Real-time prediction of the end of an epidemic wave: COVID-19 in China as a case-study

Quentin Griette[(a)], Zhihua Liu[(b)], Pierre Magal[(a)]*and Robin N. Thompson[(c)]

[(a)] *Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France.*
CNRS, IMB, UMR 5251, F-33400 Talence, France.

[(b)] *School of Mathematical Sciences, Beijing Normal University, Beijing 100875, People's Republic of China*

[(c)] *Christ Church, University of Oxford, St Aldate's, Oxford OX1 1DP, United Kingdom.*
Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, United Kingdom.

February 1, 2021

Abstract

Forecasting when an epidemic wave is likely to end is an important component of disease management, allowing deployment of limited control resources to be planned efficiently. Here, we report an analysis that we conducted in real-time during the first COVID-19 epidemic wave in mainland China. We developed a mathematical model to construct bounds on the end date of the first epidemic wave there, assuming that strong quarantine and testing measures remained in place until the epidemic wave was confirmed over. We used reported data on case numbers in China from January 20 to April 9, 2020. We first developed an analytic approach, obtaining a formula describing the probability distribution of the epidemic wave end date using a combination of deterministic modelling and the theory of continuous-time Markov processes. Then, we ran simulations of an individual-based model to demonstrate that our analytic predictions were accurate. We found that the predicted end date of the first epidemic wave in China depended on the proportion of infected individuals that are symptomatic and appear in case notification data, as opposed to remaining asymptomatic throughout their courses of infection. We therefore provide an easy-to-use approach for predicting the ends of epidemic waves, as well as a clear demonstration that predicted end-of-epidemic times depend on the extent of asymptomatic infection. Our framework can be applied to predict the ends of epidemic waves during future outbreaks of a wide range of pathogens.

Keywords: Mathematical modelling; COVID-19; end of epidemic; reported and unreported cases; control measures.

1 Introduction

The COVID-19 pandemic has now spread worldwide, causing over one million deaths and 40 million reported cases so far (as of 25 October, 2020 [37]). SARS-CoV-2, the virus that causes COVID-19, emerged in China at the end of 2019. In early 2020, the Chinese government imposed strong public health measures, including enhanced epidemiological surveys and surveillance, travel restrictions, quarantine, contact tracing and isolation [27]. These intense interventions were sufficient to bring the epidemic wave under control, and since mid-March case numbers have remained low.

A key challenge in infectious disease epidemiology is forecasting the progression of an epidemic. Significant attention has been directed towards developing methods for estimating future numbers of cases and deaths, as well as forecasting the timing of the epidemic peak [2, 3, 4, 5, 16, 17, 18, 19, 20, 28, 32]. Predicting the ends of epidemic waves, on the other hand, has received considerably less attention [23], despite the fact that the end of an epidemic wave signals an opportunity to relax costly public health measures. Some previous studies have estimated the probability that an epidemic is over as a function of the time since the last observed case using renewal equation models [22, 14] or stochastic compartmental

*Corresponding author: pierre.magal@u-bordeaux.fr

models [31]. However, predicting the end of the first COVID-19 epidemic wave in China was particularly challenging for two key reasons. First, evidence emerged early in the COVID-19 pandemic that infected individuals could transmit the virus prior to displaying symptoms ("presymptomatic infection"). Second, some infected individuals never display symptoms or display only mild symptoms, and therefore do not report disease ("asymptomatic infection"). It is now widely accepted that these presymptomatic and asymptomatic hosts play a significant role in SARS-CoV-2 transmission [6, 29, 30, 12, 33].

Early evidence for asymptomatic transmission included a study by Nishiura et al. [24], which reported early in the pandemic that 13 evacuees on charter flights from Wuhan (China) were infected and four of these individuals never developed symptoms. Chowell et al. [21] estimated the proportion of asymptomatic infections to be 17.9%. Research by Li et al. [15] generated an estimate that 86% of all infections were undocumented (95% CI: [82%-90%]) prior to the introduction of travel restrictions in China on January 23, 2020, and a team in China [35] suggested that there were 37,400 cases in Wuhan that authorities were unaware of by February 18, 2020. More recently, Ferretti et al. [6] split the reproduction number into components corresponding to transmission from symptomatic, presymptomatic and asymptomatic infectious individuals, as well as environmental transmission. Unreported cases, largely due to presymptomatic and asymptomatic infections, were a key driver of the rapid geographic spread of SARS-CoV-2 and explain why early containment of the virus was impossible (compared to, e.g. SARS [7]). In [4], we consider the symptomatic reported and unreported patients and we prove that it is hopeless to estimate the fraction of reported (or unreported) patients by using SI models. In other words, several values of the fraction of reported symptomatic patients give the exact same fit to the data. Finally, a study based on several cohorts of patients was conducted in Oran et al. [26].

Here, we consider a compartmental model characterising SARS-CoV-2 transmission, and parameterise it using data from the first (yet unique) epidemic wave in China. Our model incorporates key features of this epidemic wave, including explicit inclusion of public health measures designed to mitigate the severity of the epidemic, as well as presymptomatic and asymptomatic infections. When we conducted our analysis in real-time, the proportion of infected individuals that were symptomatic and reported disease was unknown (and, in fact, the precise value remains uncertain even now), so we consider a range of values of that parameter (f). We derive an analytic expression for predicting when an epidemic wave is likely to end, under the assumption that public health measures that are in place remain fixed until the epidemic wave is over. We use this expression to show how the predicted end of epidemic wave date changed as the epidemic wave continued, and compare these results to equivalent results obtained using model simulations. Not only do we provide a framework for predicting the ends of epidemic waves, but we also show that the times at which epidemic waves end depend on the proportion of detected cases. This emphasises the importance of intense surveillance to find infectious cases, including those who do not display clear symptoms.

2 Methods

2.1 Data

We use cumulative data describing daily numbers of cases in mainland China from January 20, 2020 to March 18, 2020, obtained from the National Health Commission of the People's Republic of China and Chinese Center for Disease Control and Prevention [38, 39]. Up until February 10 2020, cases in the dataset were only those that were confirmed by laboratory testing. From February 11 to February 15, data were available not only for cases confirmed by laboratory testing, but also for cases that were clinically diagnosed based on medical imaging. From February 16 onwards, these two data types were combined in the dataset, so that it was impossible to distinguish between laboratory confirmed and clinically diagnosed cases. Changing case definitions in response to changes in case numbers is necessary and commonplace [34], however such changes make inferring epidemic trends based on case numbers challenging. To account for this and remove the substantial jump in cases on February 16 due to changes in testing practices, we calculated the cumulative number of clinically diagnosed cases between February 11 and February 15, and subtracted this from the cumulative numbers of cases from February 16 onwards. We therefore obtained approximate numbers of confirmed cases throughout the period from January 20 to March 18, 2020. The dataset, accounting for this adjustment, is shown in the Supplementary Information (Table 4).

We note that, on January 23, mainland China began implementing lockdowns, beginning with a lockdown in the city of Wuhan.

2.2 Mathematical model

To characterise changes in observed case numbers from January 20 to March 18 in mainland China, we considered a compartmental model in which we track the number of individuals that are either susceptible to the virus ($S(t)$), in early infection and infectious ($I(t)$) and in later infection and reporting disease ($R(t)$) or in later infection and not reporting disease ($U(t)$) [11, 17]. Individuals that are in later infection and not reporting disease include those that are asymptomatic and those who develop only mild symptoms and so do not adhere to interventions targeting symptomatic individuals. The model is therefore given by:

$$\begin{cases} S'(t) = -\tau(t)S(t)[I(t) + U(t)], \\ I'(t) = \tau(t)S(t)[I(t) + U(t)] - \nu I(t), \\ R'(t) = \nu f I(t) - \eta R(t), \\ U'(t) = \nu(1-f)I(t) - \eta U(t), \end{cases} \tag{2.1}$$

with initial data

$$S(t_0) = S_0 > 0,\ I(t_0) = I_0 > 0,\ R(t_0) = R_0 \geq 0 \text{ and } U(t_0) = U_0 \geq 0. \tag{2.2}$$

In this model, $t \geq t_0$ is time in days and t_0 is the start date of the epidemic wave. A schematic illustrating the different model compartments is shown in Figure 1 and the model parameters - including whether the parameter values were assumed or obtained via model fitting - are listed in Table 1. It has previously been demonstrated that the latent period for COVID-19 is short [18], and COVID-19 patients have been found to have high viral loads early in infection [36, 13], so we do not include individuals who are presymptomatic and not yet infectious in the model. However, explicit inclusion of individuals who are infected but not yet infectious would be a straightforward extension of our model [19].

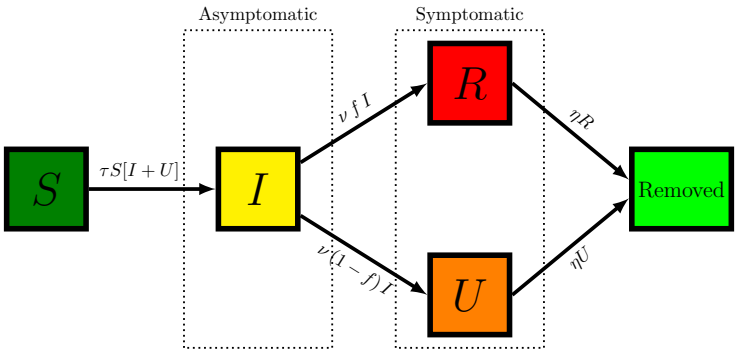

Figure 1: *Schematic showing the different compartments and transition rates in the model given by system of equations (2.1).*

Symbol	Interpretation	Method
t_0	Epidemic start time	fitted
S_0	Number susceptible at time t_0	fixed
I_0	Number in early infection and infectious at time t_0	fitted
U_0	Number in later infection and not reporting disease at time t_0	fitted
$\tau(t)$	Transmission rate at time t, accounting for public health measures	fitted
$1/\nu$	Average duration of early infection	fixed
f	Fraction of infected individuals that go on to report disease	fixed
$1/\eta$	Average duration of later infection	fixed

Table 1: *Parameters and initial conditions of the model.*

Early infection (which corresponds to the incubation period, for individuals who develop clear symptoms) is assumed to last for an average period of $1/\nu$ days. The infectious period is assumed to be $1/\nu + 1/\eta$ days, although we assume that individuals that report disease do not transmit the virus during their symptomatic infectious period (i.e. they adhere to public health measures that are effective at reducing transmission). A fraction, f, of infected hosts report disease, whereas a fraction $1 - f$ do not report disease at any stage of their infection.

In the model, the transmission rate at time t, accounting for public health measures in place at that time, is denoted by $\tau(t)$. During the exponential growth phase, we assume that $\tau(t) \equiv \tau_0$ is constant. We then use a time-dependent decreasing transmission rate $\tau(t)$ to incorporate the effects of the strong measures taken by Chinese authorities to control the epidemic wave (see Introduction for a description of the different measures that were introduced):

$$\begin{cases} \tau(t) = \tau_0, \, 0 \leq t \leq N, \\ \tau(t) = \tau_0 \, \exp\left(-\mu \left(t - N\right)\right), \, t > N. \end{cases} \tag{2.3}$$

The date N and the value of μ are chosen so that daily numbers of cumulative reported cases in the numerical simulation of the epidemic align with the analogous values in the dataset.

The cumulative number of reported cases at time t is given by

$$CR(t) = \nu f \int_{t_0}^{t} I(\sigma) d\sigma, \text{ for } t \geq t_0, \tag{2.4}$$

and the cumulative number of unreported cases at time t is given by

$$CU(t) = \nu(1-f) \int_{t_0}^{t} I(\sigma) d\sigma, \text{ for } t \geq t_0. \tag{2.5}$$

The daily number of reported cases can be obtained by computing the solution of the following equation:

$$DR'(t) = \nu \, f \, I(t) - DR(t), \text{ for } t \geq t_0 \text{ and } DR(t_0) = 0. \tag{2.6}$$

2.3 Parameter values

Since there is substantial uncertainty surrounding the proportion of cases that are symptomatic and report disease for COVID-19, the value of f is unknown. Since intense interventions were introduced in China during the first epidemic wave, and the full extent of asymptomatic transmission was unknown, we assume in the baseline version of our analysis that $f = 0.8$. However, we checked the robustness of our results to this assumption by also considering different values ($f = 0.2, 0.4$ and 0.6).

We assume that the durations of early and late infection are $\nu = 1/7$ days and $\eta = 1/7$ days, respectively. By assuming that the mean duration of early infection (i.e. duration of infection prior to symptoms, for individuals that go on to develop symptoms) is 7 days, the expected generation time for individuals that develop symptoms might be expected to be around 3.5 days. This lies within the range of estimated generation times for COVID-19 (see e.g. [8]). COVID-19 patients have been found to shed virus up to around one week after hospitalisation, thereby motivating our assumed value of η [36].

To determine the initial conditions (equations (2.2)), we assumed that in the initial exponential growth phase of the epidemic wave (the earliest stages of the epidemic, which is assumed to be between January 19 and January 26, 2020), $CR(t)$ took the form:

$$CR(t) = \chi_1 \exp(\chi_2 \, t) - \chi_3, \; t \geq t_0. \tag{2.7}$$

Following [16], expressions for I_0, U_0, R_0 can be obtained:

$$I_0 = \frac{\chi_2}{f(\nu f + \nu_2)}, \;\; U_0 = \left(\frac{(1-f)(\nu f + \nu_2)}{\eta + \chi_2} \right) I_0, \;\; R_0 = 0. \tag{2.8}$$

Furthermore, the transmission rate during this exponential growth phase of the epidemic wave is given by the constant value

$$\tau(t) = \tau_0 = \left(\frac{\chi_2 \, + \nu f + \nu_2}{S_0} \right) \left(\frac{\eta \, + \, \chi_2}{\nu(1-f) \, + \, \eta + \, \chi_2} \right), \tag{2.9}$$

the epidemic start time is

$$t_0 \, = \, \frac{1}{\chi_2} \left(\log(\chi_3) \, - \, \log(\chi_1) \right), \tag{2.10}$$

and the value of the basic reproductive number is

$$\mathcal{R}_0 \, = \, \left(\frac{\tau_0 S_0}{\nu f + \nu_2} \right) \left(1 \, + \, \frac{\nu_2}{\eta} \right). \tag{2.11}$$

In the above, the value of $\chi_3 = 30$ is assumed and the values of χ_1 and χ_2 are obtained by fitting equation (2.7) to data on the cumulative numbers of cases per day using least squares estimation. Specifically, we use the "polyfit" Matlab function to estimate χ_1 and χ_2. The population size is assumed to be large, so that the initial number of susceptible individuals, S_0, corresponds to the total population size.

3 Results

3.1 Fitting the model to data

We first estimated the values of χ_1 and χ_2 using data on the cumulative number of confirmed cases in the earliest stages of the epidemic wave (January 19 to January 26, 2020). The values of τ_0 and the initial conditions (I_0, U_0 and t_0) are then obtained using formulae (2.8)-(2.10). The fitted parameter values are shown in Table 2. Analogous results for different values of the reporting fraction, f, are also shown.

χ_1	χ_2	χ_3	t_0	f	μ	N	I_0	U_0	S_0	τ_0
0.2601	0.3553	30	13.3617	0.8	0.1480	Jan. 26	93.2785	5.3494	1.40005×10^9	3.3655×10^{-10}
0.2601	0.3553	30	13.3617	0.6	0.1531	Jan. 26	124.3550	14.2646	1.40005×10^9	3.1920×10^{-10}
0.2601	0.3553	30	13.3617	0.4	0.1574	Jan. 26	186.5325	32.0953	1.40005×10^9	3.0358×10^{-10}
0.2601	0.3553	30	13.3617	0.2	0.1612	Jan. 26	373.0650	85.5875	1.40005×10^9	2.8942×10^{-10}

Table 2: *Values of parameters obtained by fitting to cumulative data from the initial exponential phase of the mainland China epidemic wave. The values of I_0 U_0, τ_0, and t_0 are obtained using formulae (2.8)-(2.10). Here we take $\chi_3 = 30$ in order to obtain non-zero integer values of I_0 and U_0.*

We then used the mathematical model (2.1) with these parameter values and initial conditions to project the cumulative number of reported cases forwards (black line in Figure 2 (left)), choosing μ so that $CR(t)$ matched the observed data (red dots in Figure 2 (left)). The inferred cumulative numbers of unreported cases are also shown in Figure 2 (left), assuming that $f = 0.8$. Daily numbers of reported cases corresponding to this forward projection are shown in Figure 2 (right).

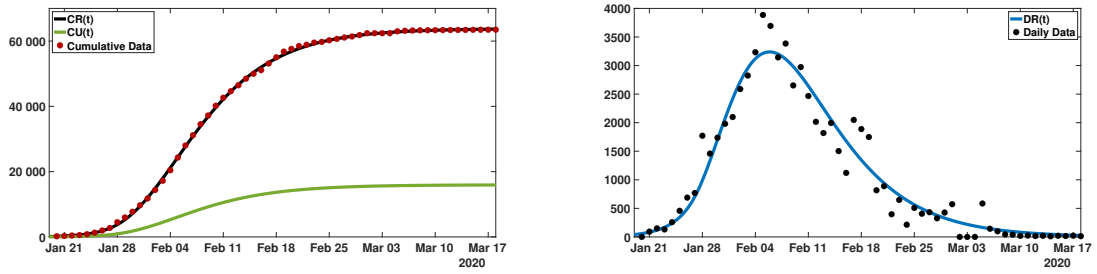

Figure 2: *Comparison of the model output with the data for mainland China. The parameter values and initial conditions are listed in Table 2, and $f = 0.8$. On the left hand side we plot the cumulative data (red dots), the simulated cumulative reported cases $CR(t)$ (black line) and unreported cases $CU(t)$ (green line). On the right hand side, we plot data on the daily numbers of cases (black dots) and the inferred daily number of cases using the model, $DR(t)$ (blue line).*

3.2 Predicting the end of the epidemic wave

To predict the end of the epidemic wave, we are particularly interested in the time period in which cases are fading out and very few new infections are occurring. We consider a scenario in which the current time is day t_1, and we are attempting to predict when the epidemic will end. As long as t_1 is sufficiently long after the peak of the epidemic wave that the quantity $\tau(t)S(t) \leq \tau(t)S_0$ is small, the approximation

$$I'(t) \simeq -\nu I(t),$$

can be used instead of the second equation in system (2.1) when $t > t_1$. For the parameter values used in our model, temporal changes in $S_0\tau(t)$ are shown in the Supplementary Information (Figure 6), highlighting that $\tau(t)S(t)$ is small from the second half of March, 2020, onwards).

Hence, to obtain an analytic expression describing the predicted end of the epidemic wave, we considered the following approximate system of equations whenever $t \geq t_1$:

$$\begin{cases} I'(t) = -\nu I(t), \\ R'(t) = \nu f\, I(t) - \eta R(t), \\ U'(t) = \nu(1-f)\, I(t) - \eta U(t). \end{cases} \tag{3.1}$$

This system is supplemented by the initial data

$$I(t_1) = I_1, U(t_1) = U_1 \text{ and } R(t_1) = R_1. \tag{3.2}$$

where I_1, U_1 and R_1 are the values of the solutions of the original system (2.1)-(2.2) on day t_1. A schematic for the approximate model (3.1) is shown in the Supplementary Information (Figure 7).

The error between the original model and the approximate model is shown in the Supplementary Information (Figure 8), where the error is given by

$$\text{err}(t_1) = \sup_{t \geq t_1} \max\left(|I(t) - I_1(t)|, |U(t) - U_1(t)|\right). \tag{3.3}$$

In this expression, $I(t)$ and $U(t)$ are the solutions of the original system (2.1), and $I_1(t)$ and $U_1(t)$ are solutions of the approximate model. In both cases, the models are fitted to observed data on cumulative numbers of reported cases (hence, this error formula does not involve $R(t)$ which is very similar for the two models). When applied in the later stages of the epidemic wave, the approximate model is more accurate than earlier in the epidemic wave.

By considering the analogous continuous-time Markov chain to the approximate model (3.1), the probability that the epidemic is over on different future dates can be estimated analytically (see Supplementary Information section 5 for additional details). Specifically, the probability that no individuals remain in the I or U compartments can be calculated at different times in future:

$$\begin{aligned} &\mathbb{P}(I(s) + U(s) = 0 \text{ for } s \geq t \,|\, I(t_1) = I_1, U(t_1) = U_1) \\ &\qquad = \left(1 - e^{-\eta(t-t_1)}\right)^{U_1} \times \left(1 - e^{-\nu(t-t_1)} - (1-f)\nu(t-t_1)e^{-\eta(t-t_1)}\right)^{I_1}. \end{aligned} \tag{3.4}$$

The predictions generated by equation (3.4) for different values of t_1 are shown in Figure 3. We note that, as t_1 increases, the probability distribution of the date of extinction converges to a limit profile.

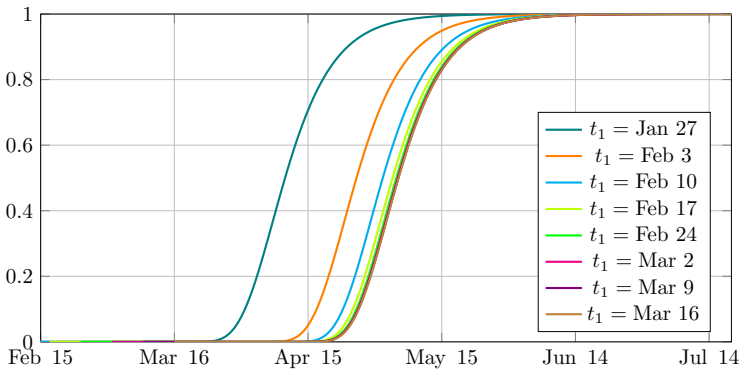

Figure 3: *Estimated extinction probabilities (using equation (3.4)). The numerical values for I_1 and U_1 were computed from the ODE model, considering t_1 values at 7 day intervals. In this figure, we assume that $f = 0.8$ (other parameter values are listed in Table 2).*

Furthermore, we also computed the earliest dates that corresponded to at least 90%, 95% and 99% probabilities that the epidemic was over, for different values of t_1, using equation (3.4). The results of this analysis are shown in Figure 4.

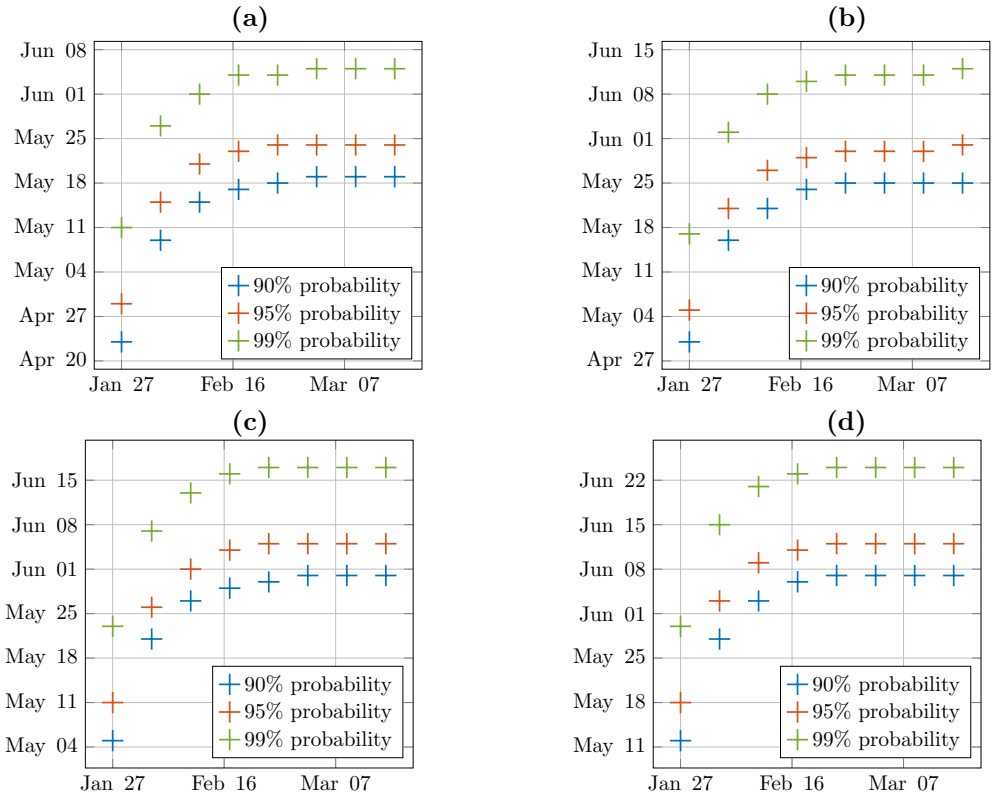

Figure 4: *For each panel, the x-axis corresponds to the day t_1 and the y-axis corresponds to the dates of epidemic wave extinction at different probability levels (*90%, 95% *and* 99%*) computed by using (3.4). Different panels correspond to different values of the parameter f ((a) $f = 0.8$; (a) $f = 0.6$; (a) $f = 0.4$; (a) $f = 0.2$). The values of I_1 and U_1 are computed by solving system of equations (2.1) numerically up to the time $t = t_1$. Parameter values are listed in Table 2.*

3.3 Comparing the analytic predictions with stochastic simulations

To investigate the accuracy of our analytic predictions, we also estimated the end of epidemic time using simulations of the analogous stochastic model to the system of equations (2.1). Specifically, as before, the deterministic model was fitted to the data on cumulative numbers of confirmed cases and used up until time t_1. Then from time t_1 onwards, stochastic simulations were run using the direct method version of the Gillespie stochastic simulation algorithm [9].

In Figure 5, we plot the cumulative distribution for the epidemic wave extinction probability obtained using the stochastic simulations. As can be seen in that figure, since the stochastic simulations involve using the exact model (equations (2.1)) rather than the approximate model, the predicted end dates of the epidemic wave are independent of t_1. The graph in Figure 5 corresponds to the limit profile discussed at the end of the previous section (i.e. the analytic prediction when t_1 is sufficiently late in the epidemic that the analytic prediction is accurate). From Figure 3, it can be seen that that this approximation is accurate when t_1 is February 17, 2020, or later.

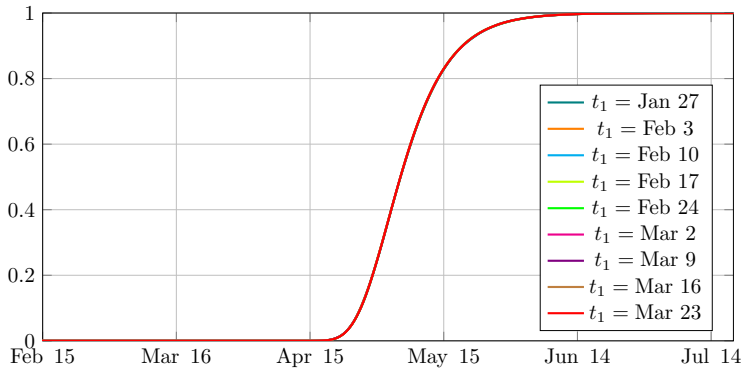

Figure 5: *Estimated cumulative probability distribution for the end of epidemic wave date obtained using stochastic simulations. Results are shown for different values of* t_1*, although as expected the different lines in this graph lie on top of each other. Initial conditions for the stochastic simulations were computed by rounding the solutions of equations (2.1) at* $t = t_1$ *to the nearest integers. 150,000 simulations were run for each value of* t_1*. In this figure,* $f = 0.8$*. Other parameter values are shown in Table 2.*

We also computed the error between the analytic end of epidemic time prediction and the analogous quantity using the stochastic simulations. More precisely, we computed the quantity

$$\text{diff}(t_1) = \sup_{t \geq t_1} |f_{IBM}(t) - f_{\text{analytic}}(t)| \tag{3.5}$$

for each value of t_1 presented in Figures 3 and 5, where f_{IBM} is the cumulative distribution computed by stochastic simulations (Figure 5) and f_{analytic} is the cumulative distribution given by equation (3.4) (Figure 3). The results are shown in the Supplementary Information (Table 6).

Finally, we compared the mean outputs from the stochastic simulations to the numerical solutions of the original model (system of equations (2.1)). Unsurprisingly, these quantities match closely (Figure 9). In Figure 10, we show the variability between different stochastic simulations obtained when the stochastic simulations are run throughout the epidemic (i.e. starting on day t_0). This high variability observed between different simulations is largely due to the small number of individuals infected initially; when instead stochastic simulations were run from day t_1 onwards, the variability between different stochastic simulations reduced (see Supplementary Information, Table 7).

4 Discussion

Despite receiving surprisingly little attention from epidemiological modellers, predicting the ends of epidemic waves is important for estimating how long intense interventions are likely to be required [22, 14, 31, 23]. In this study, we developed a framework for predicting the ends of epidemic waves using compartmental epidemiological models. This involving fitting a compartmental model to case notification data and using an analytic expression to estimate when the epidemic wave is likely to end. We also compared our analytic prediction to analogous results obtained via model simulations, thereby demonstrating that our results are accurate whenever the underlying epidemiological model provides a realistic reflection of pathogen transmission.

In Table 3, we show the results that we obtained using this framework in real-time to predict the end of the first COVID-19 epidemic wave in China. Specifically, the results in this table correspond to those shown in Figure 3, after the end of epidemic wave probability converged to the limit profile (i.e. using values of t_1 from approximately mid-February onwards). Importantly, the predicted epidemic wave end date depended on the assumed proportion of infectious cases that report disease (f). Since this quantity was unknown, and remains uncertain even now, we conclude that accurate estimation of the reporting fraction is essential to forecast the ends of epidemic waves accurately.

Level of risk	10%	5%	1%
Extinction date ($f = 0.8$)	May 19	May 24	June 5
Extinction date ($f = 0.6$)	May 25	May 31	June 12
Extinction date ($f = 0.4$)	May 31	June 5	June 17
Extinction date ($f = 0.2$)	June 7	June 12	June 24

Table 3: *The predicted end of epidemic wave date inferred when t_1 was March 16, 2020, for different levels of risk aversion. For example, assuming $f = 0.8$, our model predicted a 10% chance that the epidemic wave would persist beyond May 19, 2020.*

Our intention here was to develop a basic modelling approach for predicting when an epidemic wave is likely to end. To improve the accuracy of predictions, this approach would require adjustments to account for important features of real-world epidemic waves. As well as uncertainty in the reporting fraction, another key assumption was that public health measures remained in place until the end of the epidemic wave. Of course, if measures such as isolation of infectious cases are relaxed before an epidemic wave has ended, then the epidemic end date is likely to be different to the one predicted using our modelling framework. In that scenario, relaxation of interventions could in theory be integrated explicitly into the underlying model, and model simulations used to predict the end of epidemic waves. Since interventions are often included in compartmental models [3, 5, 32, 28], this is a straightforward extension of the research presented here. We also note that, if interventions are relaxed following the end of an epidemic wave, then additional cases could begin a second wave - a phenomenon that is now arguably being observed in a range of countries worldwide for COVID-19.

We note that there were very few cases in mainland China after mid-March, 2020. As a result, our modelling framework tended to estimate later end of epidemic wave dates than turned out to be the case. The most likely explanation for this is that, by characterising the impacts of control interventions using equation (2.3), public health measures did not have a sufficiently strong effect in the model. Testing the effects of different possible characterisations of the effects of public health measures is left as future work.

Since the precise method of parameter inference was not central to our framework, we used a basic approach to estimate the values of pathogen transmission parameters here, namely least squares estimation. Many different methods are used to estimate transmission parameters in real-time during epidemics [25, 1], and our modelling framework could be extended to use these more sophisticated methods.

Despite the many simplifications in our modelling approach as presented here, we have provided an initial framework for predicting the ends of epidemic waves, and demonstrated the key principle that the end date of an epidemic wave depends sensitively on the proportion of infectious cases that report disease. Extending this framework to include additional epidemiological realism, so that ends of epidemic waves can be forecasted as accurately as possible, is an important target for future research. This will allow public health decision makers to plan control interventions effectively during infectious disease epidemics.

Author contributions

All authors conceived and designed the study. Q.G. and P.M. analysed the data, carried out the analysis and performed the numerical simulations, Z.L. and P.M. conducted the literature review. All authors participated in writing and editing the manuscript.

Acknowledgements

The numerical simulations presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr/).

Funding: This research was funded by the National Natural Science Foundation of China (grant number: 11871007 (ZL)), NSFC and CNRS (Grant number: 11811530272 (ZL, PM)) and the Fundamental Research Funds for the Central Universities (ZL). This research was also funded by the Agence Nationale de la Recherche in France (Project name : MPCUII (QG, PM)), and Christ Church, Oxford, via a Junior Research Fellowship (RNT).

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] T. Britton and E. Pardoux, *Stochastic Epidemic Models with Inference*, Springer (2019).

[2] R.M. Cotta, C.P. Naveira-Cotta and P. Magal (2020), Modelling the COVID-19 epidemics in Brasil: Parametric identification and public health measures influence, Biology 2020, 9(8), 220.

[3] N.G. Davies et al., Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study, *Lancet Public Health*, **5(7)** (2020), e375-e385.

[4] J. Demongeot, Q. Griette and P. Magal (2020) SI epidemic model applied to COVID-19 data in mainland China Royal Society Open Science (2020), 7:201878.

[5] N.M. Ferguson et al., Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand, (2020), see: See www. imperial.ac.uk/mrc-global-infectious-disease-analysis/ covid-19/report-9-impact-of-npis-on-covid-19/.

[6] L. Ferretti et al., Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing, *Science*, **368(6491)** (2020), eabb6936.

[7] C. Fraser et al., Factors that make an infectious disease outbreak controllable, *PNAS* **101(16)** (2004), 6146-6151.

[8] T. Ganyani et al., Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020, *Eurosurveillance* **25** (2020), 2000257.

[9] D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.* **8** (1977), 2340-2361.

[10] Q. Griette and P. Magal, Clarifying predictions for COVID-19 from testing data: the example of New-York State, (2021) Volume 6 (2021), 273-283.

[11] Q. Griette, P. Magal and O. Seydi, Unreported cases for Age Dependent COVID-19 Outbreak in Japan, *Biology* **9** (2020), 132.

[12] W. Guan et al., Clinical Characteristics of Coronavirus Disease 2019 in China, *New England Journal of Medicine*, (2020). Published on February 28, 2020, PMID: 32109013. https://doi.org/10.1056/NEJMoa2002032.

[13] X. He et al., Temporal dynamics in viral shedding and transmissibility of COVID-19, *Nature Medicine*, **26** (2020), 672-675.

[14] H. Lee and H. Nishiura, Sexual transmission and the probability of an end of the Ebola virus disease epidemic. *Journal of theoretical biology*, **471** (2019), 1-12.

[15] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang and J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* (2020). https://doi.org/10.1126/science.abb3221

[16] Z. Liu, P. Magal, O. Seydi and G. Webb, Understanding unreported cases in the 2019-nCov epidemic outbreak in Wuhan, China, and the importance of major public health interventions, *Biology*, **9(3), 50** (2020). https://doi.org/10.3390/biology9030050

[17] Z. Liu, P. Magal, O. Seydi and G. Webb, Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data, *Mathematical Biosciences and Engineering* **17(4)** (2020), 3040-3051. https://doi.org/10.3934/mbe.2020172

[18] Z. Liu, P. Magal, O. Seydi and G. Webb, A COVID-19 epidemic model with latency period, *Infectious Disease Modelling* **5** (2020), 323-337.

[19] Z. Liu, P. Magal, O. Seydi and G. Webb, A model to predict COVID-19 epidemics with applications to South Korea, Italy, and Spain, *SIAM News* May 01 2020.

[20] Z. Liu, P. Magal, G. Webb, Predicting the number of reported and unreported cases for the COVID-19 epidemics in China, South Korea, Italy, France, Germany and United Kingdom, *Journal of Theoretical Biology*, Volume 509, 21 (2021).

[21] K. Mizumoto, K. Kagaya, A. Zarebski and G. Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveill.* **25(10)** (2020). https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180

[22] H. Nishiura, Y. Miyamatsu and K. Mizumoto, Objective determination of end of MERS outbreak, South Korea, 2015, *Emerging infectious diseases*, **22(1)** (2016), 146.

[23] H. Nishiura, Methods to determine the end of an infectious disease epidemic: A short review, In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases* (eds. G. Chowell, J.M. Hyman), (2016), 291-301.

[24] H. Nishiura et al., Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19), *International Journal of Infectious Diseases*, (2020). Published:March 13, https://doi.org/10.1016/j.ijid.2020.03.020.

[25] P.D. O'Neill, Introduction and snapshot review: Relating infectious disease transmission models to data, *Statistics in Medicine*, **29** (2010), 2069-2077.

[26] D. P. Oran, & E. J. Topol, (2020) Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Annals of internal medicine*, **173(5)**, 362-367.

[27] A. Pan et al., Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China, *JAMA*, **323(19)** (2020), 1915-1923.

[28] K. Prem et al., The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study, *Lancet Public Health*, **5** (2020), e261-e270.

[29] J. Qiu, Covert coronavirus infections could be seeding new outbreaks, *Nature*, (2020). https://www.nature.com/articles/d41586-020-00822-x

[30] C. Rothe et al., Transmission of 2019-nCoV infection from an asymptomatic contact in Germany, *New England Journal of Medicine*, (2020). https://doi.org/10.1056/NEJMc2001468

[31] R. N. Thompson, O. W. Morgan and K. Jalava, Rigorous surveillance is necessary for high confidence in end-of-outbreak declarations for Ebola and other infectious diseases, *Philosophical Transactions of the Royal Society B*, **374(1776)** (2019), 20180431. https://doi.org/10.1098/rstb.2018.0431

[32] R. N. Thompson, Epidemiological models are important tools for guiding COVID-19 interventions, *BMC Medicine*, **18** (2020), 152.

[33] R. N. Thompson, F. A. Lovell-Read and U. Obolski, Time from Symptom Onset to Hospitalisation of Coronavirus Disease 2019 (COVID-19) Cases: Implications for the Proportion of Transmissions from Infectors with Few Symptoms. *Journal of Clinical Medicine*, **9(5)** (2020), 1297.

[34] T. K. Tsang et al., Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. *Lancet Public Health*, **5** (2020), e289-296.

[35] C. Wang et al., Evolving Epidemiology and Impact of Non-pharmaceutical Interventions on the Outbreak of Coronavirus Disease 2019 in Wuhan, China, *medRxiv*. https://doi.org/10.1101/2020.03.03.20030593

[36] R. Wölfel et al., Virological assessment of hospitalized patients with COVID-2019, *Nature*, (2020). https://doi.org/10.1038/s41586-020-2196-x

[37] Worldometer, Covid-19 Coronavirus Pandemic, (2020), see: See www.worldometers.info/coronavirus/.

[38] The National Health Commission of the People's Republic of China http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml(accessed on 10 April 2020)

[39] Chinese Center for Disease Control and Prevention. http://www.chinacdc.cn/jkzt/crb/zl/szkb_11803/jszl_11809/ (accessed on 10 April 2020)

5 Supplementary Information

5.1 Formula to compute the probability distribution of the extinction date

We use continuous-time Markov processes to compute the exact distribution of the date of end of the epidemic after the transmission rate is effectively taken as zero. We start on t_1 with initial values I_1, U_1, and R_1 for I-individuals, U-individuals and R-individuals, respectively. The evolution of each individual is guided by independent exponential processes, and we have the following:

(i) Each individual I will change state following an exponential clock of rate ν. When I changes its state, it will be transferred to the class of R-individuals with probability f and to the class of U-individuals with probability $(1-f)$;

(ii) Each individual in the state U will change state following an exponential clock with rate η and become removed individual;

(iii) Each individual in the state R will change state following an exponential clock with rate η and become removed individual

Since the class I has only outgoing fluxes, the law of extinction for the I-individuals is

$$\mathbb{P}(I(t)=0\,|\,I(t_1)=I_1)=\left(\int_{t_1}^{t}\nu e^{-\nu(s-t_1)}\mathrm{d}s\right)^{I_1}=\left(1-e^{-\nu(t-t_1)}\right)^{I_1},$$

and the probability to have some I-individual left at time t is

$$\mathbb{P}(I(t)=I\,|\,I(t_1)=I_1)=(1-e^{-\nu(t-t_1)})^{I_1-I}e^{-\nu I(t-t_1)}.$$

For the U-individuals and the R-individuals, the situation is more intricate. Indeed, the U-individuals and the R-individuals vanish at a constant rate η but new individuals appear from the I class at rate $(1-f)\nu$ and $f\nu$, respectively, depending on the remaining stock of I. Therefore the probability that U gets extinct before t also depends on the number of remaining I. It is actually easier to compute directly the extinction property for the sum $I+U$, which is our aim anyways.

When $\nu \neq \eta$, we obtain

$$\begin{aligned}
&\mathbb{P}(I(s)+U(s)=0\,\forall s\geq t\,|\,I(t_1)=I_1,U(t_1)=U_1)\\
&=\left(1-e^{-\eta(t-t_1)}\right)^{U_1}\times\left(\int_{t_1}^{t}\mathbb{P}(U\to RR\text{ before }t\,|\,I\to U\text{ at }s)\mathbb{P}(I\to U\text{ at }s)+\mathbb{P}(I\to R\text{ at }s)\mathrm{d}s\right)^{I_1}\\
&=\left(1-e^{-\eta(t-t_1)}\right)^{U_1}\times\left(\int_{t_1}^{t}\left(1-e^{-\eta(t-s)}\right)\times(1-f)\nu e^{-\nu(s-t_1)}+f\nu e^{-\nu(s-t_1)}\mathrm{d}s\right)^{I_1}\\
&=\left(1-e^{-\eta(t-t_1)}\right)^{U_1}\times\left((1-f)\left(1-e^{-\nu(t-t_1)}-\nu\frac{e^{-\nu(t-t_1)}-e^{-\eta(t-t_1)}}{\eta-\nu}\right)+f(1-e^{-\nu(t-t_1)})\right)^{I_1}\\
&=\left(1-e^{-\eta(t-t_1)}\right)^{U_1}\times\left(1-e^{-\nu(t-t_1)}-(1-f)\nu\frac{e^{-\nu(t-t_1)}-e^{-\eta(t-t_1)}}{\eta-\nu}\right)^{I_1},
\end{aligned}$$

where the RR-individuals are the removed individuals.

Similarly when $\eta=\nu$, we obtain

$$\begin{aligned}
&\mathbb{P}(I(s)+U(s)=0\,\forall s\geq t\,|\,I(t_1)=I_1,U(t_1)=U_1)\\
&\qquad=\left(1-e^{-\eta(t-t_1)}\right)^{U_1}\times\left(1-e^{-\nu(t-t_1)}-(1-f)\nu(t-t_1)e^{-\eta(t-t_1)}\right)^{I_1}. \qquad (5.1)
\end{aligned}$$

5.2 Cumulative distribution of the date of end of the epidemic

The stochastic simulations introduced in section 3.3 can be used, in particular, to precisely estimate the cumulative probability distribution of the date of end of the epidemic, defined as the last time at which the quantity $I+U$ is positive.

In order to get a measure of the precision we remark that the values taken by the cumulative probability distribution $f(t)$ can be estimated by the average of independent measures of the random variable

$$X=\mathbb{1}_{t_{ext}\leq t},$$

which follows an Bernouilli distribution of parameter $f(t)$. Consecutive runs of the individual-based simulations yield independent observations X_n of this distribution. By Hoeffding's inequality we have for all $\varepsilon>0$ and $n\in\mathbb{N}$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_n-f(t)\right|\geq\varepsilon\right)\leq 2\exp\left(-2\varepsilon^2 n\right)=:\alpha,$$

and we achieved an error of at most $\varepsilon=10^{-3}$ at risk $\alpha\leq 10^{-3}$ by running $n=-\frac{2}{\varepsilon^2}\ln\left(\frac{\alpha}{2}\right)\approx 15201805$ independent individual-based simulations to estimate the probability distribution of the extinction time (Figure 5, $t_1=82$ *i.e.* March 23). Other curves are esimated on the basis of 152019 independent simulations, which amouts to an error of at most 10^{-2} at risk 10^{-3}.

Since the curves presented in Figure 3 are so similar that it is difficult to see any difference between them, we computed the absolute error between each curve and the "reference" of $t_1=82$. We present the numerical values in Table 5. Notice that the error is actually below the estimated precision of the approximation.

5.3 Supplementary figures

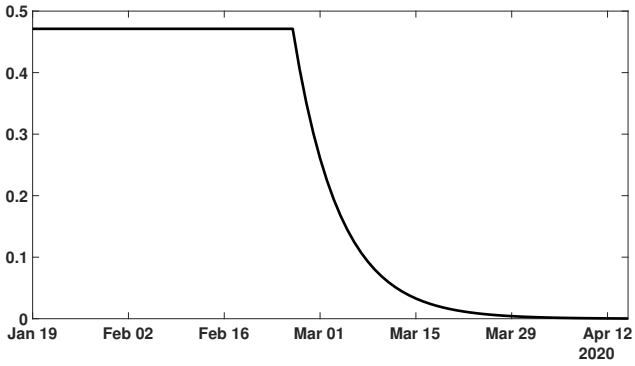

Figure 6: *Graph of* $\tau(t)S_0 = \tau_0 S_0 \exp(-\mu \max(t-N,0))$ *with* $S_0 = 1.40005 \times 10^9$, $\tau_0 = 3.3655 \times 10^{-10}$, $N =$ *Jan.* 26, *and* $\mu = 0.148$. *The transmission rate is very small in the second half of March onwards. The parameter values correspond to the baseline case that we considered* ($f = 0.8$) *see Table 2.*

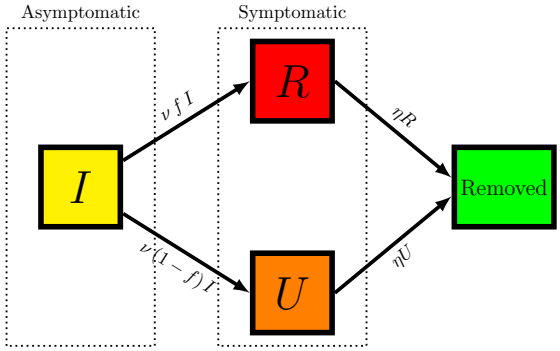

Figure 7: *Schematic of the model (3.1).*

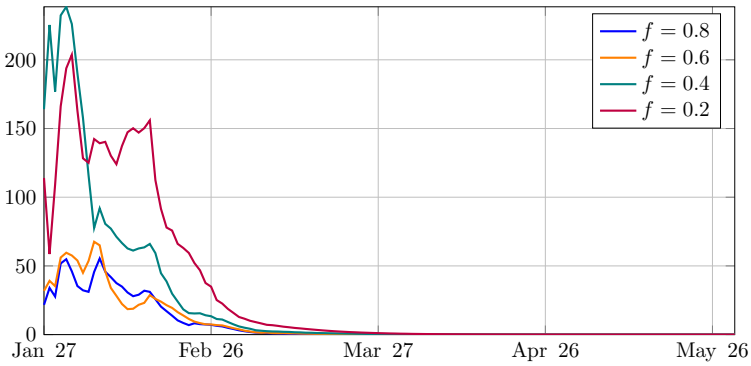

Figure 8: *In this figure the x-axis corresponds to* t_1 *and the y-axis correponds to the error* $err(t_1)$ *defined in (3.3). We observe that the smaller* f, *the larger the error. Parameter values are listed in Table 2.*

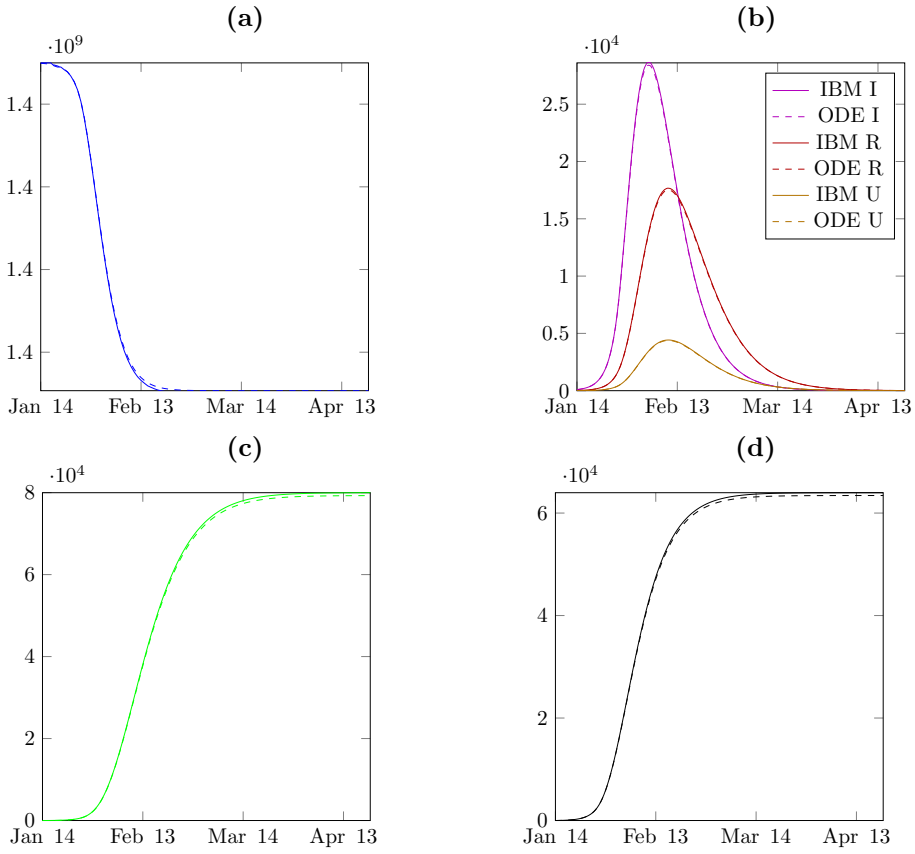

Figure 9: *In figure (a) we plot a comparison between the average S (susceptible) computed from the IBM and the S component of the solution of (2.1). In figure (b) we plot a comparison between the average I (asymptomatic), R (reported) and U (unreported) computed from the IBM and the components I, R and U of the solution of (2.1). In figure (c) we plot a comparison between the average RR (removed) computed from the IBM and the components RR of the solution of (2.1). In figure (d) we plot a comparison between the average CR (cumulative reported cases) computed from the IBM and the curve CR computed by (2.1)-(2.4). In this figure* 500 *independent runs of the IBM simulations are used and the corresponding components of the ODE model start from the same initial condition (at $t = t_0$). The parameters we used for both computations are the following: $I_0 = 93$, $U_0 = 5$, $S_0 = 1.40005 \times 10^9 - (I_0 + U_0)$, $R_0 = RR_0 = CR_0 = 0$ and $f = 0.8$, $\tau_0 = 3.3655 \times 10^{-10}$, $N = 26$, $\mu = 0.148$, $\nu = \frac{1}{7}$, $\eta = \frac{1}{7}$, $t_0 = 13.3617$.*

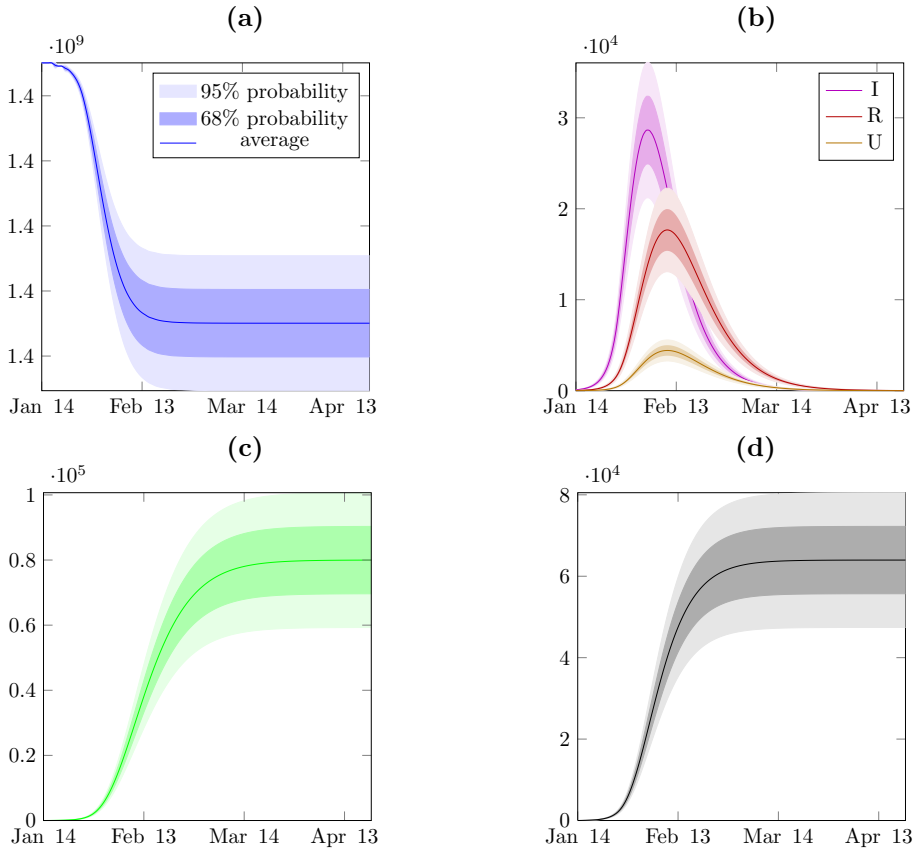

Figure 10: *In figure (a) we plot the mean value and variance of S (susceptible) computed from the IBM. The dark blue area contains 68% of the trajectories, and the light blue area 95%. In figure (b) we plot the mean value and variance of I (infected), R (reported) and U (unreported) computed from the IBM. The dark areas contains 68% of the trajectories, and the light areas 95%. In figure (c) we plot the mean value and variance of RR (removed) computed from the IBM. The dark green area contains 68% of the trajectories, and the light green area 95%. In figure (d) we plot the mean value and variance of CR (cumulated reported) computed from the IBM. The dark gray area contains 68% of the trajectories, and the light gray area 95%. We use* 500 *independent runs of the IBM simulations. The parameters we used for both computations are the following:* $I_0 = 93$, $U_0 = 5$, $S_0 = 1.40005 \times 10^9 - (I_0 + U_0)$, $R_0 = RR_0 = CR_0 = 0$ *and* $f = 0.8$, $\tau_0 = 3.3655 \times 10^{-10}$, $N = 26$, $\mu = 0.148$, $\nu = \frac{1}{7}$, $\eta = \frac{1}{7}$, $t_0 = 13.3617$.

5.4 Supplementary tables

We use cumulative reported data from the National Health Commission of the People's Republic of China and the Chinese CDC for mainland China. Before February 11, the data was based on laboratory confirmations. From February 11 to February 15, the data included cases that were not tested for the virus, but were clinically diagnosed based on medical imaging (patients that showed signs of pneumonia). There were 17,409 such cases from February 11 to February 15. The data from February 11 to February 15 specified both types of reported cases. From February 16, the data did not separate the two types of reporting, but reported the sum of both types. We therefore subtracted 17,409 cases from the cumulative reported cases after February 15 to obtain approximate data for the cumulative numbers of reported cases based only on laboratory confirmations after February 15.The data is given in Table 4 with this adjustment.

January						
19	20	21	22	23	24	25
198	291	440	571	830	1287	1975
26	27	28	29	30	31	
2744	4515	5974	7711	9692	11791	
February						
1	2	3	4	5	6	7
14380	17205	20438	24324	28018	31161	34546
8	9	10	11	12	13	14
37198	40171	42638	44653	46472	48467	49970
15	16	17	18	19	20	21
51091	70548 − 17409	72436 − 17409	74185 − 17409	75002 − 17409	75891 − 17409	76288 − 17409
22	23	24	25	26	27	28
76936 − 17409	77150 − 17409	77658 − 17409	78064 − 17409	78497 − 17409	78824 − 17409	79251 − 17409
29						
79824 − 17409						
March						
1	2	3	4	5	6	7
79824 − 17409	79824 − 17409	79824 − 17409	80409 − 17409	80552 − 17409	80651 − 17409	80695 − 17409
8	9	10	11	12	13	14
80735 − 17409	80754 − 17409	80778 − 17409	80793 − 17409	80813 − 17409	80824 − 17409	80844 − 17409
15	16	17	18			
80860 − 17409	80881 − 17409	80894 − 17409	80928 − 17409			

Table 4: *Cumulative data describing confirmed cases in mainland China from January 20, 2020 to March 18, 2020.*

t_1	26	33	40	47	54	61
date	Jan. 27	Feb. 3	Feb. 10	Feb. 17	Feb. 24	Mar. 2
$\text{diff}(t_1)$	2.9×10^{-3}	2.1×10^{-3}	2.9×10^{-3}	1.8×10^{-3}	2.5×10^{-3}	1.4×10^{-3}
t_1	68	75	82			
date	Mar. 9	Mar. 16	Mar. 23			
$\text{diff}(t_1)$	1.6×10^{-3}	1.2×10^{-3}	0.00			

Table 5: *Absolute difference between the cumulative distribution given by the stochastic simulations and the reference simulation $t_1 = 82$. For each t_1 we computed the error as $\text{diff}(t_1) = \sup_{t \geq t_1} |f_{t_1}(t) - f_{81}(t)|$, where f_{t_1} is the estimated distribution computed simulations, for which the initial condition correspond to the components of (2.1) at $t = t_1$ rounded to the closest integer.*

t_1	26	33	40	47	54	61
date	Jan. 27	Feb. 3	Feb. 10	Feb. 17	Feb. 24	Mar. 2
$\text{diff}(t_1)$	8.6×10^{-1}	4.4×10^{-1}	1.7×10^{-1}	6.4×10^{-2}	2.5×10^{-2}	8.1×10^{-3}
t_1	68	75	82			
date	Mar. 9	Mar. 16	Mar. 23			
$\text{diff}(t_1)$	3.5×10^{-3}	8.5×10^{-4}	5.7×10^{-4}			

Table 6: *Absolute difference between the cumulative distribution given by the stochastic simulations and the analytic approximation using the approximate model (3.1), computed using equation (3.5).*

t_1	t_0	18	22	26	33	40
date	Jan. 14	Jan. 19	Jan. 23	Jan. 27	Feb. 3	Feb.10
$\max_{t \geq t_1} \sigma(t)$	3717	1685	787	401	186	106

Table 7: *Maximal standard deviation for the components I, R and U computed by stochastic simulations started at date t_1 with initial condition given by the solution to (2.1) with the parameters from Table 2. The ODE model (2.1) is solved up to $t = t_1$, and we take the solution to (2.1) at $t = t_1$ as initial condition for the stochastic simulations. $\sigma(t)$ is the maximum, at time t, of the standard deviations of the quantities $I(t)$, $R(t)$ and $U(t)$ in a sample of $n = 1000$ independent simulations started at $t = t_1$, and is expressed in number of individuals. We took $f = 0.8$ and other parameters are taken from Table 2.*