Epidemics 9 (2014) 52–61

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

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

V. Marmara[a,*], A. Cook[b], A. Kleczkowski[a]

[a] University of Stirling, Stirling FK9 4LA, United Kingdom
[b] National University of Singapore, Singapore 119246, Singapore

ARTICLE INFO

Article history:
Received 22 August 2013
Received in revised form 2 September 2014
Accepted 29 September 2014
Available online 6 October 2014

Keywords:
Epidemiology
Compartmental modelling
Bayesian inference
Markov chain methods
Reproduction ratio

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.

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 (Chowell et al., 2011a). Shortly afterwards the influenza started to spread in the European countries (Flasche et al., 2011). 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 (Flasche et al., 2011). A lot of uncertainty about this influenza existed especially during the initial stages of the influenza, but the availability of data sets 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 (WHO, 2010; Ishak et al., 2011) 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. (2009). 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 (Reed et al., 2009).

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

* Corresponding author at: University of Stirling, Computing Science and Mathematics, Room 4B59, Cottrell Building Stirling FK94LA, United Kingdom. Tel.: +35699891390.
E-mail addresses: vam@cs.stir.ac.uk, vincentmarmara@gmail.com (V. Marmara).

http://dx.doi.org/10.1016/j.epidem.2014.09.010

for epidemics and in particular on the estimation of the effective reproduction ratio, R_t (Ong et al., 2010; Chowell et al., 2006, 2011b; Griffin et al., 2011; Boëlle et al., 2009; Fraser et al., 2009; Hsieh et al., 2011; Clancy, 2008; White et al., 2009; Katriel et al., 2011). In order to do so, for the case of the H1N1, several papers considered and compared different datasets coming from different states and countries (Chowell et al., 2011a; Flasche et al., 2011; Flahault et al., 2009; Opatowski et al., 2011; Kenah et al., 2011).

Parameter estimation for epidemiological models has so far been mostly based upon positive cases of H1N1 (laboratory-tested-positive) (Flasche et al., 2011; Hsieh et al., 2011; Buckley and Bulger, 2011; Nishiura et al., 2009; Chang et al., 2010) although some analyzed swabbed cases (Influenza-Like-Illness) (Correia et al., 2010; Rizzo et al., 2010) and others compared swabbed and positive cases (Chowell et al., 2011a; Opatowski et al., 2011). Many datasets were analysed with resolution varying from weekly reporting (Rizzo et al., 2010; Yu et al., 2012) to daily datasets (Flasche et al., 2011; Chowell et al., 2011b).

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. 6000 in December to as much as ca. 50,000 in August. Malta is a densely populated country with circa 1311 inhabitants per square kilometre.

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 (Fig. 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 °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 4544 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 increase 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 °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 °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.

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 know 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.

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 (Ong et al., 2010; Anderson and May, 1991) was used to estimate the parameters. The model includes four compartments, susceptible (S), exposed (E) (infected but not infectious), infectious (I) and

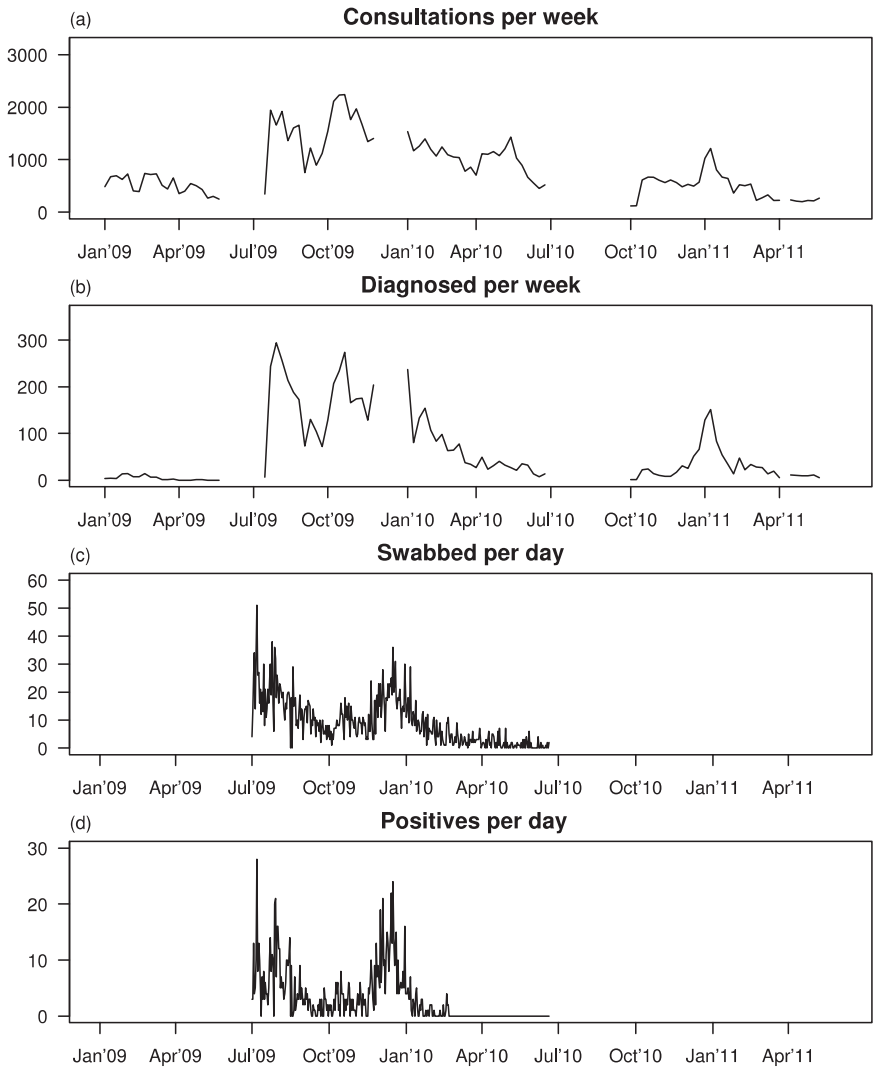

Fig. 1. The epidemiological data from Malta covering the period from January 2009 to May 2011. Consultations and diagnosed were reported weekly by 8 sentinel doctors selected by MHPD. During the H1N1 epidemic, data were collected daily for swabbed and positive patients from risk groups; data collected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

recovered (R). The SEIR model describes the flow of individuals between the compartments

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

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:

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

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 $\theta = \left(\beta, \varepsilon, \alpha, \tau\right)$ and the current state $\Sigma_t = \left\{S_t, E_t, I_t, R_t\right\}$ are unknown. Observations, D_t, are assumed to be Poisson distributed with mean $N_t \delta_{d(t)} \left\{\phi + (I_t/300)\right\}$ 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:

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

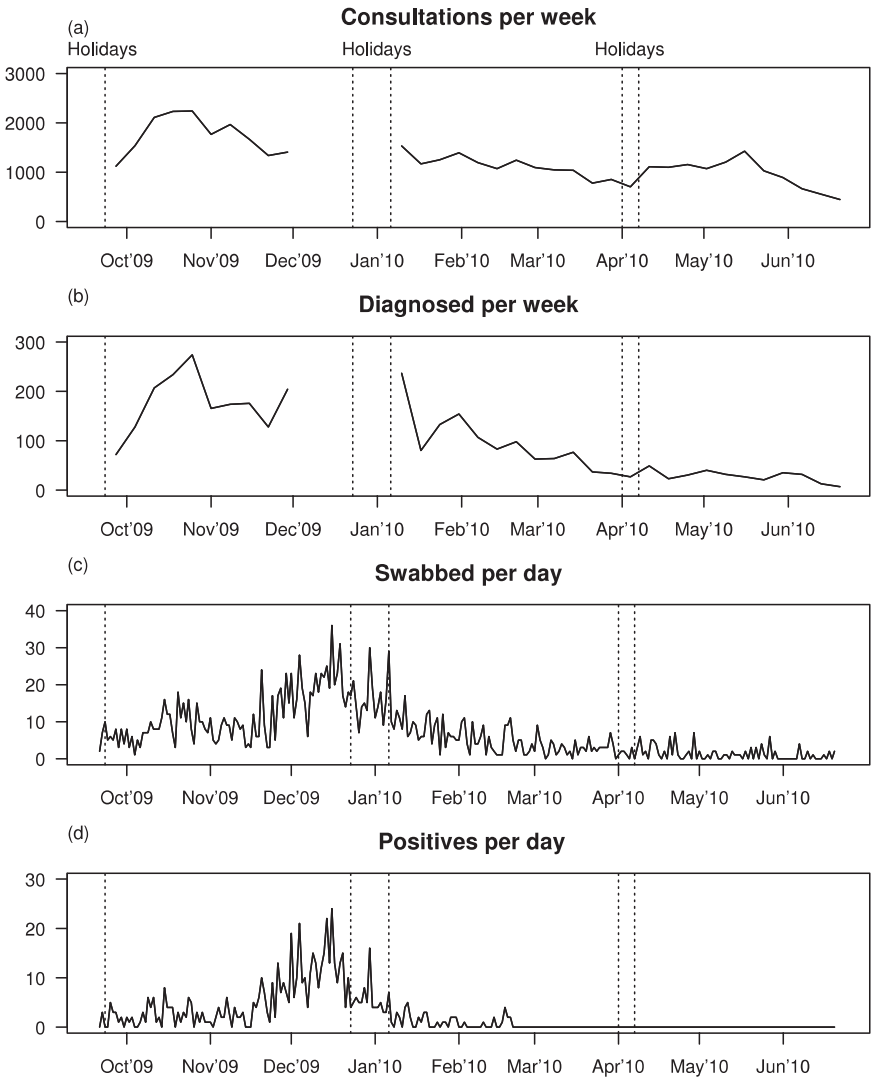

Fig. 2. Malta influenza data used in the analysis. The dotted lines denote Malta's holidays; no apparent correlation with holidays was found in the data.

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 (Doucet et al., 2000, 2001) 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 (Ong et al., 2010) and we only provide a short summary here.

The particle filter algorithm (Doucet et al., 2000, 2001) 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 Ong et al. (2010) 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, were $x_t^p = (\Sigma_t, \vartheta)$ 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: $w_{t+1}^{*(p)} = w_{t+1}^{*(p)} / \sum_P^{q=1} w_{t+1}^{*(q)}$.

Re-sampling (Doucet et al., 2001) 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 $\{1, 2, \ldots, P\}$ 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 (Ong et al., 2010; Trenkel et al., 2000). 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. (2010) and were generally very broad. For the daily data sets 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. (2010). 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.

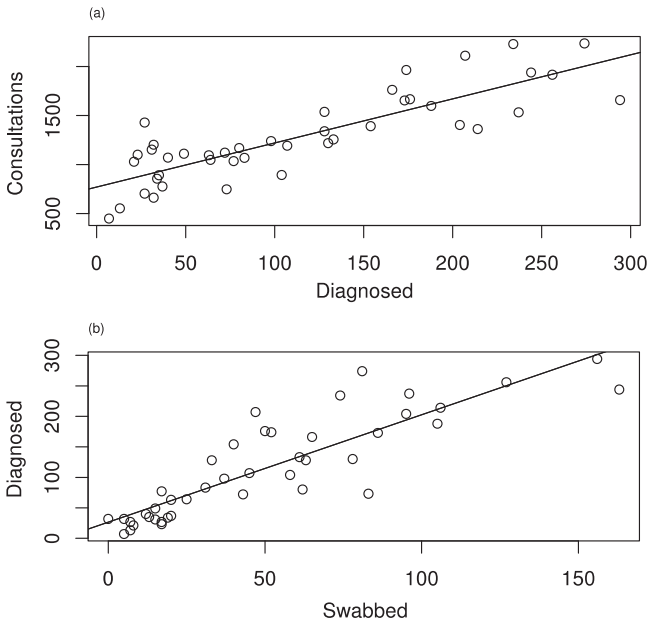

Fig. 3. Relationship between the number of consultations and diagnosed (a) and the number of diagnosed and swabbed (b) over the period shown in Fig. 2. Lines of best least-squares fit are used to 'reconstruct' the missing data. *Consultations = 772.32 + 4.49 (diagnosed)*, $R^2 = 0.76$ and *diagnosed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

Simulation parameters

The performance of the simulations depends on the size of the data sets. The memory and time constraints limit the number of particles that can realistically be used for large data sets. 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 (R Development Core Team, 2010) 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 following. 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, Fig. 2(a) and (b), and the relationship can be approximated by a linear function ($R^2 = 0.71$), Fig. 3(a) (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.

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 (Hsieh et al., 2011; Kenah et al., 2011; Correia et al., 2010; Poletti et al., 2011; Omori and Nishiura, 2011; Nishiura, 2011; Fierro, 2011). 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 (Fig. 2c and d), some of which can be associated with the day of the week, see Fig. 5.

The model successfully represents the main features of all data sets, both for the weekly data sets (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

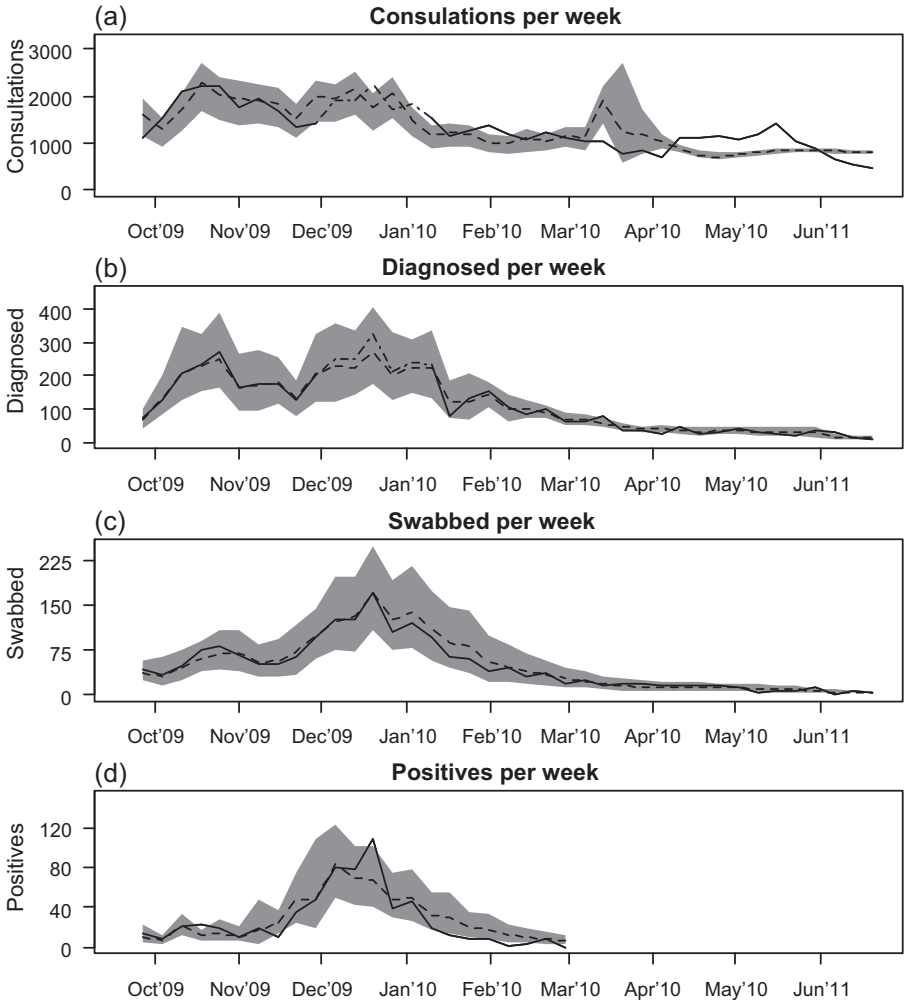

Fig. 4. Comparison of weekly (consultations, (a), and diagnosed, (b)) and weekly-aggregated (swabbed, (c), and positive, (d)) data, solid line, with the results of model fit, dashed line (mean) and shaded area (95% high predictive density regions). The 'reconstructed' data for consultations and diagnosed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

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.

The estimates of individual parameters vary widely among different data sets 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 data sets for the weekly sampling, Fig. 6. They are also consistent with other data sets available in the literature, for example see Ong et al. (2010). 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.

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. (2010) estimates. The estimates for τ based on daily data are more consistent with Ong et al. (2010) (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. (2010). Finally, the background consultation rate is high for the consultations data reflecting the need for accounting for non-ILI patients, whereas for other data sets 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 date. Unfortunately, this situation is rare in human epidemiology of influenza and similar diseases as we always struggle with incomplete data coming

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.

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 (45.03–1118.06)	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

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 data set from the 2009–2010 H1N1 epidemic in Malta. The SIR 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.

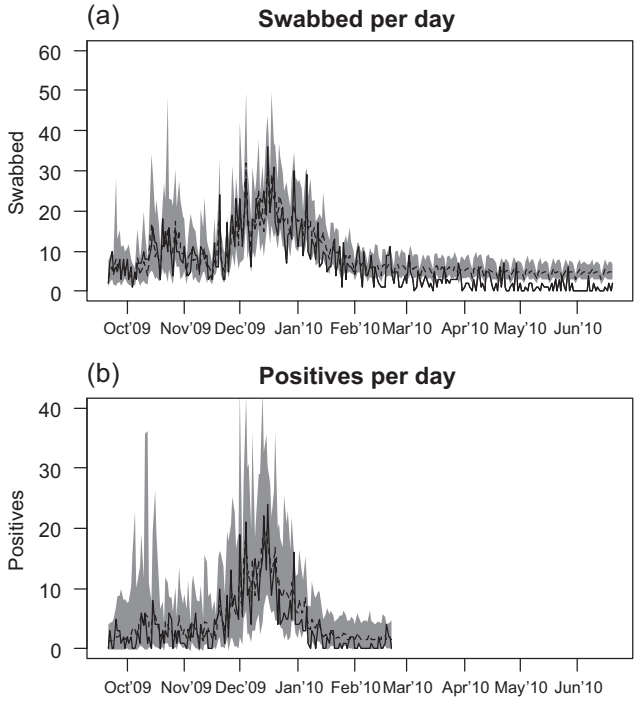

Fig. 5. Comparison of daily (swabbed and positive) data, solid line, with the results of model fit, dashed line (mean) and shaded area (95% high predictive density regions).

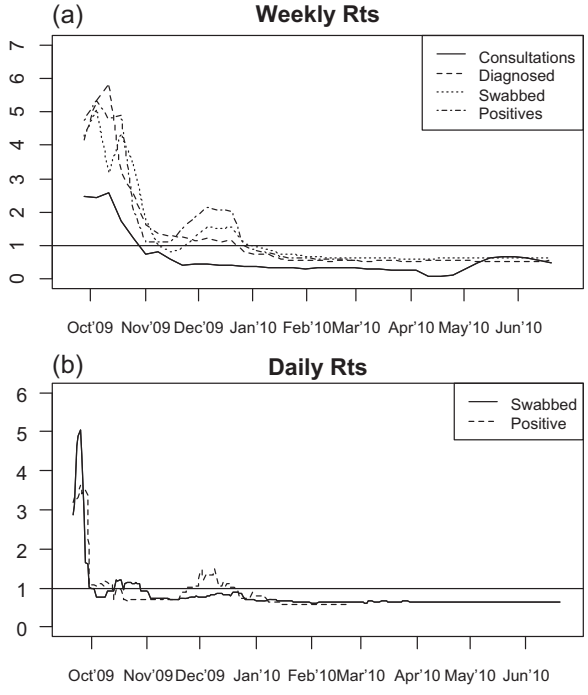

Fig. 6. Estimation of the effective reproduction ratio at any given point of the epidemic for different data sets, including weekly (consultations and diagnosed) and weekly-aggregated (swabbed and positives) data, (a), and daily (swabbed and positives) data, (b). Horizontal line corresponds to $R_t =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

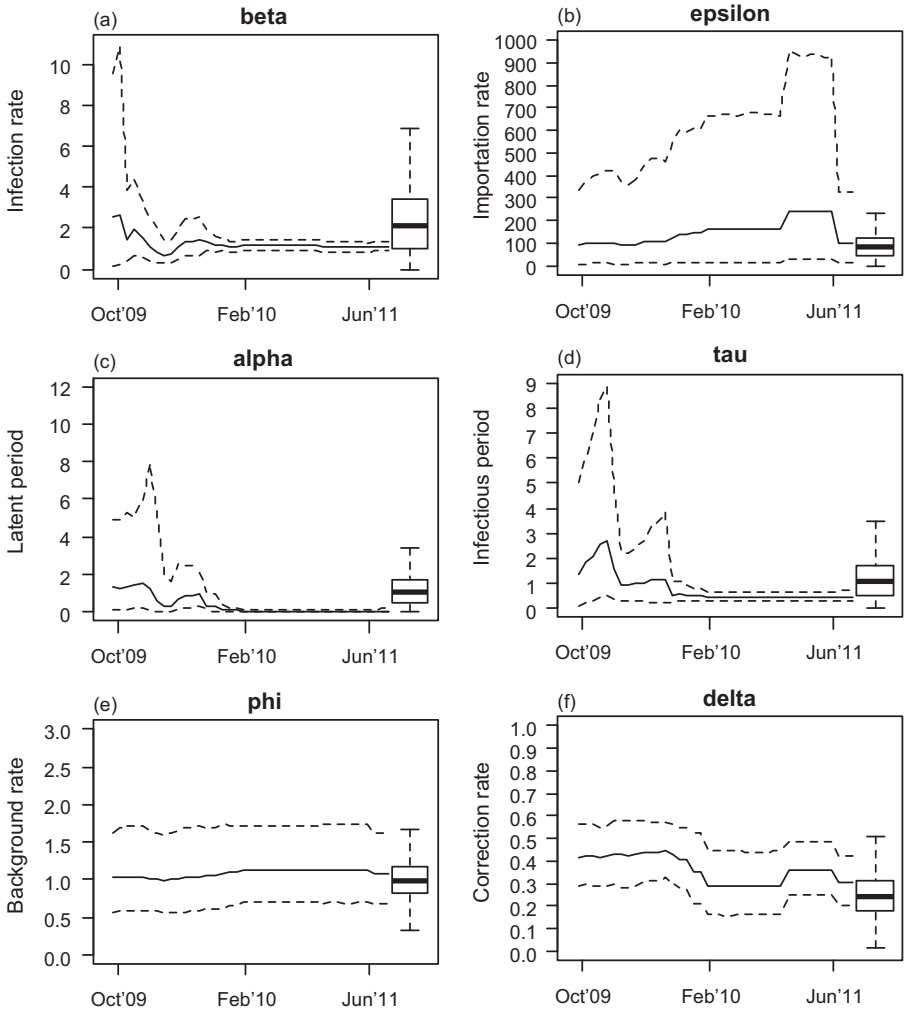

Fig. 7. Posterior and priors parameter distributions for the swabbed weekly data (for illustration). The box-plot represents on right represents the prior distribution, whereas the graph shows the evolution of the posterior distribution over time (solid line represents the mean and the dotted lines show 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 data sets 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 data sets 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, Figs. 3 and 8.

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 Figs. 2 and 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

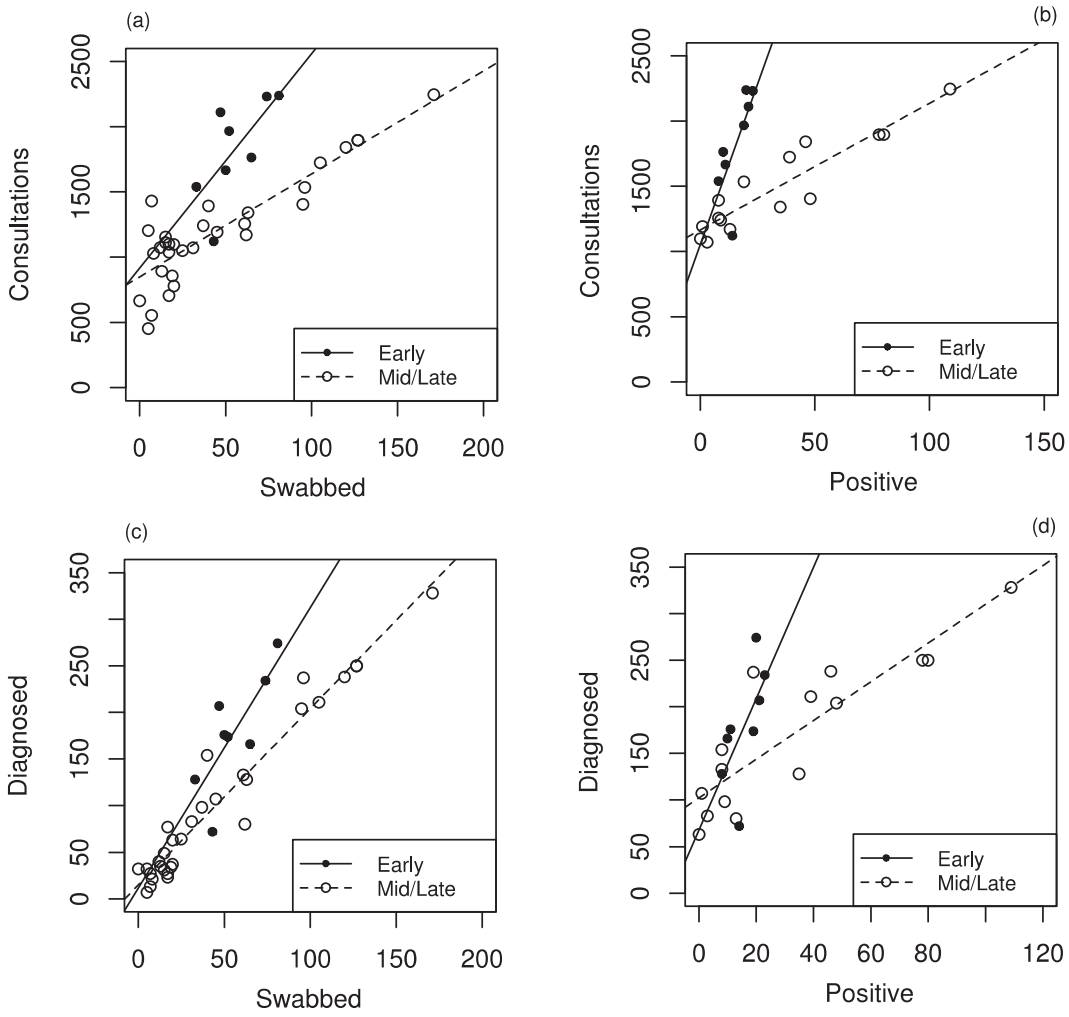

Fig. 8. Relationship between weekly and weekly-aggregated data for different periods in the epidemic timeline. Early period (weeks 39/2009 to 46/2009) is characterised by high overall levels and high variability of consultations and diagnosed cases as compared to swabbed and positive.

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.

Acknowledgments

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

References

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

Boëlle, P.Y., Bernillon, P., Desenclos, J.C., 2009. A preliminary estimation of the reproduction ratio for new influenza A(H1N1) from the outbreak in Mexico, March–April 2009. Euro Surveill. 14 (19), pii=19205, Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19205

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

Chang, C.Y., Cao, C.X., Wang, Q., Chen, Y., Cao, Z., Zhang, H., Dong, L., Zhao, J., Xu, M., Gao, M., 2010. The novel H1N1 influenza A global airline transmission and early warning without travel containments. Chin. Sci. Bull. 2010 (55), 3030–3036.

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

Chowell, G., Echevarria-Zuna, S., Vibound, C., Simonsen, L., Tamerius, J., Miller, M.A., Borja-Aburto, V.H., 2011a. Characterizing the epidemiology

of the 2009 influenza A/H1N1 pandemic in Mexico. PLoS Med. 8 (5), http://dx.doi.org/10.1371/journal.pmed.1000436, e1000436.
Chowell, G., Viboud, C., Munayco, C.V., Gomez, J., Simonsen, L., Miller, M.A., Tamerius, J., Fiestas, V., Halsey, E.S., Laguna-Torres, C.A., 2011b. Spatial and temporal characteristics of the 2009 A/H1N1 influenza pandemic in Peru. PLoS One 6 (6), e21287.
Clancy, D., O'Neill, P.D., 2008. Bayesian estimation of the basic reproduction number in stochastic epidemic models. Int. Soc. Bayesian Anal. 3, 737–758.
Correia, A.M., Queiros, L., Dias, J., 2010. Pandemic influenza A (H1N1) in the North of Portugal: how did the autumn–winter wave behave? Rev. Port Pneumol. 16 (6), 880–886.
Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. 10, 197–208.
Doucet, A., De Freitas, N., Gordon, N., 2001. Sequential Monte Carlo methods in practice, first ed. Springer Verlag, New York, NY.
Fierro, A., 2011. A simple stochastic lattice gas model for H1N1 pandemic. Application to the Italian epidemiological data. Eur. Phys. J. E: Soft Matter 34, http://dx.doi.org/10.1140/epje/i2011-11011-2.
Flahault, A., Vergu, E., Boëlle, P.Y., 2009. Potential for a global dynamic of influenza A(H1N1). BMC Infect. Dis. 9, 129.
Flasche, S., Hens, N., Boelle, P.Y., Mossong, J., Ballegooijen, W.M.V., Nunes, B., Rizzo, C., Popovici, F., Santa-Olalla, P., Hruba, F., Parmakova, K., Baguelin, M., Hoek, A.J.V., Desenclos, J.C., Bernillon, P., Camara, A.L., Wallinga, J., Asikainen, T., White, P.J., Edmunds, W.J., 2011. Different transmission patterns in the early stages of the influenza A(H1N1)v pandemic: a comparative analysis of 12 European countries. Epidemics 3, 125–133.
Fraser, C., Donnelly, C.A., Cauchemez, S., Hanage, W.P., Van Kerkhove, M.D., Hollingsworth, T.D., Griffin, J., Baggaley, R.F., Jenkins, H.E., Lyons, E.J., Jombart, T., Hinsley, W.R., Grassly, N.C., Balloux, F., Ghani, A.C., Ferguson, N.M., Rambaut, A., Pybus, O.G., Lopez-Gatell, H., Alpuche-Aranda, C.M., Chapela, I.B., Zavala, E.P., Guevara, D.M.E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324, 1557–1561.
Griffin, J.T., Garske, T., Ghani, A.C., 2011. Joint estimation of the basic reproduction number and generation time parameters for infectious disease outbreaks. Biostatistics 12 (2), 303–312.
Hsieh, Y., Cheng, K., Wu, T., Liz, T., Cheng, C., Chen, J., Lin, M., 2011. Transmissibility and temporal changes of 2009 pH1N1 pandemic during summer and fall/winter waves. BMC Infect. Dis. 11, 332.
Ishak, A., Tee, D., Nawmar, I., Pang, L.K., Ruslan, N., Che Mansor, N., Gam, L., 2011. H1N1 influenza: a viral infection. Infect. Dis. 2 (12), http://dx.doi.org/10.9754/journal.wmc.2011.002736, WMC002736; WebmedCentral.
Katriel, G., Yaari, R., Huppert, A., Roll, U., Stone, L., 2011. 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, 856–867.
Kenah, E., Chao, D.L., Matrajt, L., Halloran, M.E., Longini Jr., I.M., 2011. The global transmission and control of influenza. PLoS One 6 (5), e19515.
Nishiura, H., 2011. Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009). Biomed. Eng. Online 10 (15), http://dx.doi.org/10.1186/1475-925X-10-15.
Nishiura, H., Klinkenberg, D., Roberts, M., Heesterbeek, J.A.P., 2009. Early epidemiological assessment of the virulence of emerging infectious diseases: a case study of an influenza pandemic. PLoS One 4 (8), e6852.
Omori, R., Nishiura, H., 2011. Theoretical basis to measure the impact of short-lasting control of an infectious disease on the epidemic peak. Theor. Biol. Med. Modell. 8 (2), http://dx.doi.org/10.1186/1742-4682-8-2.
Ong, J.B.S., Chen, M.I.C., Cook, A.R., Lee, H.C., Lee, V.J., Lin, R.T.P., Tambyah, P.A., Goh, L.G., 2010. Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. PLoS One 5 (4), e10036.
Opatowski, L., Fraser, C., Griffin, J., de Silva, E., Van Kerkhove, M.D., Lyons, E.J., Cauchemez, S., Ferguson, N.M., 2011. Transmission characteristics of the 2009 H1N1 influenza pandemic: comparison of 8 southern hemisphere countries. PLoS Pathog. 7 (9), e1002225.
Poletti, P., Ajelli, M., Merler, S., 2011. The effect of risk perception on the 2009 H1N1 pandemic influenza dynamics. PLoS One 6 (2), e16460.
R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, ⟨http://www.R-project.org⟩.
Reed, C., Angulo, F.J., Swerdlow, D.L., Lipsitch, M., Meltzer, M.I., Jarnigan, D., Finelli, L., 2009. Estimates of the prevalence of pandemic (H1N1) 2009, United States, April–July 2009. Emerg. Infect. Dis. 15 (12), 2004–2007.
Rizzo, C., Rota, M.C., Bella, A., Giannitelli, S., De Santis, S., Nacca, G., Pompa, M.G., Vellucci, L., Salmaso, S., Declich, S., 2010. Response to the 2009 influenza A(H1N1) pandemic in Italy. Euro Surveill. 15 (49), pii=19744, Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19744
Trenkel, V.M., Elston, D.A., Buckland, S.T., 2000. Fitting population dynamics models to count and cull data using sequential importance sampling. J. Am. Stat. Assoc. 95, 363–374.
White, L.F., Wallinga, J., Finelli, L., Reed, C., Riley, S., Lipsitch, M., Pagano, M., 2009. 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 Respir. Viruses 3 (6), 267–276.
WHO, 2010. Pandemic (H1N1) 2009—Update 100. World Wide Web Electronic Publication, ⟨http://www.who.int/csr/don/2010_05_14/en/index.html⟩.
Yu, H., Cauchemez, S., Donnelly, C.A., Zhou, L., Feng, L., Xiang, N., Zheng, J., Ye, M., Huai, Y., Liao, Q., Peng, Z., Feng, Y., Jiang, H., Yang, W., Wang, Y., Ferguson, N.M., Feng, Z., 2012. Transmission dynamics, border entry screening, and school holidays during the 2009 influenza A (H1N1) pandemic, China. Emerg. Infect. Dis. 18 (5), 758–766.