbar+h*(ns-xbar)
  x_new=pmax(x_new,rep(0,length(x_new)))
 CI[t,]=quantile(x_new,probs)
}
write.table(round(CI,2),paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code","/output/v3_",DAYSTOREAD,".txt",sep=""),col.names=FALSE,row.names=FALSE)


errors=c()
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/read_data.r",sep=""))

#Read and store errors, posterior absolute deviation between predicted and observed
averages over one week period following the time forecast is made
for(DAYSTOREAD in 1:(MAXDAYSTOREAD-1))
{
 er=read.table(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/output/error_",DAYSTOREAD,".txt",sep=""),header=FALSE)
 errors[DAYSTOREAD]=as.numeric(er)
}


today=DAYSTOREAD
predictuntil=today+PREDICTIONSPAN
print(paste("  ...Time",today),quote=FALSE)

t_length=predictuntil
Hist$t=today
Hist=onestepahead(Hist)
Hist=loglikelihooder(Hist,dataset)
Hist=Reweighting(Hist)
Hist=Resampling(Hist)
Hist=KernelSmoothing(Hist)
for(i in today:(predictuntil-1))
{
 Hist=onestepahead(Hist)
 Hist$t=Hist$t+1
 print(paste("    ...predicting",Hist$t),quote=FALSE)
}

Initialization=function(n_particles=10,t_length=52,dataset)
{
 #Create a list containing state and parameter matrices, loglikelihood and weights
 Hist = list()
 Hist$t = 1
 Hist$n_particles=n_particles
 Hist$N=dataset$popn_size
 Hist$S=matrix(0,t_length,n_particles)
```

```
  Hist$E=matrix(0,t_length,n_particles)
  Hist$I=matrix(0,t_length,n_particles)
  Hist$R=matrix(0,t_length,n_particles)
  Hist$D=matrix(0,t_length,n_particles)
  Hist$loglikelihood=matrix(0,t_length,n_particles)

  #Tracking index assigned to each particle
  Hist$parent=matrix(0,t_length,n_particles)
  Hist$urparent=matrix(0,t_length,n_particles)

  Hist$weight=matrix(0,t_length,n_particles)

  for(i in 1:n_particles) Hist$weight[1,i]=1
  Hist$weight[1,]=Hist$weight[1,]/sum(Hist$weight[1,])

  Hist$beta = matrix(0,t_length,n_particles)      #probability of infection
  Hist$epsilon = matrix(0,t_length,n_particles)    #importation rate
  Hist$lambda = matrix(0,t_length,n_particles)     #infectious rate
  Hist$gamma = matrix(0,t_length,n_particles)      #recovery rate
  Hist$falseflu = matrix(0,t_length,n_particles)   #background rate of ppl having same
symptoms but not H1N1

  Hist$delta1 = matrix(0,t_length,n_particles)  #Week

  #Assign values to initial parameters
  Hist=populate.priors(Hist,1,1)

  Hist
}

library(MASS)

KernelSmoothing=function(H)
{
  h=0.7
  t=H$t
  H2=H

  #Construct transition matrix, find column mean and covariance matrix
  transmat=cbind(log(H$beta[t,]),
                 logit(H$delta1[t,]),
                 log(H$epsilon[t,]),
                 log(H$gamma[t,]),
                 log(H$lambda[t,]),
                 log(H$falseflu[t,]),
                 logit((1+H$E[t,])/(2+dataset$popn_size)),
                 logit((1+H$I[t,])/(2+dataset$popn_size)),
                 logit((1+H$R[t,])/(2+dataset$popn_size)))
  mn=colMeans(transmat)
  si=cov(transmat)
```

```
 #Implement kernel smoothing on state and parameter values
 eit=mvrnorm(dim(transmat)[1],0*mn,si)
 x_new=eit*sqrt(1-h*h)
 transmat2=transmat
 for(k in 1:dim(transmat)[2])transmat2[,k]=x_new[,k]+mn[k]+h*(transmat[,k]-mn[k])

 #Reassign smoothed parameters back into parameter matrices
 k=1;H2$beta[t,]=exp(transmat2[,k])
 k=2;H2$delta1[t,]=inv.logit(transmat2[,k])
 k=3;H2$epsilon[t,]=exp(transmat2[,k])
 k=4;H2$gamma[t,]=exp(transmat2[,k])
 k=5;H2$lambda[t,]=exp(transmat2[,k])
 k=6;H2$falseflu[t,]=exp(transmat2[,k])
 k=7;H2$E[t,]=round((2+dataset$popn_size)*inv.logit(transmat2[,k]))-1
 k=8;H2$I[t,]=round((2+dataset$popn_size)*inv.logit(transmat2[,k]))-1
 k=9;H2$R[t,]=round((2+dataset$popn_size)*inv.logit(transmat2[,k]))-1
 H2$S[t,]=dataset$popn_size-H2$E[t,]-H2$I[t,]-H2$R[t,]

 #If number of susceptibles are <0 or >total population,states assume un-smoothed values
 for(i in 1:length(H$beta[t,]))
 {
  REJECT=0
  if(H2$S[t,i]<0)REJECT=1
  if(H2$S[t,i]>dataset$popn_size)REJECT=1
  if(REJECT==1)
  {
   H2$S[t,i]=H$S[t,i]
   H2$E[t,i]=H$E[t,i]
   H2$I[t,i]=H$I[t,i]
   H2$R[t,i]=H$R[t,i]
   H2$beta[t,i]=H$beta[t,i]
   H2$delta1[t,i]=H$delta1[t,i]
   H2$epsilon[t,i]=H$epsilon[t,i]
   H2$gamma[t,i]=H$gamma[t,i]
   H2$lambda[t,i]=H$lambda[t,i]
   H2$falseflu[t,i]=H$falseflu[t,i]
  }
 }
 H2
}

loglikelihooder=function(H,dataset)
{
 t=H$t
 H$loglikelihood[t,]=0
 if(dataset$ndr_priv[t]>0)
 {
  #Use corresponding consultation rate for each day of the week
  p1 = H$delta1[t,]

  p2 = 1    #propn going to private doctors
```

```
   p3 = 1/300  #propn cases from private practice making it into data
   pall = p1*p2*p3

   #Poisson approximation
   meanrate = (H$I[t,]*pall+ H$falseflu[t,]*p1)*dataset$ndr_priv[t]
   H$loglikelihood[t,]=H$loglikelihood[t,]+dpois(dataset$ILI_priv[t],meanrate,log=TRUE)
  }
  H$loglikelihood[t,]=as.numeric(sub(-Inf,-20000,H$loglikelihood[t,]))  #just in case, but
shouldn't need
  H$loglikelihood[t,]=as.numeric(sub(NaN,-20000,H$loglikelihood[t,]))   #just in case, but
shouldn't need
  H$loglikelihood[t,]=pmax(H$loglikelihood[t,],-20000)
  if(mean(H$loglikelihood[t,])==-20000)print("WARNING! All particles have too few cases for
data")
  H
}

rootdir="C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook Code/output/predictions"
rootdir="C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/output/realtime_estimation/resubmission"

#Particle filter source codes
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/InitializationFunction.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/KernelSmoothingFunction.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/OneStepAheadFunction.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/Resampling.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/loglikelihooder.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/ReweightingFunction.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/populatepriors.r",sep=""))

seed=666
set.seed(seed)
library(boot)

#Number of data points to be input into model and number of days in future to be predicted
MAXDAYSTOREAD=33
PREDICTIONSPAN=2

source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/read_data.r",sep=""))
n_particles=20000
Hist=Initialization(n_particles=n_particles,t_length=PREDICTIONSPAN+MAXDAYSTOREAD,data
set)
```

```
#Get data, execute particle filtering routine, save output
for(DAYSTOREAD in 1:MAXDAYSTOREAD)
{
 print(paste("Day",DAYSTOREAD,"of",MAXDAYSTOREAD,": lambda =
",mean(Hist$lambda[DAYSTOREAD,])),quote=FALSE)

 source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/filtering.r",sep=""))
 source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook Code/dump.r",sep=""))
}

for(i in 1:10){if(dev.cur()!=1)dev.off()}

#Calculate and store prediction errors and posterior
MAXDAYSTOREAD=33
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook Code/error.r",sep=""))
source(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/calculate_posteriors.r",sep=""))

onestepahead=function(H)
{
 #takes one particle with history H, currently at time H$t, simulates
 #forward one step and evaluates the likelihood for the data at time t+1

 #Assign all state and parameter values at time t+1 to be the same as that at time t
 t=H$t
 n_particles = H$n_particles
 H$weight[t+1,]=H$weight[t,]
 H$loglikelihood[t+1,]=H$loglikelihood[t,]
 H$beta[t+1,]=H$beta[t,]
 H$epsilon[t+1,]=H$epsilon[t,]
 H$gamma[t+1,]=H$gamma[t,]
 H$tau[t+1,]=H$tau[t,]
 H$falseflu[t+1,]=H$falseflu[t,]
 H$lambda[t+1,]=H$lambda[t,]

 H$delta1[t+1,]=H$delta1[t,]

 H$parent[t+1,] = H$parent[t,]
 H$urparent[t+1,] = H$urparent[t,]

 H$S[t+1,] = H$S[t,]
 H$E[t+1,] = H$E[t,]
 H$I[t+1,] = H$I[t,]
 H$R[t+1,] = H$R[t,]

 #Use the parameters in time t to calculate state values in time t+1, under SEIR model
 prob.recover=1-exp(-1/H$gamma[t,])
 recoveries=rbinom(n_particles,H$I[t,],prob.recover)
 H$R[t+1,] = H$R[t+1,] + recoveries
 H$I[t+1,] = H$I[t+1,] - recoveries
```

```
  prob.infectious=1-exp(-1/H$lambda[t,])
  infectiousnesses=rbinom(n_particles,H$E[t,],prob.infectious)
  H$I[t+1,] = H$I[t+1,] + infectiousnesses
  H$E[t+1,] = H$E[t+1,] - infectiousnesses

  prob.infection=1-exp(-(H$epsilon[t,]+H$I[t,]*H$beta[t,])/H$N)
  infections=rbinom(n_particles,H$S[t,],prob.infection)
  H$E[t+1,] = H$E[t+1,] + infections
  H$S[t+1,] = H$S[t+1,] - infections

  H
}

populate.priors=function(Hist,starttime,currenttime)
{
#Generate initial parameter values from Normal distribution
  Hist$D[starttime,]=rep(dataset$ILI_priv[starttime],n_particles)
  Hist$beta[starttime,]=abs(rnorm(n_particles,2.0,2.0))
  Hist$epsilon[starttime,]=abs(rnorm(n_particles,80.0,60.0))
  Hist$lambda[starttime,]=abs(rnorm(n_particles,1.0,1.0))
  Hist$gamma[starttime,]=abs(rnorm(n_particles,1.0,1.0))
  Hist$falseflu[starttime,]=abs(rnorm(n_particles,1.0,0.25))
  ##Hist$q1[starttime,]=lm1$coefficient[1]
  ##Hist$q2[starttime,]=lm1$coefficient[2]

  #Generate parameters from Beta distribution
  tempa=15
  tempb=5
  Hist$delta1[starttime,] = rbeta(n_particles,tempa, tempb)

  Hist$parent[starttime,]=1:n_particles
  Hist$urparent[starttime,]=1:n_particles

  #Generate initial state values from Normal distribution
  minI=dataset$ILI_priv[1]
  Hist$E[starttime,]=round(abs(rnorm(n_particles,minI*37.5,minI*20)))
  Hist$I[starttime,]=round(abs(rnorm(n_particles,minI*37.5,minI*20)))
  Hist$R[starttime,]=rep(250000,n_particles)

  Hist$S[starttime,]=Hist$N-Hist$E[starttime,]-Hist$I[starttime,]-Hist$R[starttime,]

  #If not starting from first day, assign state values to be equal to previous state in the same
particle
  if(starttime>1)
  {
    for(i in 1:n_particles)
    {
      ancestor=i
      counter=currenttime
      for(counter in currenttime:(starttime+1)){ancestor=Hist$parent[counter,ancestor]}
```

```
      Hist$E[starttime,i]=Hist$E[starttime,ancestor]
      Hist$I[starttime,i]=Hist$I[starttime,ancestor]
      Hist$R[starttime,i]=Hist$R[starttime,ancestor]
      Hist$
S[starttime,i]=Hist$N-Hist$E[starttime,i]-Hist$I[starttime,i]-Hist$R[starttime,i]
   }
  }

  Hist
}

maxdays=52
options(warn=-1)

#Create "dataset" to store data, read the number of ILIs in private clinics and polyclinics
respectively
dataset=list()
v=read.table(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/ILIs.txt",sep=""),sep=',')
dataset$ndr_priv=v[[1]]
dataset$ILI_priv=v[[2]]
dataset$other_priv=v[[3]]
dataset$ndr_poly=v[[4]]
dataset$ILI_poly=v[[5]]
dataset$other_poly=v[[6]]
dataset$day=1:maxdays;rm(v)

#Read population size
v=read.table(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/population_size.txt",sep=""))
dataset$popn_size=v[[1]];rm(v)

#Read first day and store index for the day of week for first day in "dow"
v=read.table(paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook
Code/firstday.txt",sep=""))
dataset$date.day=v[[1]]
if(v[[1]]=="1" | v[[1]]=="1")dataset$dow=1
if(v[[1]]=="2" | v[[1]]=="2")dataset$dow=1
if(v[[1]]=="3" | v[[1]]=="3")dataset$dow=1
if(v[[1]]=="4" | v[[1]]=="4")dataset$dow=1
if(v[[1]]=="5" | v[[1]]=="5")dataset$dow=1
if(v[[1]]=="6" | v[[1]]=="6")dataset$dow=1
if(v[[1]]=="7" | v[[1]]=="7")dataset$dow=1
if(v[[1]]=="8" | v[[1]]=="8")dataset$dow=1
if(v[[1]]=="9" | v[[1]]=="9")dataset$dow=1
if(v[[1]]=="10" | v[[1]]=="10")dataset$dow=1
if(v[[1]]=="11" | v[[1]]=="11")dataset$dow=1
if(v[[1]]=="12" | v[[1]]=="12")dataset$dow=1
if(v[[1]]=="13" | v[[1]]=="13")dataset$dow=1
if(v[[1]]=="14" | v[[1]]=="14")dataset$dow=1
if(v[[1]]=="15" | v[[1]]=="15")dataset$dow=1
```

```
if(v[[1]]=="16" | v[[1]]=="16")dataset$dow=1
if(v[[1]]=="17" | v[[1]]=="17")dataset$dow=1
if(v[[1]]=="18" | v[[1]]=="18")dataset$dow=1
if(v[[1]]=="19" | v[[1]]=="19")dataset$dow=1
if(v[[1]]=="20" | v[[1]]=="20")dataset$dow=1
if(v[[1]]=="21" | v[[1]]=="21")dataset$dow=1
if(v[[1]]=="22" | v[[1]]=="22")dataset$dow=1
if(v[[1]]=="23" | v[[1]]=="23")dataset$dow=1
if(v[[1]]=="24" | v[[1]]=="24")dataset$dow=1
if(v[[1]]=="25" | v[[1]]=="25")dataset$dow=1
if(v[[1]]=="26" | v[[1]]=="26")dataset$dow=1
if(v[[1]]=="27" | v[[1]]=="27")dataset$dow=1
if(v[[1]]=="28" | v[[1]]=="28")dataset$dow=1
if(v[[1]]=="29" | v[[1]]=="29")dataset$dow=1
if(v[[1]]=="30" | v[[1]]=="30")dataset$dow=1
if(v[[1]]=="31" | v[[1]]=="31")dataset$dow=1
if(v[[1]]=="32" | v[[1]]=="32")dataset$dow=1
if(v[[1]]=="33" | v[[1]]=="33")dataset$dow=1
if(v[[1]]=="34" | v[[1]]=="34")dataset$dow=1
if(v[[1]]=="35" | v[[1]]=="35")dataset$dow=1
if(v[[1]]=="36" | v[[1]]=="36")dataset$dow=1
if(v[[1]]=="37" | v[[1]]=="37")dataset$dow=1
if(v[[1]]=="38" | v[[1]]=="38")dataset$dow=1
if(v[[1]]=="39" | v[[1]]=="39")dataset$dow=1
if(v[[1]]=="40" | v[[1]]=="40")dataset$dow=1
if(v[[1]]=="41" | v[[1]]=="41")dataset$dow=1
if(v[[1]]=="42" | v[[1]]=="42")dataset$dow=1
if(v[[1]]=="43" | v[[1]]=="43")dataset$dow=1
if(v[[1]]=="44" | v[[1]]=="44")dataset$dow=1
if(v[[1]]=="45" | v[[1]]=="45")dataset$dow=1
if(v[[1]]=="46" | v[[1]]=="46")dataset$dow=1
if(v[[1]]=="47" | v[[1]]=="47")dataset$dow=1
if(v[[1]]=="48" | v[[1]]=="48")dataset$dow=1
if(v[[1]]=="49" | v[[1]]=="49")dataset$dow=1
if(v[[1]]=="50" | v[[1]]=="50")dataset$dow=1
if(v[[1]]=="51" | v[[1]]=="51")dataset$dow=1
if(v[[1]]=="52" | v[[1]]=="52")dataset$dow=1

Resampling=function(H)
{
 #Resample the particles according to weights and then assign uniform weights to the new
sample
 H2=H
 t=H$t

 if(sd(Hist$weight[t,])/mean(Hist$weight[t,])>1)
 {
  resample=sample(1:H$n_particles,H$n_particles,replace=TRUE,prob=H$weight[t,])

  H2$S[t,]=H$S[t,resample]
  H2$E[t,]=H$E[t,resample]
```

```
   H2$I[t,]=H$I[t,resample]
   H2$R[t,]=H$R[t,resample]
   H2$D[t,]=H$D[t,resample]
   H2$beta[t,]=H$beta[t,resample]
   H2$epsilon[t,]=H$epsilon[t,resample]
   H2$gamma[t,]=H$gamma[t,resample]
   H2$lambda[t,]=H$lambda[t,resample]
   H2$falseflu[t,]=H$falseflu[t,resample]
   H2$loglikelihood[t,]=H$loglikelihood[t,resample]
   H2$parent[t,]=resample
   H2$urparent[t,]=H$urparent[t,resample]

   H2$delta1[t,]=H$delta1[t,resample]

   H2$weight[t,]=1/H$n_particles
  }
  H2
}

Reweighting=function(H)
{
  #Calculation of weights based on loglikelihoods
  t=H$t
  weightMax = max(H$loglikelihood[t,])
  H$weight[t,] = exp(H$loglikelihood[t,]-weightMax)
  cp = 0; cp = sum(H$weight[t,])
  H$weight[t,] = H$weight[t,] / cp
  H
}
```

Appendix D

Joint model

```
##ILIs per 8 GPs
CI=matrix(0,predictuntil,Nbounds)
CI1=matrix(0,predictuntil,Nbounds)
sa = sample(1:Hist$n_particles,20000,replace=TRUE)
Ancestors=sa
errorsample=0
for(t in predictuntil:1)
{
 ns=Hist$I[t,Ancestors]
 #ns1=772+4.49*Hist$I[t,Ancestors]
 p1 = Hist$delta1[t,]

 p2 = 1 #propn going to private
 p3 = 1/300 #propn cases from private practice making it into data
 ps = p1*p2*p3
 xs = ns*ps#rbinom(length(ns),ns,ps)
 #xs1 = ns1*ps#rbinom(length(ns1),ns1,ps)

  h=0.9;xbar=mean(xs);sigma=var(xs);x_new=rnorm(length(xs),0,sqrt(sigma))*sqrt(1-h*h)
  x_new=x_new+xbar+h*(xs-xbar);x_new=pmax(x_new,rep(0,length(x_new)))
  x_new=x_new+Hist$falseflu[t,Ancestors]*p1

# Creating the model based on the number of actual ILIs being reported by GPs
   x_new=x_new *dataset$ndr_priv[1:t]
   lm1=lm(dataset$other_priv[1:t]~dataset$ILI_priv[1:t])
   parA=lm1$coefficient[1]
   parB=lm1$coefficient[2]
   x_new1=parA+parB*x_new
   if(t>DAYSTOREAD & t<=(DAYSTOREAD+52))
   {
    target=(dataset$ILI_priv/dataset$ndr_priv)[t]
    target1=(dataset$other_priv/dataset$ndr_priv)[t]
    diffs=abs(x_new-target)
    diffs1=abs(x_new1-target1)
    errorsample=errorsample+mean(diffs)
    errorsample1=errorsample+mean(diffs1)
    #print(mean(diffs))
   }
 CI[t,]=quantile(x_new,probs,na.rm=TRUE)
 CI1[t,]=quantile(x_new1,probs,na.rm=TRUE)
 if(t<predictuntil)Ancestors=Hist$parent[t,Ancestors]
}
write.table(round(CI,4),paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook Code","/output/v1_",DAYSTOREAD,".txt",sep=""),col.names=FALSE,row.names=FALSE)
write.table(round(CI1,4),paste("C:/Users/Vincent/Desktop/Vincent - PhD/H1N1 - Cook Code","/output/v11_",DAYSTOREAD,".txt",sep=""),col.names=FALSE,row.names=FALSE)
```

Appendix E

The parameters of the Linear Regression model

(Chapter 4 – Joint model)

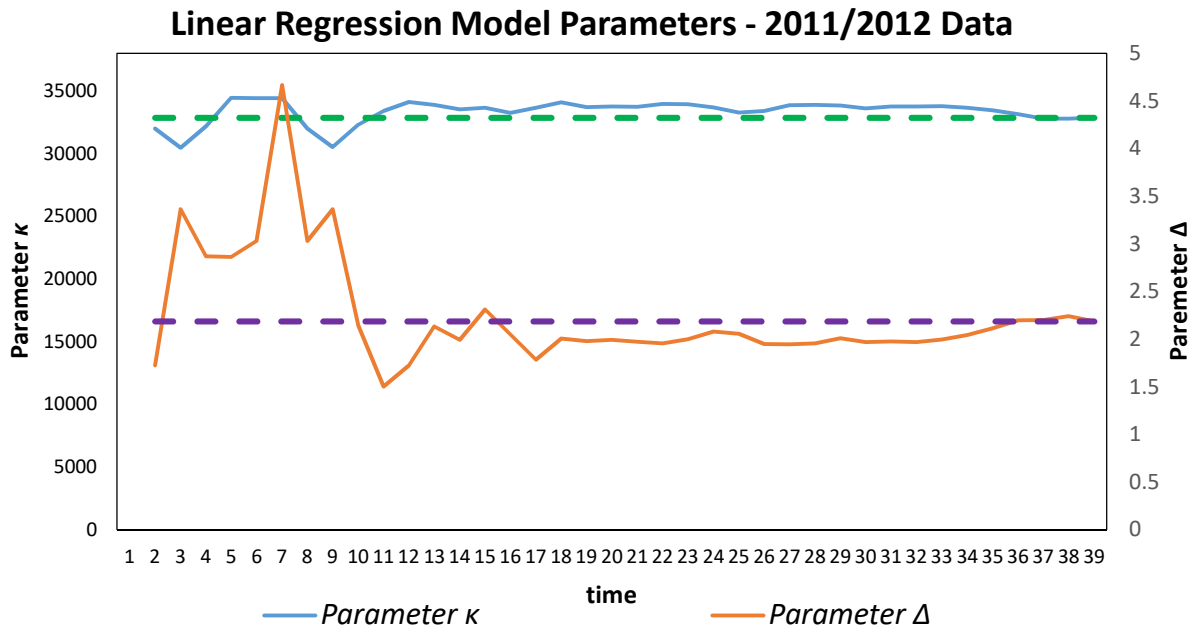


Figure E.1 –Parameter values for the linear regression between the weekly consultations (dependent variable) and the weekly diagnosed (independent variable) of the 2011/2012 outbreak. These parameters were updated at each individual time point during the course of the outbreak. The green dashed line is the general parameter κ for the above 2011/2012 linear regression model (Table 4.1) and the violet dashed line is the general parameter Δ for the above 2011/2012 linear regression model (Table 4.1).

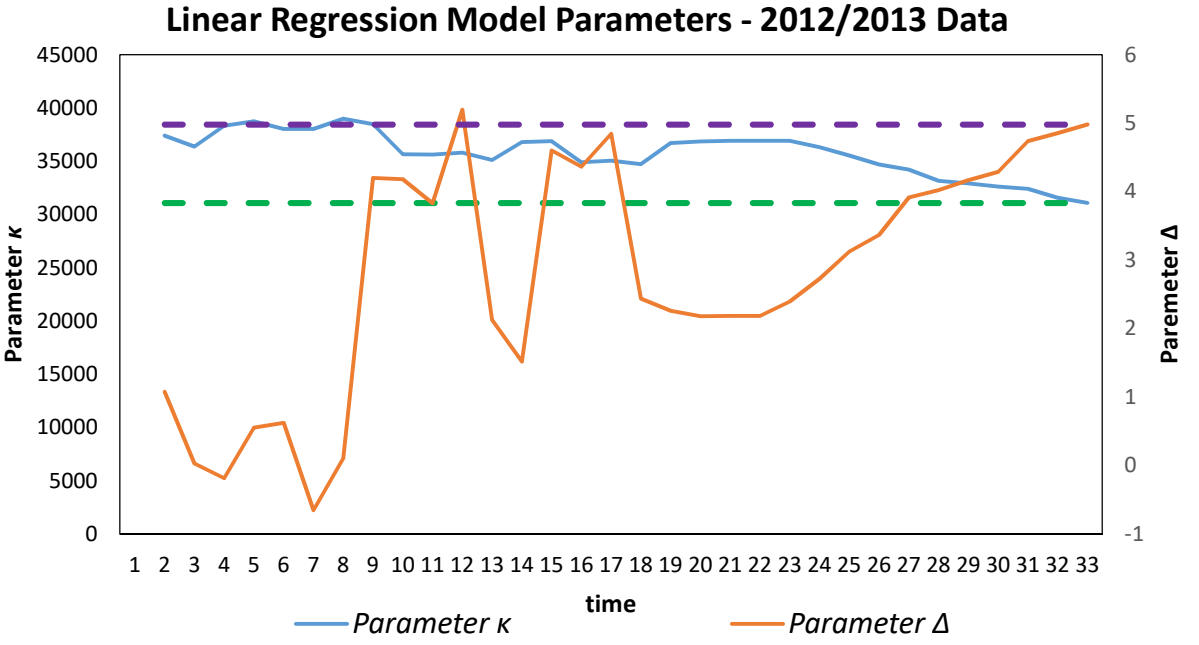


Figure E.2 –Parameter values for the linear regression between the weekly consultations (dependent variable) and the weekly diagnosed (independent variable) of the 2012/2013 outbreak. These parameters were updated at each individual time point during the course of the outbreak. The green dashed line is the general parameter κ for the above 2012/2013 linear regression model (Table 4.1) and the violet dashed line is the general parameter Δ for the above 2012/2013 linear regression model (Table 4.1).

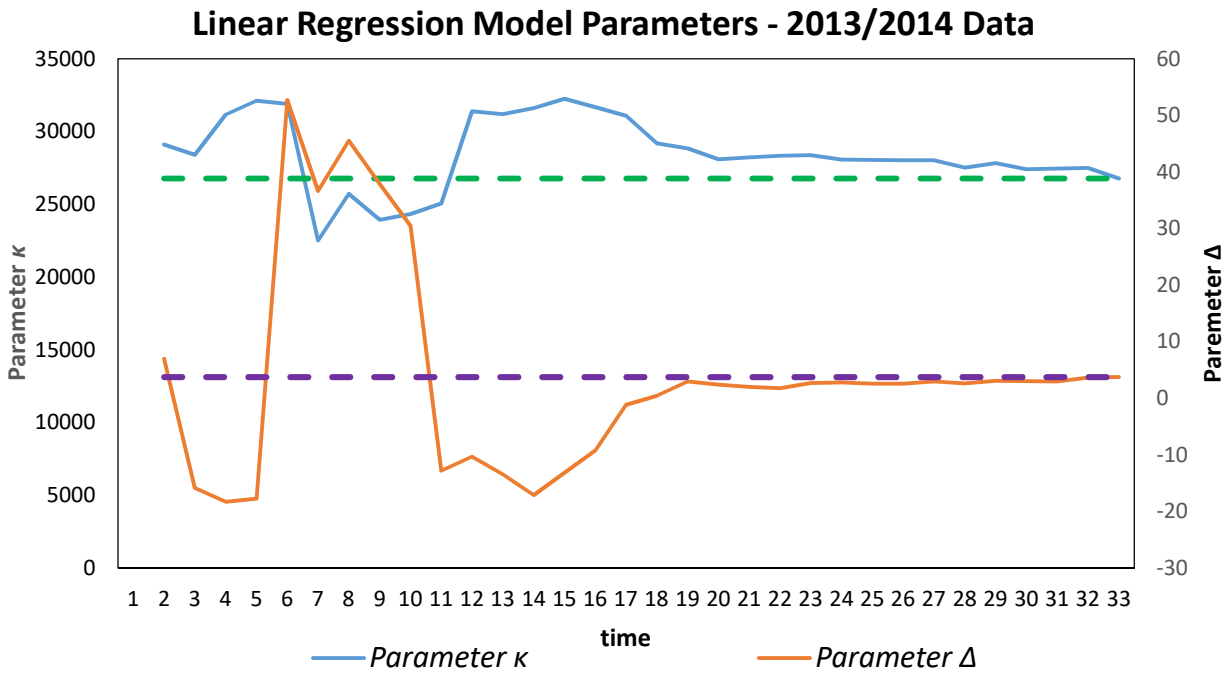


Figure E.3 –Parameter values for the linear regression between the weekly consultations (dependent variable) and the weekly diagnosed (independent variable) of the 2013/2014 outbreak. These parameters were updated at each individual time point during the course of the outbreak. The green dashed line is the general parameter κ for the above 2013/2014 linear regression model (Table 4.1) and the violet dashed line is the general parameter Δ for the above 2013/2014 linear regression model (Table 4.1).

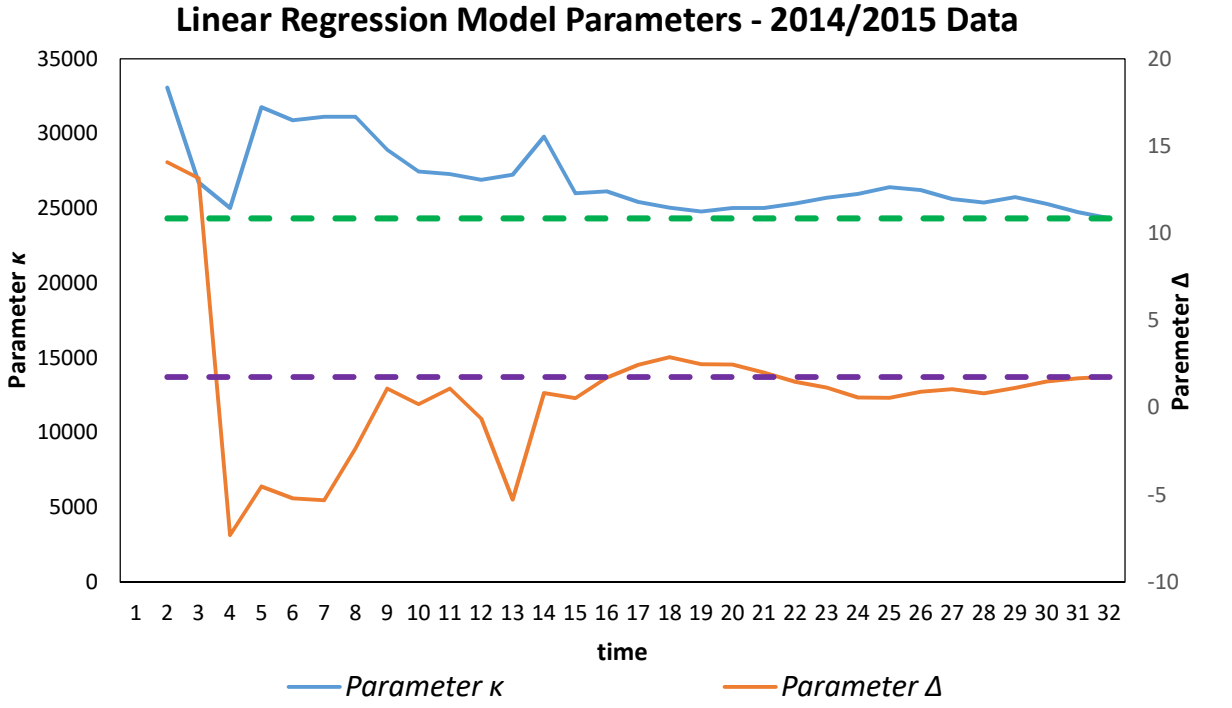


Figure E.4 –Parameter values for the linear regression between the weekly consultations (dependent variable) and the weekly diagnosed (independent variable) of the 2014/2015 pandemic outbreak. These parameters were updated at each individual time point during the course of the outbreak. The green dashed line is the general parameter κ for the above 2014/2015 linear regression model (Table 4.1) and the violet dashed line is the general parameter Δ for the above 2014/2015 linear regression model (Table 4.1).

Appendix F

Ethics form for the cross-sectional survey

PSYCHOLOGY DIVISION ETHICAL APPROVAL FORM

Check one box: ☐ STAFF project ☐ UNDERGRADUATE project ☒ POSTGRADUATE project

Title of project Understanding the under-reporting of the Seasonal Influenza

Name of Researcher(s): Vincent Marmara

Email Address: vam@cs.stir.ac.uk

Name of Supervisor(s) (for student research) Prof Adam Kleczkowski

Date: 28th July 2015

Postgraduate and Staff Projects

Please indicate your source of funding (Division, Research Council, Govt, Charity, etc)
Self-funded

		Yes	No	N/A
1	Will you tell participants that their participation is voluntary?	☒	☐	☐
2	Will you tell participants that they may withdraw at any time and for any reason?	☒	☐	☐
3	Will you obtain written consent for participation?	☐	☒	☐
4	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?	☒	☐	☐
5	If an experiment, will you describe the main experimental procedures to participants in advance, so that they are informed about what to expect?	☐	☐	☒
6	With questionnaires, will you give participants the option of omitting any questions they do not want to answer?	☒	☐	☐
7	If the research is observational, will you ask participants for their consent to being observed?	☐	☐	☒
8	Will you debrief participants at the end of their participation (i.e. give them a brief explanation of the study)?	☒	☐	☐

If you have ticked **No** to any of Q1-8, you should normally **tick box B** overleaf; if not, please give an explanation on a separate sheet.. [Note: N/A = not applicable]

			Yes	No	N/A
9	Will your project involve deliberately misleading participants in any way?		☐	☒	☐
10	Is there any realistic risk of you or any participants experiencing either physical or psychological discomfort, distress or harm?		☐	☒	☐
11	Will you be administering drugs or other substances to your participants, or taking fluid or other samples from them?		☐	☒	☐
12	Does your project involve work with animals?		☐	☒	☐
13	Do participants fall into any of the following special groups? If they do please refer to BPS guidelines, and **tick box B** overleaf. Note that you may also need to obtain satisfactory CRB clearance (or equivalent for overseas students)	Schoolchildren (under 18yrs)	☐	☒	☐
		People with learning or communication difficulties	☐	☒	☐
		Patients	☐	☒	☐
		People in custody	☐	☒	☐
		People engaged in illegal activities (e.g. drug taking)	☐	☒	☐

If you have ticked **Yes** to any of Q9 - 13 you should normally **tick box B** overleaf; if not, please give an explanation on a separate sheet.

DECLARATION

I am familiar with the BPS Guidelines for ethical practices in psychological research. I understand that there is an obligation on the lead researcher to bring to the attention of the Ethics Committee any issues with ethical implications not clearly covered by the checklist.

Please tick to confirm: ☒

PLEASE TICK **EITHER** BOX A OR BOX B BELOW AND **PROVIDE THE DETAILS REQUIRED** IN SUPPORT OF YOUR APPLICATION.

	Please tick
A. I consider that this project has **no** specific ethical implications to be brought before the Division Ethics Committee	☒
Give a brief description of participants and procedure (methods, tests used etc) in around 200 words.	

Checklist for a Part A submission:-	**Please tick**
• Project title	☐
• Number of participants and how they will be recruited	☐
• Start and end dates	☐
• Brief description of methods and measurements	☐
o Where participants will be tested	☐
o How materials will be administered	☐
o Any novel questions or questionnaires are included with submission	☐
o Length of time for each participant	☐
• Information/ consent form attached	☐
o Participant allowed to withdraw at any time	☐
o All individual data will be confidential	☐
• Debriefing form attached	☐

B. I consider that this project **may** have ethical implications that should be brought before the Division committee, and/or it will be carried out with children or other vulnerable populations	☒
Please provide details on a separate sheet.	

Checklist for a Part B submission:-	**Please tick**
• Project title	☒
• Purpose of project and its academic rationale	☒
• Number of participants (age, gender, exclusion/inclusion criteria) and how they will be recruited	☒
• Start and end dates	☒
• Brief description of methods and measurements	☒
o Where participants will be tested	☒
o How materials will be administered	☒
o Any novel questions or questionnaires are included with submission	☒
o Length of time for each participant	☒
• Information/ consent form attached (***this is not required as data collection is conducted by telephone interview, further details in 'Project Summary'***)	☐
o Participant allowed to withdraw at any time	☒
o All individual data will be confidential	☒
• Debriefing form attached	☒
• A clear but concise statement of ethical considerations raised by the project and how you intend to deal with them.	☒

This form should be submitted by email to the Psychology Ethics Committee for consideration (psychethicssubs@stir.ac.uk). Please include the name of the applicant in the 'Subject' line of the email. Students should send the form to their supervisor who, after checking it, will forward it to the Psychology Ethics Committee.

Project Summary (Part B)

Understanding the under-reporting of the Seasonal Influenza

Vincent Marmara, PhD Student, University of Stirling

The main objective of this study is to understand the under-reporting of the Seasonal Influenza in Malta. During the Seasonal Influenza period, a number of people visit their GP to be tested for Seasonal Influenza. However, it is believed that a significant portion of the population still did not visit their GP to be examined (Marmara et. al., 2014; WHO, 2010; Ishak et al., 2011). Hence, further research is required to understand this important factor to be able to set better health strategies and to plan the appropriate interventions.

Why is this important to study and what are the benefits of the study for the whole population?

The last two decades have seen several large-scale epidemics of international importance, including human, animal, and plant epidemics (Fisher et. al., 2012). Notable among these are avian and swine influenza, SARS, foot-and-mouth disease, Dutch elm disease, citrus canker, sudden oak death, and rhizoctonia. There is therefore a pressing need to construct models that allow us to use all available information to predict an emerging outbreak and to control it as quickly and as efficiently as possible (Marmara et. al., 2014). Epidemic data sets are typically short and have unobserved compartments (Chong et. al., 2014). For example, when individuals are infected but do not show symptoms, it is usually impossible to estimate their number and locations. Even for patients that do exhibit symptoms, very often only a limited proportion of cases are noted by the authorities. Some locations or groups of individuals are also notoriously difficult to assess. However, in many cases we can gather auxiliary information from different sources, for example by conducting horizontal serological studies giving us a snapshot of information at a single time point but with much broader and detailed information than longitudinal studies carried over time (Laurie et. al., 2013). This research intends to improve the 'missing data' problem by acquiring further information about the actual extent of the number of influenza cases within the Maltese population. Hence, such results aim to improve the understanding of the spread of the seasonal influenza amongst the population

and thus will serve as a good basis to authorities to take the necessary steps to control the spread of influenza. Furthermore, this telephone survey will aim to improve the knowledge of symptoms and seasonal influenza amongst participants.

Methodology

In this study, a questionnaire was designed to include a number of influenza-related questions. In fact, the questionnaire includes several questions regarding whether participants had experienced the seasonal influenza and whether they had any particular symptoms. Furthermore, respondents are given a list of symptoms to evaluate whether they actually had experienced these symptoms during the past year, thus assessing to what extent citizens know the definition of seasonal influenza. These questions are then tested against several demographics and general information regarding the individual's characteristics. Before commencing the actual data collection the questionnaire will be tested on a small sample of 20 individuals to ensure that all questions are understandable and all replies are in-line with the above objective. Hence if required the questionnaire will be amended accordingly.

To ensure a good response rate, the study will be carried out through the use of telephone surveys. The interview will be conducted in Maltese, however if participants prefer to answer in English, this option will be available as well. The study will comprise a sample of 400 Maltese individuals from a population of around 349,724 individuals (National Statistics Office, Malta, 2015). Hence, the study will be carried out through a 95% confidence level and 4.9% confidence interval as shown below.

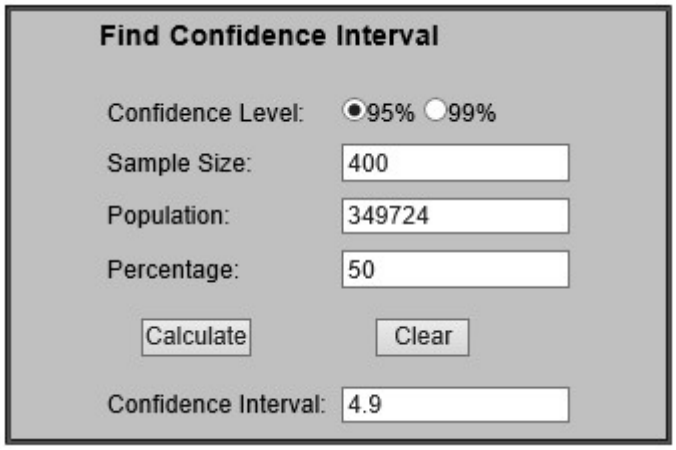


Figure 1 – Sample Size (Creative Research Systems, 2012)

The criteria for selection will include quota sampling by age, district and gender. Telephone numbers will be selected from the telephone directory using systematic sampling to ensure a representative sample of the Maltese population. As for the inclusion criteria for this study, only individuals of 18 years and older will be asked to reply to the questionnaire. Following an explanation of the main purpose of this research, individuals will be invited to participate in the study. They will be given the option to opt out from this research study at any time during the 5-minute telephone survey. Furthermore, they will be informed that their information will be kept confidential.

The following statement will be used at the starting point of each interview:

"Currently, I am carrying out a research about the seasonal influenza amongst Maltese Citizens as part of my PhD Study. Would you like to participate in this interview? This will only take 5 minutes of your time. All the information you provide will be treated in strict confidence and your identity will not be revealed at any point."

If the individual agrees to participate, the following statement will be:

"Thank you for deciding to participate in this study. Feel free to refuse to answer any questions or to terminate this interview at any point."

It is being planned that the data will be collected during August and September 2015 and analysis will be concluded by July 2016. During the whole research pathway, the researcher will ensure that all processes are being administered in line with this ethics application.

Hypothesis and main questions:

The main hypothesis of this research is:

> "The actual number of influenza cases amongst Maltese citizens is significantly higher when compared to the total number of influenza cases reported by GPs."

Since we are already in possession of seasonal influenza data from a sample of Maltese GPs, we will be able to compare the latter data with the new collected data. Moreover, this hypothesis will be compared with several variables as one can see from the attached questionnaire. In fact, the researcher will compare the above hypothesis with demographical variables, several questions related to the influenza symptoms and other

related influenza questions. Furthermore, data will be analyzed in a way to better understand several seasonality factors and hence this will serve as a good aid for influenza prediction modelling and to understand further the extent of the spread of seasonal influenza.

Analysis

Following data collection (August – September 2015), data analysis will be commenced followed by scientific models to elicit the most important factors through means testing, factor analysis and other scientific statistical techniques. Data analysis will be conducted in aggregated format only. It is estimated that this analysis will be concluded by July 2016. Throughout this research study, mainly R software will be used.

Costs

Throughout this survey, only costs for telephone calls will be incurred by the researcher as data collection will be conducted by the latter. It will only take approximately 5 minutes of the individual's time and hence this is the only envisaged burden for participants.

Local requirements and ethical considerations

In Malta, in order to conduct such a research study, it is not required to get an approval from a board since telephone numbers that are being used are public and taken randomly from the telephone directory. Moreover, individuals are free to opt out from this telephone interview or to refuse to answer any questions as explained above and in the 'debrief'. Additionally, no risks are envisaged throughout the study.

Furthermore, data will be analyzed in aggregated format and hence this research study will not be looking at data collected on a case by case basis.

References

A. Ishak , D. Tee, I. Nawmar, L.K. Pang, N. Ruslan, N. Che Mansor, L. Gam, H1N1 Influenza: A Viral Infection, WebmedCentral INFECTIOUS DISEASES, 2(12) (2011) WMC002736.

Creative Research Systems, 'The Survey System', 2012, available at: http://www.surveysystem.com/sscalc.htm

K. C. Chong, H. F. Fong, C. Y. Zee, Estimating the incidence reporting rates of new influenza pandemics at an early stage using travel data from the source country, Epidemiol. Infect. 142 (2014) 955-963.

K. L. Laurie, P. Huston, S. Riley, J. M. Katz, D. J. Willison, J. S. Tam, A. W. Mounts, K. Hoschler, E. Miller, K. Vandemaele, E. Broberg, M. D. Van Kerkhove, A. Nicoll, Influenza serological studies to inform public health action: best practices to optimise timing, quality and reporting, Influenza Other Respir Viruses 7(2) 2013 211-224.

M. C. Fisher, D. A. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, S. J. Gurr, Emerging fungal threats to animal, plant and ecosystem health, Nature 484 (2012) 186-194.

National Statistics Office, Malta (NSO), 'Malta in Figures 2014', 2014, available at: http://nso.gov.mt/en/publicatons/Pages/Publications-by-Date.aspx

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2010) URL http://www.R-project.org, ISBN 3-900051-07-0.

V. Marmara, A. Cook, A. Kleczkowski, Estimation of force of infection based on different epidemiological proxies: 2009/2010 Influenza epidemic in Malta, Epidemics 9 (2014) 52-61.

WHO (2010) Pandemic (h1n1) 2009 - update 100. World Wide Web electronic publication. URL http://www.who.int/csr/don/2010_05_14/en/index.html

Appendix G

Survey 2014/2015 results

Married

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	283	69.7	70.4	70.4
	No	119	29.3	29.6	100.0
	Total	402	99.0	100.0	
Missing	System	4	1.0		
Total		406	100.0		

Table G.1 – Respondents' marital status

Status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Employee	186	45.8	46.5	46.5
	Student	30	7.4	7.5	54.0
	Housewife	86	21.2	21.5	75.5
	Unemployed	11	2.7	2.8	78.3
	Pensioner	87	21.4	21.8	100.0
	Total	400	98.5	100.0	
Missing	System	6	1.5		
Total		406	100.0		

Table G.2 – Respondents' occupational status

Level of Education reached

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Primary	75	18.5	18.7	18.7
	Secondary	219	53.9	54.6	73.3
	Diploma	54	13.3	13.5	86.8
	Tertiary	53	13.1	13.2	100.0
	Total	401	98.8	100.0	
Missing	System	5	1.2		
Total		406	100.0		

Table G.3 – Respondents' level of education

Number of individuals at your household (including you)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	45	11.1	11.1	11.1
	2	100	24.6	24.7	35.8
	3	135	33.3	33.3	69.1
	4	100	24.6	24.7	93.8
	5	21	5.2	5.2	99.0
	6	3	.7	.7	99.8
	7	1	.2	.2	100.0
	Total	405	99.8	100.0	
Missing	System	1	.2		
Total		406	100.0		

Table G.4 – Respondents' number of individuals in their household

What is your main means of transport?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Walking	16	3.9	4.0	4.0
	Bike	7	1.7	1.7	5.7
	Motorbike	6	1.5	1.5	7.2
	Car	260	64.0	64.2	71.4
	Public transport	87	21.4	21.5	92.8
	Other	29	7.1	7.2	100.0
	Total	405	99.8	100.0	
Missing	System	1	.2		
Total		406	100.0		

Table G.5 – Respondents' main means of transport

Age Group * Did you receive the flu vaccine this winter/autumn season? (2014-2015) Crosstabulation

			Did you receive the flu vaccine this winter/autumn season? (2014-2015)			
			Yes	No	Don't know	Total
Age Group	18-25	Count	12	37	4	53
		% within Age Group	22.6%	69.8%	7.5%	100.0%
	26-35	Count	23	34	1	58
		% within Age Group	39.7%	58.6%	1.7%	100.0%
	36-45	Count	24	39	0	63
		% within Age Group	38.1%	61.9%	0.0%	100.0%
	46-55	Count	26	44	1	71
		% within Age Group	36.6%	62.0%	1.4%	100.0%
	56-65	Count	38	53	0	91
		% within Age Group	41.8%	58.2%	0.0%	100.0%
	66+	Count	51	17	1	69
		% within Age Group	73.9%	24.6%	1.4%	100.0%
Total		Count	174	224	7	405
		% within Age Group	43.0%	55.3%	1.7%	100.0%

Table G.6 – Respondents' flu vaccine uptake compared with their respective age group.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	49.860[a]	10	.000
Likelihood Ratio	49.034	10	.000
Linear-by-Linear Association	26.853	1	.000
N of Valid Cases	405		

a. 6 cells (33.3%) have expected count less than 5. The minimum expected count is .92.

Table G.7 – Chi-Square test of association between flu vaccine uptakes compared with their respective age group.

If 'No', why?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		16	7.1	7.1	7.1
	Forgot to take it	7	3.1	3.1	10.3
	I am afraid	54	24.1	24.1	34.4
	I am rarely sick	1	.4	.4	34.8
	I am very young	1	.4	.4	35.3
	I do not believe in it	1	.4	.4	35.7
	I do not care	5	2.2	2.2	37.9
	I do not know	1	.4	.4	38.4
	I do not need it	4	1.8	1.8	40.2
	I do not take it	5	2.2	2.2	42.4
	I do not want to take it	2	.9	.9	43.3
	I don't know	4	1.8	1.8	45.1
	I will feel sick after taking the vaccine	24	10.7	10.7	55.8
	Not good for me	7	3.1	3.1	58.9
	Not interested	92	41.1	41.1	100.0
	Total	224	100.0	100.0	

Table G.8 – Respondents' reasons for not taking the flu vaccine.

How many times did you visit your GP (doctor) during this past year?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	55	13.5	13.5	13.5
	1	75	18.5	18.5	32.0
	2	107	26.4	26.4	58.4
	3	66	16.3	16.3	74.6
	4	36	8.9	8.9	83.5
	5	26	6.4	6.4	89.9
	6	14	3.4	3.4	93.3
	7	12	3.0	3.0	96.3
	8	4	1.0	1.0	97.3
	9	1	.2	.2	97.5
	10	8	2.0	2.0	99.5
	12	1	.2	.2	99.8
	14	1	.2	.2	100.0
	Total	406	100.0	100.0	

Table G.9 – The number of times respondents visit their GP throughout the whole year.

Age Group * Do you take regular medication for any medical conditions such as asthma, diabetes, heart disorders, kidney disorder or other? Crosstabulation

			Do you take regular medication for any medical conditions such as asthma, diabetes, heart disorders, kidney disorder or other?		
			Yes	No	Total
Age Group	18-25	Count	9	44	53
		% within Age Group	17.0%	83.0%	100.0%
	26-35	Count	5	53	58
		% within Age Group	8.6%	91.4%	100.0%
	36-45	Count	14	49	63
		% within Age Group	22.2%	77.8%	100.0%
	46-55	Count	28	43	71
		% within Age Group	39.4%	60.6%	100.0%
	56-65	Count	49	42	91
		% within Age Group	53.8%	46.2%	100.0%
	66+	Count	62	7	69
		% within Age Group	89.9%	10.1%	100.0%
Total		Count	167	238	405
		% within Age Group	41.2%	58.8%	100.0%

Table G.10 – Respondents' frequency of regular medication compared with their respective age group.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	121.105[a]	5	.000
Likelihood Ratio	133.690	5	.000
Linear-by-Linear Association	103.396	1	.000
N of Valid Cases	405		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.85.

Table G.11 – Chi-Square test of association between the frequencies of regular medication compared with their respective age group.

Gender * Do you smoke? Crosstabulation

			Do you smoke?		
			Yes	No	Total
Gender	Female	Count	33	169	202
		% within Gender	16.3%	83.7%	100.0%
	Male	Count	71	128	199
		% within Gender	35.7%	64.3%	100.0%
Total		Count	104	297	401
		% within Gender	25.9%	74.1%	100.0%

Table G.12 – Cross tabulation between the frequencies of smokers compared with gender.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	19.523[a]	1	.000		
Continuity Correction[b]	18.529	1	.000		
Likelihood Ratio	19.867	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	19.475	1	.000		
N of Valid Cases	401				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 51.61.

b. Computed only for a 2x2 table

Table G.13 – Chi-Square test of association between the frequencies of smokers compared with gender.

If 'Yes', how many cigarettes per day?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	.2	1.0	1.0
	3	4	1.0	3.9	4.9
	4	2	.5	2.0	6.9
	5	2	.5	2.0	8.8
	6	2	.5	2.0	10.8
	7	3	.7	2.9	13.7
	10	23	5.7	22.5	36.3
	13	1	.2	1.0	37.3
	15	11	2.7	10.8	48.0
	20	43	10.6	42.2	90.2
	25	2	.5	2.0	92.2
	30	5	1.2	4.9	97.1
	40	3	.7	2.9	100.0
	Total	102	25.1	100.0	
Missing	System	304	74.9		
Total		406	100.0		

Table G.14 – The number of cigarettes respondents consume per day.

Symptoms Duration (Days)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Day	19	4.7	5.7	5.7
	2 Days	38	9.4	11.3	17.0
	3 Days	53	13.1	15.8	32.8
	4 Days	47	11.6	14.0	46.9
	5 Days	27	6.7	8.1	54.9
	6 Days	11	2.7	3.3	58.2
	7 Days	57	14.0	17.0	75.2
	9-14 Days	53	13.1	15.8	91.0
	15+ Days	30	7.4	9.0	100.0
	Total	335	82.5	100.0	
Missing	System	71	17.5		
Total		406	100.0		

Table G.15 – The number of days for the influenza-like-illness symptoms to persist according to survey respondents.

Seasonal Influenza Duration (Days)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Days	2	.5	1.8	1.8
	3 Days	7	1.7	6.2	8.0
	4 Days	10	2.5	8.8	16.8
	5 Days	6	1.5	5.3	22.1
	7 Days	48	11.8	42.5	64.6
	9-14 Days	27	6.7	23.9	88.5
	15+ Days	13	3.2	11.5	100.0
	Total	113	27.8	100.0	
Missing	System	293	72.2		
Total		406	100.0		

Table G.16 – The number of days for the seasonal influenza to persist according to survey respondents.

Nights at hospital due to the seasonal influenza:

If 'Yes' for how many nights?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	10	2.5	37.0	37.0
	3	2	.5	7.4	44.4
	4	2	.5	7.4	51.9
	5	1	.2	3.7	55.6
	7	3	.7	11.1	66.7
	8	1	.2	3.7	70.4
	9	1	.2	3.7	74.1
	10	1	.2	3.7	77.8
	14	6	1.5	22.2	100.0
	Total	27	6.7	100.0	
Missing	System	379	93.3		
Total		406	100.0		

Table G.17 – The number of days of hospitalisation due to seasonal influenza.

Did any of your household members had the seasonal influenza?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	211	52.0	54.5	54.5
	No	176	43.3	45.5	100.0
	Total	387	95.3	100.0	
Missing	System	19	4.7		
Total		406	100.0		

Table G.18 – Household members that had acquired the seasonal influenza.

If 'Yes', how many members?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	121	29.8	58.5	58.5
	2	65	16.0	31.4	89.9
	3	18	4.4	8.7	98.6
	4	2	.5	1.0	99.5
	6	1	.2	.5	100.0
	Total	207	51.0	100.0	
Missing	System	199	49.0		
Total		406	100.0		

Table G.19 – The number of members within the respondents' household that had acquired the seasonal influenza.

Appendix H

The SEIR model

Applying the SEIR Model in R without applying the particle filter algorithm:

```
 tn=29
  SS      <- numeric(tn)
  II      <- numeric(tn)
  EE      <- numeric(tn)
  RR      <- numeric(tn)
  AA      <- numeric(tn)
  DD      <- numeric(tn)

  SS[1]   <- 259560
  II[1]    <-  2220
  EE[1]   <-  2220
  RR[1]  <-  150000

  N = 414000

# previous posterior time-dependent average parameter values#
# considering the parameters from (t=2) when temperature drops#
v1=matrix((scan("SeasonalParValues.txt")),byrow=T,ncol=4)
# applying the observed model Dt#
v2=matrix((scan("SeasonalParValuesObs.txt")),byrow=T,ncol=2)


  for (t in 2:tn)
{
AA = (1-exp((-v1[t,1]-v1[t,4]*II[t-1])/N))
BB = (1-exp(-1/v1[t,2]))
CC = (1-exp(-1/v1[t,3]))

   A = rbinom(1, SS[t-1], AA)
   B = rbinom(1, EE[t-1], BB)
   C = rbinom(1, II[t-1], CC)

    SS[t]    <- SS[t-1] - A
    EE[t]    <- EE[t-1] + A - B
    II[t]     <- II[t-1] + B - C
    RR[t]    <- RR[t-1] + C
    DD[t]    <- v2[t,2]*(v2[t,1]*300+II[t])
  }

par(mfrow=c(3,2),mar=c(2,5,1,4))
plot(SS)
plot(EE)
plot(II)
plot(RR)
plot(DD)
```

Appendix I

Forecast of the spread of the seasonal influenza based on the SEIR model

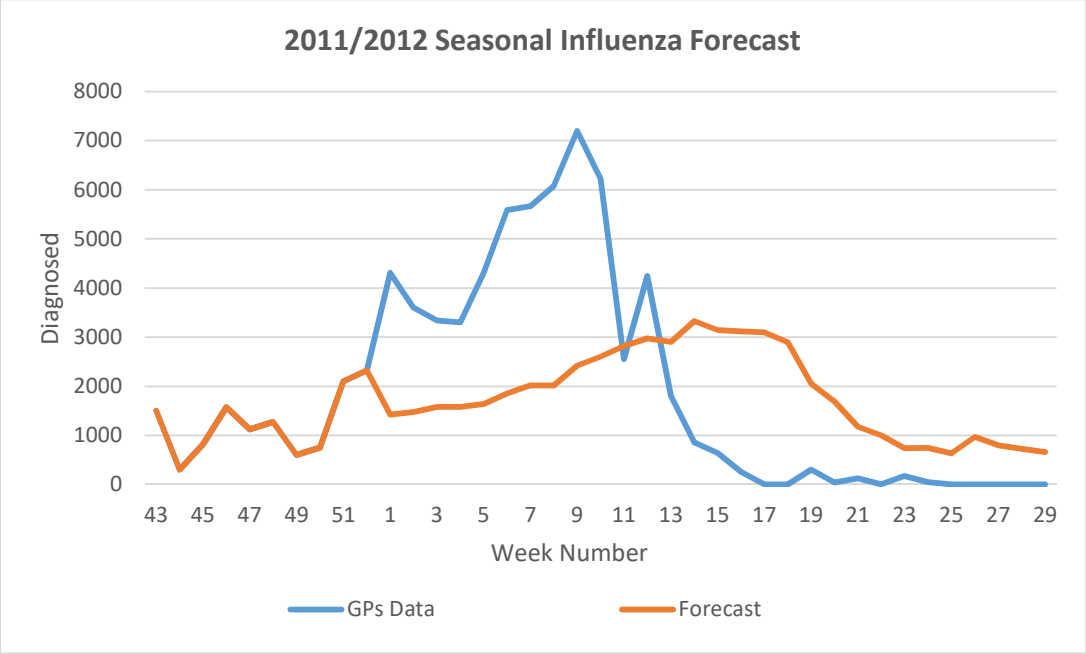


Figure I.1 - 2011/2012 diagnosed ILI forecasts through the SEIR model

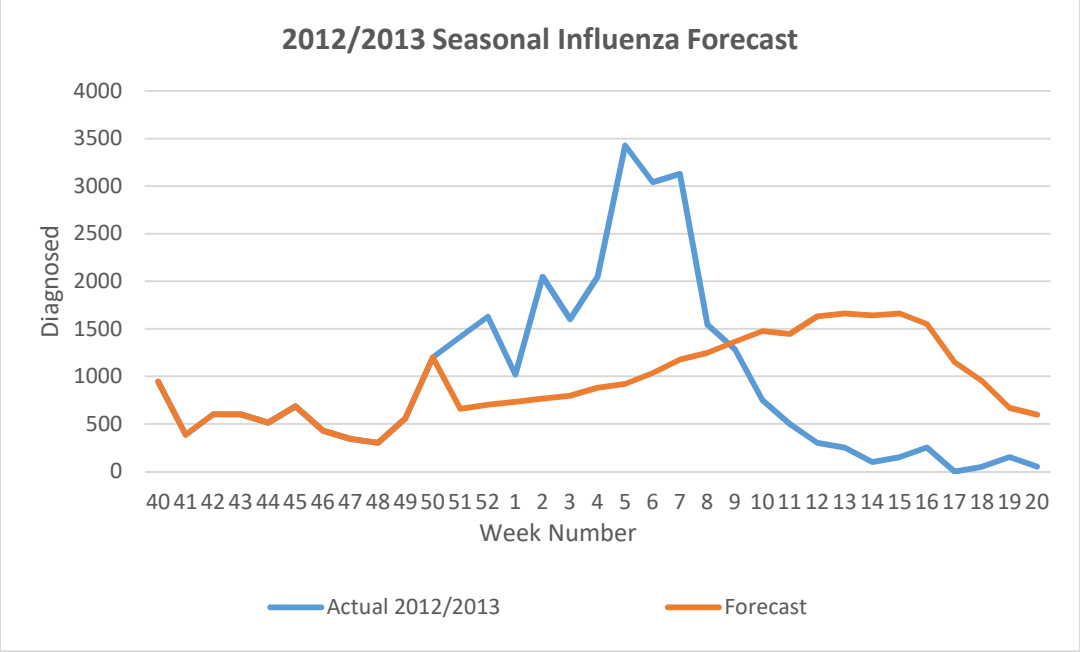


Figure I.2 - 2012/2013 diagnosed ILI forecasts through the SEIR model

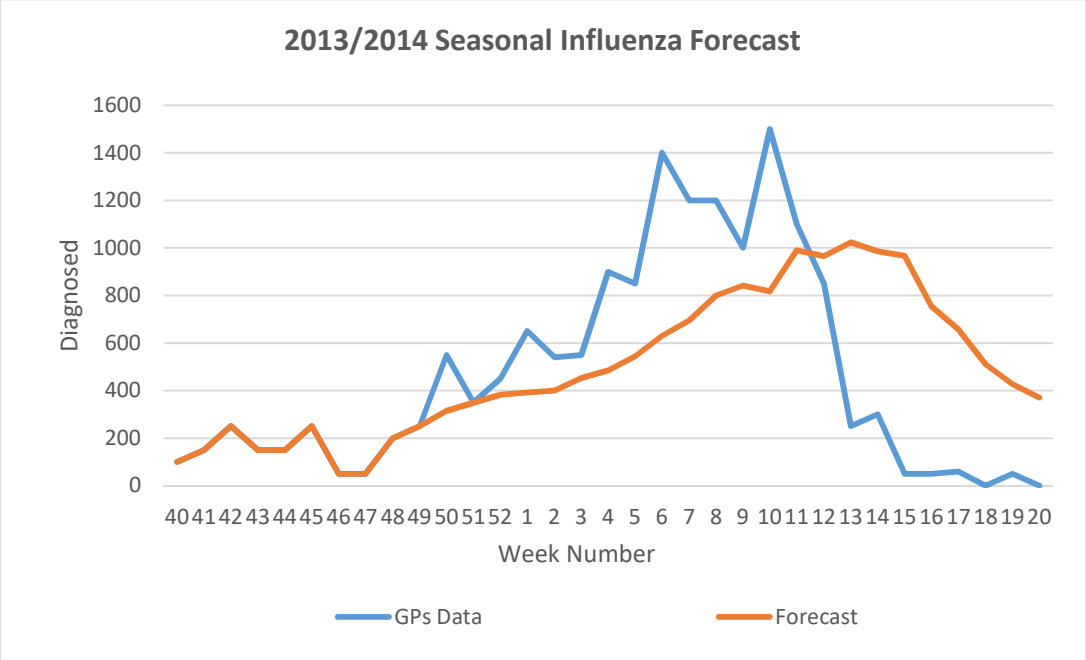


Figure I.3 - 2013/2014 diagnosed ILI forecasts through the SEIR model

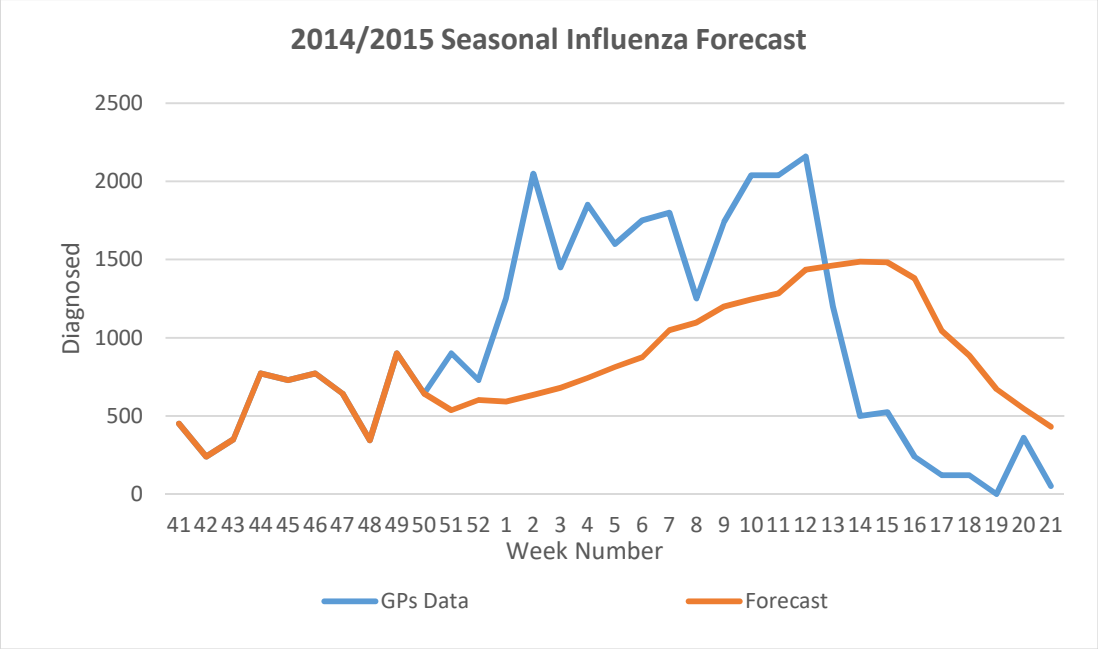


Figure I.4 - 2014/2015 diagnosed ILI forecasts through the SEIR model

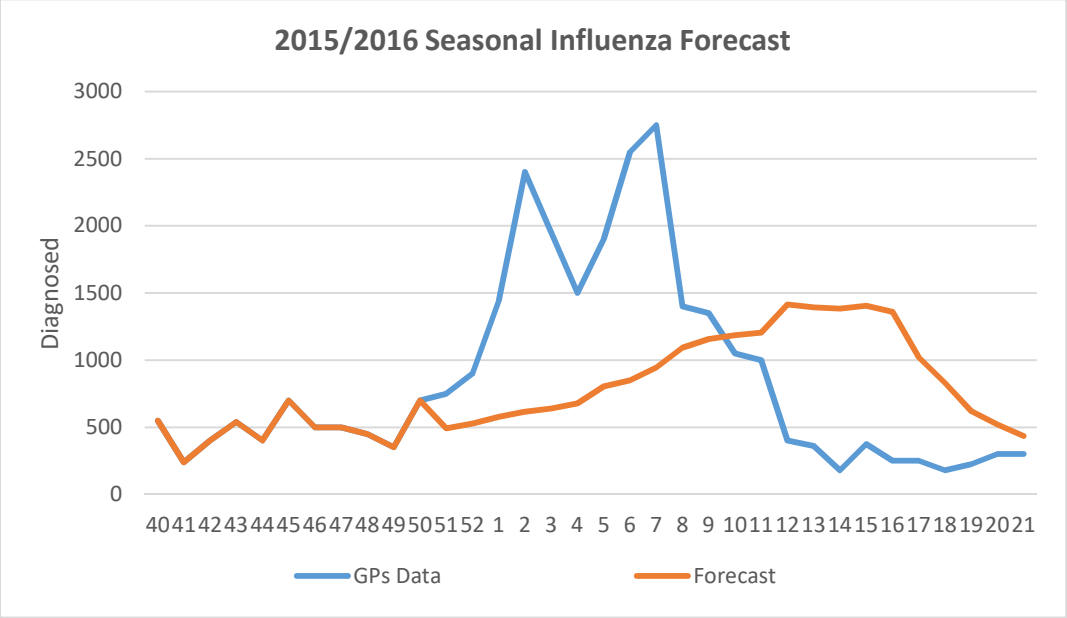


Figure I.5 - 2015/2016 diagnosed ILI forecasts through the SEIR model

Appendix J

My research paper as reported by the 'Times of Malta'

Statistician figures out epidemics risk

Claudia Calleja

A Maltese mathematician mapped out the way swine flu spread across the island in 2009 and 2010 and came up with a set of models that can help predict the outbreak of future epidemics.

The models, based on four data sets, allow for an early warning of an oncoming epidemic and for the prediction of the rate at which it will spread, Vincent Marmará, who is also a statistician, said.

This, he added, could help ensure timely action is taken and the right amount of vaccines bought.

Together with researchers Alex Cook and Adam Kleczkowski, Mr Marmará outlined the findings in a paper entitled 'Estimation of force of infection based on different epidemiological proxies: 2009/2010 influenza epidemic in Malta'. In the paper, published in the journal *Epidemics*, and forming part of his PhD in mathematics (epidemiology), he focuses on prediction and control of epidemic outbreaks based on limited information.

Mathematician Vincent Marmará came up with a set of models that predict epidemics.

The researchers resorted to data about influenza collected by the Malta Health Promotion Department during the swine flu (H1N1) epidemic when the first cases emerged in Malta in 2009. The researchers focused on data collected by eight physicians selected by the department between September 2009 and June 2010. During that period, 52,016 patients consulted the eight doctors and 4,544 were diagnosed with common influenza.

The amount of patients swabbed for swine flu amounted to 1,847 and those who tested positive reached 622. Five patients died.

Between January and February 2010, vaccine became available to everyone, so March of that year was considered to mark the end of the epidemic.

Equipped with such information the researchers drew up four data sets based on consultations, patients diagnosed (with the common flu), patients swabbed and positive cases (for swine flu).

They also took into consideration several factors including under-reporting.

Using established and up-to-date scientific modelling techniques, the four data sets were used to create four models, which could be applied to predict epidemic patterns in Malta. These models can be tailored to particular diseases by inserting certain parameters such as information about the duration of an illness.

"The innovative part of the research is that it allows for the use of the four different data sets to confirm results and predict the actual outbreak," Mr Marmará said.

So, for example, if a case of Ebola is detected in Malta, the amount of consultations – together with information about the disease drawn from the experience of other countries – can help predict the rate at which it will spread.

This can then be cross-checked mathematically with data about patients diagnosed with symptoms.

Overnight guest stays up 3% over past year

The number of people who stayed in hotels and guesthouses increased by three per cent in October, compared to October 2013.

Official data also shows that the total number of nights spent increased by two per cent, although the average length of stay decreased.

According to the National Statistics Office, arrivals in collective accommodation establishments exceeded 152,300 while the number of nights spent totalled more than 843,200.

All hotel categories, apart from those in the two-star category, recorded increases in total nights. The largest share of guest nights was reported in four-star hotels, accounting for about 414,800 nights, or 49 per cent of the total.

The average length of stay in collective accommodation establishments went down from 5.6 nights in October of last year to 5.5 nights this October.

Data for the first 10 months of the year show there were almost 1.4 million arrivals, up 7.2 per