Forecasting COVID-19: Using SEIR-D quantitative modelling for healthcare demand and capacity

Eduard Campillo-Funollet*[1], James Van Yperen[2], Phil Allman[3], Michael Bell[4], Warren Beresford[5], Jacqueline Clay[6], Graham Evans[7], Matthew Dorey[6], Kate Gilchrist[4], Pannu Gurprit[8], Ryan Walkley,[6], Mark Watson[9], and Anotida Madzvamuse†[2]

[1]School of Life Sciences, Centre for Genome Damage and Stability, University of Sussex, BN1 9QH, Brighton, UK
[2]School of Mathematical and Physical Sciences, Department of Mathematics, University of Sussex, BN1 9QH, Brighton, UK
[3]NHS Sussex Commissioners, Wicker House, Worthing, UK
[4]Public Health Intelligence team, Public Health, Health and Adult Social Care, Brighton & Hove City Council, 2nd Floor, Hove Town Hall, Norton Road, Hove, BN3 3BQ, UK
[5]Planning and Intelligence, Brighton and Hove, Sussex Commissioners, East Sussex, UK
[6]Public Health and Social Research Unit, West Sussex County Council, 1st Floor, The Grange, Tower Street, Chichester, West Sussex, PO19 1RG, UK
[7]Public Health Intelligence, East Sussex County Council, St Annes Crescent, Lewes, BN7 1UE, UK
[8]East Sussex Health Care Trust (banker) on behalf of Our Care Connected, Sussex Health and Care Partnership, First Floor, Millview Hospital, Nevill Avenue, Hove, East Sussex, BN3 7HY, UK
[9]Sussex Health & Care Partnership, 36-38 Friars Walk, Lewes, BN7 2PB, UK

Abstract

Rapid evidence-based decision-making and public policy based on quantitative modelling and forecasting by local and regional National Health Service (NHS-UK) managers and planners in response to the deadly severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2), a virus causing COVID-19, has largely been missing. In this pilot study, we present a data-driven epidemiological modelling framework that allows to integrate quantitative modelling, validation and forecasting based on current available local and regional datasets to investigate and mitigate the impact of COVID-19 on local NHS hospitals in terms of healthcare demand and capacity as well as allowing for a systematic evaluation of the predictive accuracy of the modelling framework for long-term forecasting. We present an epidemiological model tailored and designed to meet the needs of the local health authorities, formulated to be fitted naturally to datasets which incorporate regional and local demographics. The model yields quantitative information on the healthcare demand and capacity required to manage and mitigate the COVID pandemic at the regional level. Furthermore, the model is rigorously validated using partial historical datasets, which is then used to demonstrate the forecasting power of the model and also to quantify the risk associated with the decision taken by healthcare managers and planners. Model parameters are fully justified, these are derived purely based on the time series data available at the regional level, with minimal assumptions. Using these inferred parameters, the model is able to make predictions under which secondary waves and re-infection scenarios could occur. Hence, our modelling approach addresses one of the major criticisms associated with the lack of transparency and precision of current COVID-19 models. Our approach offers a robust quantitative modelling framework where the probability of the model giving a wrong or correct prediction can be quantified.

*Corresponding author: E.Campillo-Funollet@sussex.ac.uk
†Corresponding author: a.madzvamuse@sussex.ac.uk

1 Introduction

The world is at the mercy of the severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2), the virus causing COVID-19, whose roots originated from Wuhan, China where it was identified in December 2019 [12]. Since then, COVID-19 has swiftly and rapidly spread to all countries in the world, becoming an ongoing global world pandemic that has required unprecedented international, national and regional interventions to try and contain its spread [12, 5]. Unlike the 1918-19 H1N1 pandemic which is considered one of the greatest medical disasters of the 20th century [12], the spread of COVID-19 has taken place in front of our own eyes, live on multimedia platforms with realtime updates, statistics and with remarkable reporting accuracy [10] and yet reliable, accurate and data-validated epidemiological modelling with forecasting and prediction capabilities remains largely out of reach [4, 12, 13, 18, 28]. Given the lack of pharmaceutical interventions such as vaccinations and antiviral drugs, epidemiological modelling has been thrust to the forefront of world organisations and governments' response, rapid decision-making and public policy [3, 4, 12]. Until these pharmaceutical interventions become widely available, the only measures for infection prevention are self or group-isolation, contact tracing, quarantine, physical distancing, decontamination, and hygiene measures. A lot of these unprecedented decisions have resulted in complete lockdowns of countries, economies and a halt to 21st-century lifestyle as we know it and yet these decisions were based on qualitative predictions based on national datasets alien to the countries imposing the lockdowns. A fair criticism of the underlying approach has been the lack of rigorous model validation given the datasets available at the time of the study, the lack of risk assessment associated with the decisions and their impact on the healthcare delivery, demand and capacity and subsequently the lack of precision forecasting that is driven by data. Another criticism is the arbitrary choice of the assumptions and parameters inherent in the epidemiological models which makes it almost impossible to validate the predictions when these were made [3, 4, 12, 18, 28]. At the forefront of these epidemiological models that have played a pivotal role in guiding public policy and national healthcare responses that include the current social distance measures, contact tracing, isolation, and quarantine measures include the now well documented Imperial College London model [12]. The economic impact of these decisions have hardly been quantified, only estimates in the range of trillion of dollars to the world economy are reported [11, 17]. A few models dealing with decision-making within the COVID-19 crisis have been reported [1, 3], however, these lack the power of model prediction and forecasting based on appropriate datasets.

In order to understand the temporal dynamics of COVID-19, a lot of modelling work has been undertaken focusing primarily on national datasets from China, Italy, Spain, UK, the USA and so on [3, 4, 9, 12, 15, 18, 19, 28]. Given the inhomogeneous nature of such datasets, accurate predictions and forecasting of the spread of COVID-19, have so far not been possible and where such predictions were made, footnotes accompanied these predictions simply because of the lack of rigorous mathematical and statistical validation of the models and the lack of robust data on which mathematical assumptions are based on [3, 4, 12, 15, 18, 19, 28]. Forecasting requires ample historical datasets, which were lacking during the first wave of COVID-19. Current-state-of-the-art forecasting models are based, on one hand, on time series analysis without an underlying dynamic epidemiological model [18, 21]. On the other hand, where forecasting is based on epidemiological models [18], these lack rigorous validation, sensitivity analysis, analysis with respect to identifiability of parameters and therefore have limited forecasting power. An interesting approach is proposed in [4] where three models were presented depending on the forecasting timescales; an exponential growth model, a self-exciting branching process, and the classical susceptible-infected-recovered (SIR) compartmental model. The exponential growth model is assumed valid at the early stages of the pandemic, the self-exciting branching process models the individual count data going into the development of the pandemic, and the SIR is a macroscopic mean-field model the describes the pandemic dynamics as it approaches the peak of the disease. In that study, model parameters are inferred by fitting the models to local datasets using maximum Poisson likelihood regression coupled with grid search techniques and a nonparametric expectation maximization algorithm to fit the model to the branching processes [4]. The study however, does not indicate what model-data validation approaches were used, if any, nor does it address the issue of parameter identifiability for the models. Another interesting and alternative approach is to build

machine learning and artificial intelligence techniques on top of epidemiological models to allow for model predictions and forecasting [18]. This approach, so far, has been applied to national datasets from the USA and no regional modelling of this type has been undertaken. This forms part of our current studies by extending our proposed methodology to couple with machine learning and artificial intelligence approaches. As mentioned above, the biggest challenge to model validation, prediction and forecasting remains the existence of appropriate datasets that are readily amenable to modelling.

The use of regional datasets is critical for managing and mitigating COVID-19 secondary waves and re-infection within the local communities. Already there is ample evidence that local modelling could help local authorities to plan local lockdowns, restrictions and so. For example, in the USA, all 50 states had started to reopen and relax lockdown restrictions with bars, beaches, cafes, nightclubs and gyms all opening, however, several states are now either putting on hold their efforts to open fully or started to backtrack given the resurgence in COVID-19 infections and the start of secondary waves. Here in the UK, cities such as Leicester, Bradford, Oldham and others are in the midst of experiencing secondary COVID-19 waves and re-infection. In Australia, the city of Melbourne has now gone into a fresh six-week lockdown after a spike of coronavirus cases, with a further surge in infections. During the first wave, Australia was hailed as a global success story in suppressing the spread of Covid-19 and even at the height of the initial outbreak it only reported a little over 600 infections a day. A similar story is emerging in Spain with regions in Catalona undergoing secondary lockdowns. It is not clear in all these countries the usefulness of national models in terms of being able to predict and forecast the emergence of such waves or re-infections locally until they have already taken place, which is already too late. We propose therefore an alternative quantitative predictive approach which gives local (and national) authorities an ability to predict and forecast COVID-19 scenarios based on their current historical datasets to see future dynamic temporal trends of the disease progression for healthcare planning purposes.

In this study, we want to demonstrate the usefulness and utility of a regionally data-driven epidemiological model based on recently acquired regional datasets (involving hospital deaths and patients recovering in hospitals) from the Sussex and Surrey NHS Trusts and Local Authorities (Brighton and Hove City Council, East and West County Councils) to make predictions and forecasting. The approach is based on a modified SIR-type model, that has been formulated to reflect the dynamics of the regional population (of approximately 1.7 million) which is compartmentalised into Susceptible, denoted by $S(t)$, Exposed, $E(t)$, Infected, $I(t)$, Undetected (rather than the usual Asymptomatic), $U(t)$, Recovered, $R(t)$ and the Dead, $D(t)$ (SEIR-D), respectively (see Figure 1), where t denotes time (days, weeks, or months).

The aim of our study is therefore to propose a systematic modelling approach that addresses healthcare demand and capacity within the South East region. Our goal is to carry out healthcare demand modelling that naturally leads to a standardised framework to quantify demand supressed and generated as a result of COVID-19. We seek to make long-term forecasting and predictions to investigate the impact of COVID-19 on healthcare provision and planning within the South East region of the UK and to mitigate long-term changes in local hospital demands as a result of further COVID-19 secondary waves and economic downturn. We use the local datasets collected as a result of the first wave, starting from 24th March 2020 and this data includes cumulative hospitalisation, recovery and deaths. Our approach differs substantially from current-state-of-the-art modelling-forecasting approaches where unknown parameters driving epidemiological models have been based on various assumptions which vary substantially from one model to the other as well as variations between the domain-expertise of the researchers involved in making the assumptions. Instead, we propose that we know nothing about the values or rates of the model parameters, instead, these are inferred through an inverse modelling approach by requiring the model to fit to data in an optimal sense [4, 18, 21]. In this way, by fitting our SEIR-D model to data we obtain the best values of the unknown model parameters (all the parameters shown in Figure 1), accurate to some degree of confidence [7, 26].

This paper is therefore structured as follows. In Figure 1 we have colour coded two separate areas of data provided by the NHS hospitals in the South East region of the UK. The blue arrow denotes death data in hospitals that we use to fit a linear regression between D_H and H as described in Section 2. The red arrows, each of which denotes deaths outside of hospital, admissions to hospital

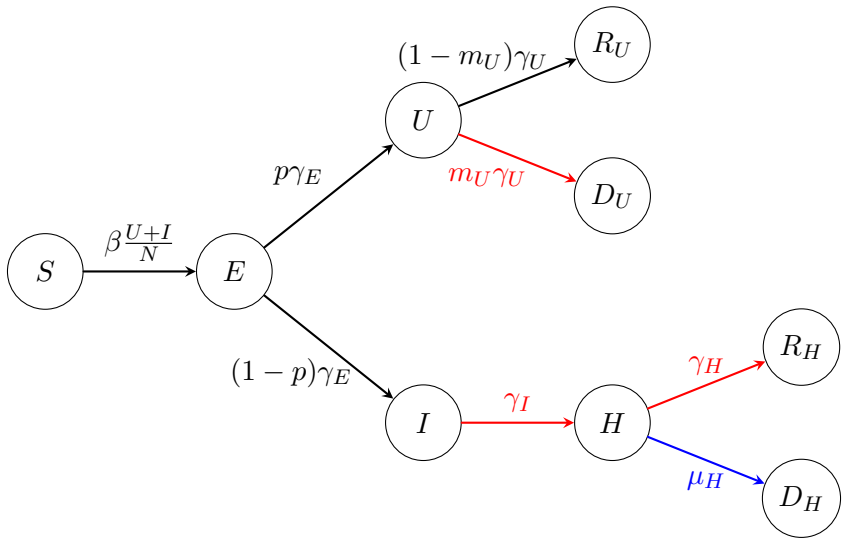

Figure 1: Schematic diagram illustrating the transmission dynamics within the population in the South East region of the UK (Brighton and Hove, East and West Sussex). All arrows indicate the flow of data from one compartment to the other. Colour code: Black arrows indicate the flow with no reliable datasets, red and blue arrows indicate the flow with reliable datasets (hospital and non-hospital). All parameters in the diagram are assumed unknown and must be obtained as part of the model solution procedure (see Sections 3 for details).

and discharges from hospital, represent data used to fit the model as outlined in Section 3. In Section 3.1 we look at the predictive capabilities of the model combined with the inference by comparing model output to new datasets to demonstrate the forecasting accuracy of the proposed approach and it is here where we quantity the accuracy of our forecasting. Section 4 summarises the main findings of our study. Discussions about the wider implications of the data-driven and parameter inference approach for model predictions and forecasting in the context of COVID-19 in particular and epidemiological modelling in general are presented in Section 5.

2 Results

We use the compartmental model described in Figure 1. We designed the model accounting for the available data from hospitals (red) and the death registers (blue) in the area of East Sussex, West Sussex and Brighton and Hove. From the schematic diagram shown in Figure 1 we are interested in finding the best or optimal set of eight model parameters: β, γ_E, p, γ_U, γ_I, γ_H, m_U and μ_H such the SEIR-D model given by equations (3.1)-(3.9) best-fits the observed data. Note that this approach corresponds to finding the maximum likelihood estimation (MLE) for a model with additive noise, where the model itself is deterministic and described by the compartmental model. We estimate the parameters in the model in two steps. First, we exploit the linear relationship, arising from the mathematical model, between mortality in hospitals and discharged patients. Thus we find the parameter $\eta = \mu_H\gamma_H^{-1}$ by fitting the model equation involving the death $D_U(t)$ and $D_H(t)$ compartments, and the recovered $R_U(t)$ and $R_H(t)$ compartments only. The second step is to find the rest of the parameters by reducing the model to a system involving the terms of the populations $U(t)$, $I(t)$ and $H(t)$ only, since it is for these compartments that data is available. See Section 3 for details on the reduced model. The reduced model ensures that all the parameters of interest can be identified from the available data. By means of the minimisation algorithm L-BFGS-B [14, 20], we find the maximum likelihood estimation (MLE) corresponding to the negative log-likelihood given by equation (.37) in Appendix A. We summarise the parameter values in Table 1, where we note that we used $\tilde{\mu}_H = 1 + \mu_H\gamma_H^{-1}$ along with Table 3 to gain μ_H. Figures 2 – 4 show the daily number of patients admitted to hospital who are infected (Figure 2), the daily number of patients who are discharged after recovering while in hospital (Figure 3), and the weekly total number of deaths outside of hospitals, for example, those

dying while at home or in care homes (Figure 4). To demonstrate the accuracy of the fitting procedure we super-impose the observed data sets and their continuum mathematical counterparts, denoted by $C_{Ud}(t)$, $C_{Dis}(t)$ and $C_{D_U}(t)$ in the Methods Section 3, as well their 95% confidence intervals for these curves respectively. One can easily verify that the fitting captures the trends of the data and fits the majority of the data within the confidence interval.

Parameter	Value	Epidemiological meaning
β	0.1420 days^{-1}	Infection rate
γ_E^{-1}	6.48 days	Average latent period
p	0.934	Fraction of undetected cases
γ_U^{-1}	6.06 days	Average infectious period (undetected cases)
γ_I^{-1}	6.31 days	Average infectious period (hospital cases)
γ_H^{-1}	10.21 days	Average hospitalisation period (recovered)
m_U	0.0301	Infected fatality ratio (undetected cases)
μ_H^{-1}	6.67 days	Average hospitalisation period (deaths)

Table 1: Optimal parameter estimates obtained computationally by fitting the SEIR-D model (3.1)-(3.9) to data through a parameter inference approach detailed in Section 3 of the Methods.

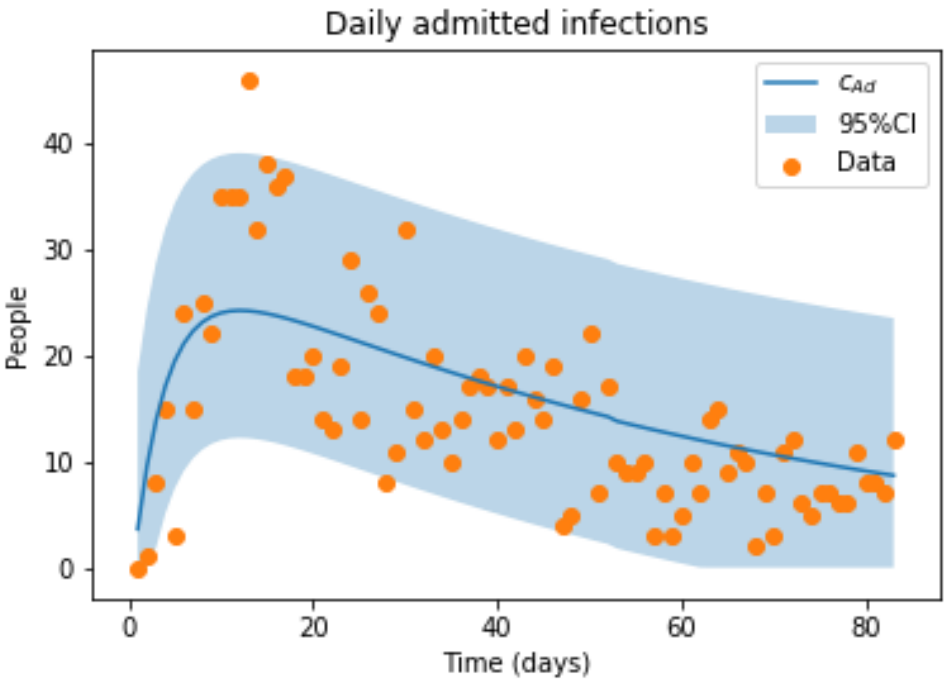

Figure 2: Results of fitting the reduced model defined by (.37) for admissions. The blue line depicts C_{Ud}, the light blue region depicts the 95% confidence interval and the orange points represent the data set C_{Ud}.

We are now in a position to compare and contrast our optimal fitted parameters to those found in the literature [12, 25]. Comparisons between our optimal parameters and those from the literature are included in Table 2. It must be noted that the physical interpretations of some of the parameters differ from one model to the other, however, the overall picture appears plausible. To proceed, we compare our optimal inferred parameters to those published in the research conducted by SAGE [25] and the Imperial College London model [12], both studies used national datasets mainly from Wuhan and other similar infectious diseases. Note that these are national datasets rather than regional datasets, which is our case. From [25] the average incubation period is estimated to be approximately 5.1 days,

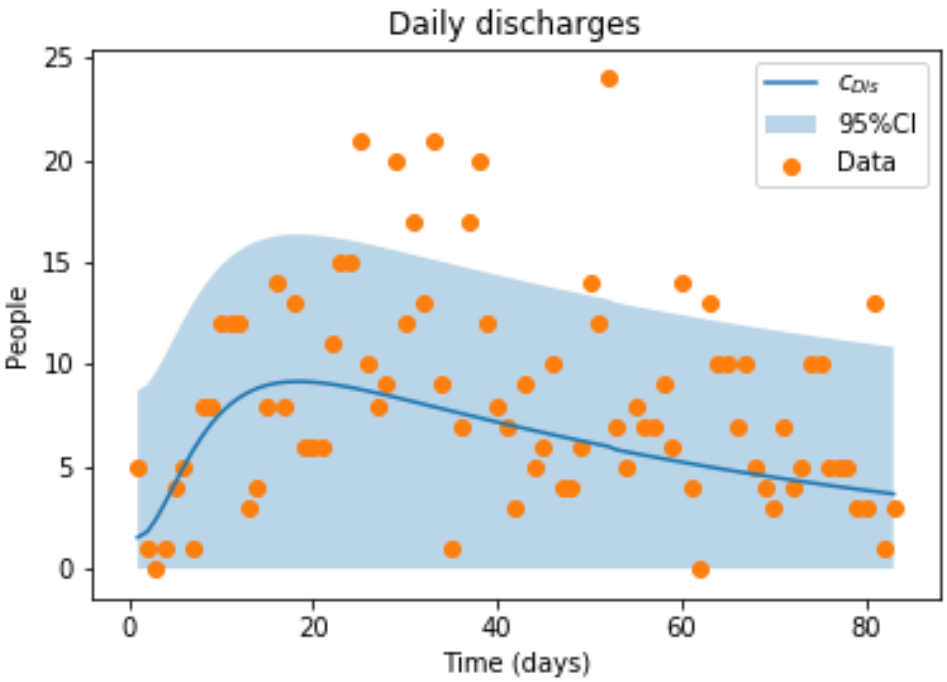

Figure 3: Results of fitting the reduced model defined by (.37) for discharges. The blue line depicts c_{Dis}, the light blue region depicts the 95% confidence interval and the orange points represent the data set C_{Dis}.

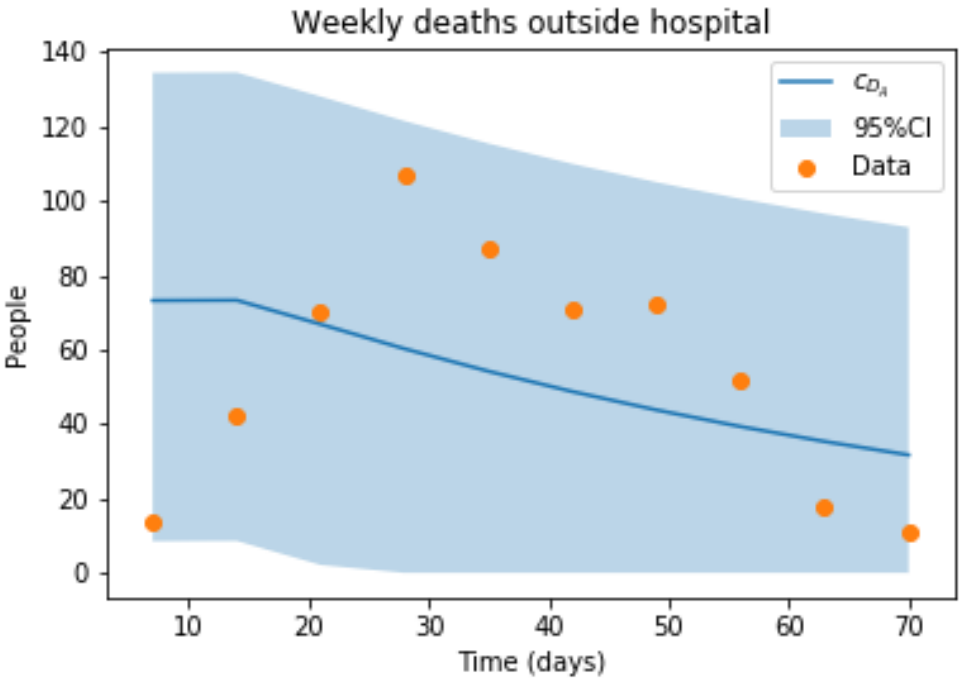

Figure 4: Results of fitting the reduced model defined by (.37) for deaths outside of hospital. The blue line depicts c_{D_U}, the light blue region depicts the 95% confidence interval and the orange points represent the data set $C^w_{D_U}$.

whereby the incubation period is estimated to be in the range of 1 day to 14 days from exposure to onset of symptoms, however the period from exposure to transmissibility is said to be shorter. We estimate that the incubation period, from exposure to transmissibility, is on average around 6.48

	Campillo-Funollet et al. (2020)	Ferguson et al. (2020) (UK-SAGE)
γ_E^{-1}	6.48 days	5.1 days
p	0.934	0.956
γ_I^{-1}	6.31 days	5 days
γ_H^{-1}	10.21 days	8-16 days
m_U	0.0301	0.009*

Table 2: Comparisons between the optimal inferred parameters and some of those currently published in the literature [12]. Note that the mortality ratio m_U in [12] is for all cases (including hospitalisation), whilst in our model refers to only non-hospital cases and is heavily influenced by the mortality in care homes.

days (given by γ_E^{-1}). From [25] the average period from the onset of symptoms to hospitalisation is estimated to be 7 days, to which we have estimated that the time taken from being infectious to being hospitalised is on average around 6.31 days (given by γ_I^{-1}). There is evidence to support that one becomes infectious before presenting symptoms [16, 30] and also that one becomes infectious after presenting symptoms [18]. From [12] the period of onset of symptoms to death is said to range between 18.8 days and 23.9 days, and Wuhan analysis suggested 17.8 days. Our estimate suggests an average of 12.98 days from becoming infectious to dying in hospital (given by $\gamma_I^{-1} + \mu_H^{-1}$) and an average of 200 days outside of hospital (given by $m_U \gamma_U^{-1}$). To our knowledge there has not been a quantified estimate of the contact rate β as well as the death rate outside of hospital, only scaled estimates based on pre-conceived values of $\mathcal{R}_0$ and the recovery rates given by [12] and similar reports. Estimates for $1-p$, the probability of needing hospital treatment, from [5, 12, 27] are around 4.4% in comparison to our inferred estimate which for $1-p$ is estimated to be approximately 6.6%. It must be noted that our set of optimal inferred parameters give a value of effective reproduction number $\mathcal{R}_{eff} = 0.81$ which is similar to that obtained in other UK county datasets (Isle of Wight, private communication).

2.1 Predictive power of the SEIR-D model

To validate the predictive power of our modified SEIR-D model described in Figure 1, we obtain new estimates for the model parameters using only a limited number of data points. We use a minimum of twelve data points since the mortality data is not available until the eleventh daily time point. We evaluate the predictive power of a parameter set by performing a prediction for the next days, starting the day after the last data point used for the parameter estimation. By comparing the prediction with the available data, we compute the percentage of days that are correctly predicted. We consider that a day is correctly predicted if the model output lies within a 95% confidence interval for the available data. Figure 5 shows the results for predictions 10, 20 and 30 days into the future.

The prediction power for admissions into hospital is high even when we only use a few data points to inform the model. For the number of discharged patients, the prediction power is irregular when we use less than 25 data points to fit the model, and is over 90% otherwise, even for predictions 30 days into the future. With only 15 data points, the predictions have an accuracy of almost 80%. Finally, the hospital capacity requires 30 data points to reach a high accuracy, but 15 suffice to reach 85% of correct predictions.

3 Methods

We are now in a position to make a detailed exposition of the methodology underpinning our epidemiological modelling of COVID-19. Our motivation is to present details of the approach that benefit the wider epidemiological community and those interested in fitting models to data, where datasets can be obtained either regionally or otherwise.

Let N denotes the total regional population in the South East region of the UK including only

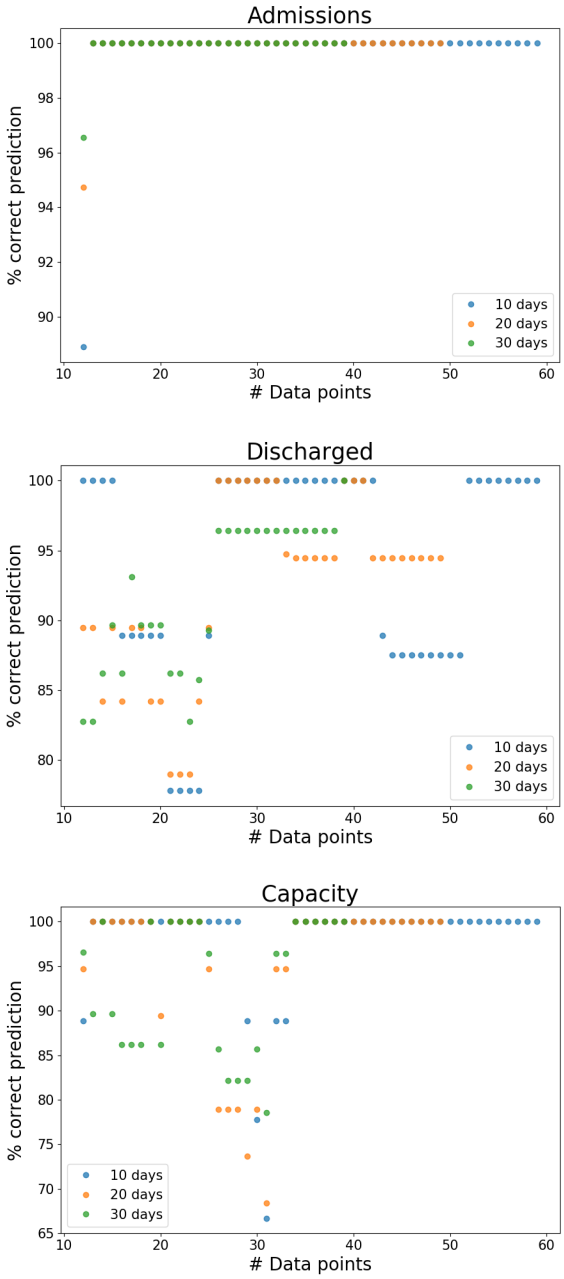

Figure 5: Predictive power of the SEIR-D model for East Sussex, West Sussex and Brighton and Hove, evaluated as the percentage of days predicted within a 95% accuracy.

Brighton and Hove, East and West Sussex (with N approximately 1.7 million). The temporal dynamics of the compartmentalised epidemiological model are depicted in Figure 1, following classical approaches for formulating SIR models [2, 5, 24]. In this setting, $S(t)$ denotes the proportion of the total population N who are susceptible to the disease, COVID-19. These become exposed to the disease to form the $E(t)$ sub-population at rate $\beta\frac{U+I}{N}$, where β is the contact rate between individuals multiplied by the probability of infection during a contact. This sub-population is in incubation period (which is assumed unknown and will be determined as part of the parameter inference approach) and can further evolve in two ways, either a proportion of this sub-population gets infected but remains undetected, denoted $U(t)$, at rate $p\gamma_E$ but does not require hospitalisation or the population gets infected at rate $(1-p)\gamma_E$ and will require hospitalisation. We denote by $I(t)$ the proportion of the total population who are infectious which will require hospitalisation. The sub-population that does not require hospitalisation can either progress to recover at rate $(1-m_U)\gamma_U$ to form the recovered population denoted by $R_U(t)$ or die at rate $m_U\gamma_U$ to form the dead population $D_U(t)$. The dead

population for this compartment comes mainly from the care homes and those that die at home due COVID-19. The infected population becomes hospitalised, denoted by $H(t)$, at rate γ_I. $H(t)$ represents the proportion of the total population that is in hospital care. Once in hospital, patients can evolve in two separate pathways, a proportion of the hospitalised population can fully recover at rate γ_H to form the sub-population $R_H(t)$. Alternatively, if they can not recover, then they die while in hospital at rate μ_H to form the dead population $D_H(t)$. In the spirit of epidemiological models of this nature [5], β denotes the contact rate between individuals, γ_E^{-1} denotes the incubation time, p denotes the proportion of infected individuals who will not require hospital treatment, γ_U^{-1} denotes the recovery time, γ_I^{-1} denotes the time from being infectious to being admitted to hospital, γ_H^{-1} denotes the recovery time for those in hospital and μ_H represents the death rate for those in hospital.

The mathematical translation or interpretation or modelling of the schematic diagram in Figure 1, given the rates described above, leads to the following temporal epidemiological dynamical system modelled by a system of ordinary differential equations supported by non-negative initial conditions

$$\dot{S} = -\beta \frac{S}{N}(U+I), \qquad t \in (0,T], \qquad S(0) = S_0, \tag{3.1}$$

$$\dot{E} = \beta \frac{S}{N}(U+I) - \gamma_E\, E, \quad t \in (0,T], \qquad E(0) = E_0, \tag{3.2}$$

$$\dot{U} = p\,\gamma_E E - \gamma_U U, \qquad t \in (0,T], \qquad U(0) = U_0, \tag{3.3}$$

$$\dot{I} = (1-p)\gamma_E E - \gamma_I I, \qquad t \in (0,T], \qquad I(0) = I_0, \tag{3.4}$$

$$\dot{H} = \gamma_I I - (\gamma_H + \mu_H)H, \quad t \in (0,T], \qquad H(0) = H_0 \tag{3.5}$$

$$\dot{R}_U = (1-m_U)\gamma_U U, \qquad t \in (0,T], \qquad R_U(0) = R_{U,0}, \tag{3.6}$$

$$\dot{R}_H = \gamma_H H, \qquad t \in (0,T], \qquad R_H(0) = R_{H,0}, \tag{3.7}$$

$$\dot{D}_U = m_U \gamma_U U, \qquad t \in (0,T], \qquad D_U(0) = D_{U,0}, \tag{3.8}$$

$$\dot{D}_H = \mu_H H, \qquad t \in (0,T], \qquad D_H(0) = D_{H,0}. \tag{3.9}$$

Model system (3.1)–(3.9) follows the general principles of SIR modelling approaches with one clear difference in that this model system is data-driven formulated where we have highlighted in colour those compartments or pathways where data is available within our local region. The physical justification of the SEIR-D model above is well grounded in the modelling literature for COVID-19 and the general theory of epidemiology [2, 5, 24]. It is known, for example, that those infected with COVID-19 have an incubation period whereby individuals are not infectious themselves [18] giving rise to the $E(t)$ compartment and that not all infected persons present themselves, i.e. individuals may be infected but do not show symptoms, giving rise to the $U(t)$ compartment. Many models have split the $U(t)$ compartment into two separate compartments (see for example [5]), one to describe individuals who are asymptomatic and the other to describe individuals who have symptoms but do not require hospitalisation. However this approach is constrained by the lack of reliable datasets in these compartments and therefore models of this nature rely purely on the merits of the simulations with no forecasting capabilities. For such models, it is very difficult to obtain reliable data on those who are asymptomatic, especially on the scale of multiple counties. Instead, we split the $U(t)$ compartment into recovered $R_U(t)$ and $D_U(t)$ as, even though we can not gather information on those who have recovered, we have access to reliable death data outside of hospitals. Similarly, the same claim holds true for individuals who die in hospital $D_H(t)$. We added a hospital compartment $H(t)$ into the model as a transition compartment since we have access to hospital admission data (i.e. those moving from being infected to seeking treatment) and hospital discharge data, (i.e. those who have recovered and move into $R_H(t)$).

3.1 Inferring model parameters given hospital datasets

Now that we have formulated the full model system to be studied, we next present the methodology for parameter inference given reliable datasets as described in the previous section. For ease of exposition of the methodology, we present two cases since this is the novel aspect of our inference approach. In both cases, we explore the relationship between model parameters where we have access to reliable

datasets to mitigate parameter identifiability issues [6, 22, 29]. In both cases, we employ linear regression methods to estimate the general relationships between model parameters where data is available. We note that Bayesian approaches [7, 26] can be treated in a similar fashion and this forms part of our current studies. An important aspect of this analysis is that the parameters involved do not depend, in terms of this analysis, on any of the other model parameters, thus allowing us to include them in the model analysis in Section 2.

4 Conclusion

In this paper we have derived a novel model based on data presented to us by NHS and Public Health England and conducted linear regression analysis as well as a novel technique to deduce a reduced model to fit the model parameters. Furthermore we have demonstrated that the fitting process and resulting parameters allow us to produce forecasts for quantities of interest such as hospital capacity, which the underlying dynamics fit the pattern of an infectious disease outbreak, rather than rely on statistically inferred parameters which have no verified ability to do long term forecasting. We look to further this work by considering more compartments as we receive more data. For example, we would like to use the care home data to get an idea of infectious spread within care homes. Given the recent research conducted by [19] we also look to include a critical care pathway to model the use of ventilators and see what impact this has on the death rate within hospitals. Another study furthering this work would be to consider model selection to fit multiple values of β and to see if we can detect changes in the public's behaviour, due to a change in policy such as lifting parts of the lockdown for example. This would also allow some scope for an early warning detection system to help forecast potential infection spikes in the system.

5 Discussion

Predicting local COVID-19 outbreaks has emerged as the number one priority by governments and local authorities around the UK and the rest of the world in trying to halt re-infection within the local and national populations. The pandemic itself has thrown to the forefront of science the role of epidemiological modelling when trying to provide novel solutions when questions of urgency, national importance and uncertainty collide thereby exposing its current limitations in terms of predictions and forecasting [23]. A comment in Nature by Salelli et al. [23] outline a manifesto highlighting five ways in which mathematical models should serve society. These include minding the assumptions (the minimal the better), be mindful of model complexities (hubris - balancing the usefulness of the model with the breath of its predictions), be mindful of the interests of the researchers (techniques and methodology can be limited in scope to the expertise of the researchers), beware of the consequences (mitigate the uncertainty), and finally be mindful of the unknowns (communicating what is unknown is as important as communicating what is known). Our approach is based on these five pillars to ensure that our research outcomes are engrained and driven by reliable local datasets with minimal assumptions and an explicit simple data-formulated model. Predictive epidemiological modelling applied to local datasets has the unique ability to offer local authorities a framework for decision-making that is based on temporal trends of these local datasets. Modelling lessons learnt at the regional level can hopefully be transferred to the national arena to help guide data acquisition such that datasets are amenable to model-data prediction approaches as well as providing avenues for short-, medium- and long-term forecasting.

To-date, a lot of models have failed to make meaningful quantitative predictions and forecasts about the impact of COVID-19 despite employing huge amounts of resources and highly sophisticated tools [4, 12, 18, 21, 27]. During the early stages of COVID-19, parallels between COVID-19 and the Spanish flu (among other influenza diseases) that killed more than 50 million people with average age of 28 years, were drawn [4, 12, 18, 21, 27]. As a result, to mitigate and prepare for COVID-19 deaths and infection, national governments and hospitals suspended or postponed important critical treatments, such as cancer treatments, mental health suffered enormously, patients with debilitating conditions avoided visiting hospitals and yet, locally, the number of COVID-19 deaths were nowhere near the

expected numbers predicted by the national models driving the national government decision-making process [4, 12, 18, 21, 27]. Recent studies have highlighted how predictions need to be transparent and humble in order to instil confidence and invite insight and not blame [23]. For a disease such as COVID-19, espoused wrong predictions can have a devastating effect on billions of people around the world in terms of the economy, health, education and societal turmoil, just to mention a few.

It is clear from the literature that the lack of predictions and forecasting is closely correlated to the underlying theoretical assumptions and the use of pre-determined values of the parameters that are alien to the models under study [23]. This in turn is driven by the lack of reliable datasets appropriate for model-data validation and sensitivity analysis. We have proposed in this study a bottom-up approach were a regional model built on local datasets has the ability to guide local decision making in terms of healthcare demand and capacity, in particular given the likelihood of COVID-19 secondary waves and re-infection. Our modelling framework is not only tailored to deal with COVID-19, but can be applied to general epidemiological winter diseases which are known to kill thousands of patients every year. Epidemic forecasting and the development of early warning systems for healthcare demand and capacity has been thrown at the forefront of epidemiological modelling, by working in close collaboration, theoreticians and local authorities and planners have a unique opportunity to bring novel approaches to healthcare decision-making and planning.

References

[1] Daron Acemoglu, Victor Chernozhukov, Iván Werning, and Michael D Whinston. A multi-risk sir model with optimally targeted lockdown. Technical report, National Bureau of Economic Research, 2020.

[2] Linda JS Allen and Amy M Burgin. Comparison of deterministic and stochastic sis and sir models in discrete time. *Mathematical biosciences*, 163(1):1–33, 2000.

[3] Fernando E Alvarez, David Argente, and Francesco Lippi. A simple planning problem for covid-19 lockdown. Technical report, National Bureau of Economic Research, 2020.

[4] Andrea L. Bertozzi, Elisa Franco, George Mohler, Martin B. Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 2020.

[5] Konstantin B Blyuss and Yuliya N Kyrychko. Effects of latency and age structure on the dynamics and containment of covid-19. *medRxiv*, 2020.

[6] Thierry Boileau, Nicolas Leboeuf, Babak Nahid-Mobarakeh, and Farid Meibody-Tabar. Online identification of pmsm parameters: Parameter identifiability and estimator comparative study. *IEEE transactions on industry applications*, 47(4):1944–1957, 2011.

[7] Eduard Campillo-Funollet, Chandrasekhar Venkataraman, and Anotida Madzvamuse. Bayesian parameter identification for turing systems on stationary and evolving domains. *Bulletin of mathematical biology.*, 81(1):81–104, 2019.

[8] Odo Diekmann, JAP Heesterbeek, and Michael G Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, 7(47):873–885, 2010.

[9] Annemarie B Docherty, Ewen M Harrison, Christopher A Green, Hayley E Hardwick, Riinu Pius, Lisa Norman, Karl A Holden, Jonathan M Read, Frank Dondelinger, Gail Carson, et al. Features of 16,749 hospitalised uk patients with covid-19 using the isaric who clinical characterisation protocol. *medRxiv*, 2020.

[10] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[11] Martin S Eichenbaum, Sergio Rebelo, and Mathias Trabandt. The macroeconomics of epidemics. Technical report, National Bureau of Economic Research, 2020.

[12] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg, et al. Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. *Technical Report*, 2020.

[13] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020.

[14] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[15] Karl J Friston, Thomas Parr, Peter Zeidman, Adeel Razi, Guillaume Flandin, Jean Daunizeau, Oliver J Hulme, Alexander J Billig, Vladimir Litvak, Rosalyn J Moran, et al. Dynamic causal modelling of covid-19. *arXiv preprint arXiv:2004.04463*, 2020.

[16] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance*, 25(17):2000257, 2020.

[17] Callum J Jones, Thomas Philippon, and Venky Venkateswaran. Optimal mitigation policies in a pandemic: Social distancing and working from home. Technical report, National Bureau of Economic Research, 2020.

[18] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science*, 368(6493):860–868, 2020.

[19] Jose Lourenco, Robert Paton, Mahan Ghafari, Moritz Kraemer, Craig Thompson, Peter Simmonds, Paul Klenerman, and Sunetra Gupta. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the sars-cov-2 epidemic. *MedRxiv*, 2020.

[20] Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[21] Fotios Petropoulos and Spyros Makridakis. Forecasting the novel coronavirus covid-19. *PloS one*, 15(3):e0231236, 2020.

[22] Andreas Raue, Johan Karlsson, Maria Pia Saccomani, Mats Jirstrand, and Jens Timmer. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*, 30(10):1440–1448, 2014.

[23] Andrea Saltelli, Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, et al. Five ways to ensure that models serve society: a manifesto, 2020.

[24] Junkichi Satsuma, R Willox, A Ramani, B Grammaticos, and AS Carstea. Extending the sir epidemic model. *Physica A: Statistical Mechanics and its Applications*, 336(3-4):369–375, 2004.

[25] Anthony C Smith, Emma Thomas, Centaine L Snoswell, Helen Haydon, Ateev Mehrotra, Jane Clemensen, and Liam J Caffery. Telehealth for global emergencies: Implications for coronavirus disease 2019 (covid-19). *Journal of telemedicine and telecare*, page 1357633X20916567, 2020.

[26] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451, 2010.

[27] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 2020.

[28] Russell M Viner, Simon J Russell, Helen Croker, Jessica Packer, Joseph Ward, Claire Stansfield, Oliver Mytton, Chris Bonell, and Robert Booy. School closure and management practices during coronavirus outbreaks including covid-19: a rapid systematic review. *The Lancet Child & Adolescent Health*, 2020.

[29] Hulin Wu, Haihong Zhu, Hongyu Miao, and Alan S Perelson. Parameter identifiability and estimation of hiv/aids dynamic models. *Bulletin of mathematical biology*, 70(3):785–799, 2008.

[30] Wei Xia, Jiaqiang Liao, Chunhui Li, Yuanyuan Li, Xi Qian, Xiaojie Sun, Hongbo Xu, Gaga Mahai, Xin Zhao, Lisha Shi, et al. Transmission of corona virus disease 2019 during the incubation period may lead to a quarantine loophole. *MedRxiv*, 2020.

Appendix A: Methodology for parameter inference

The parameter inference is a two-step approach, first we infer the parameters where data is available and model parameters can be identified independently, then the second step is to use these parameters to infer the rest of the model parameters to yield the full optimal set of model parameters.

Step 1 – Using regression analysis to infer hospitalisation discharge and death rates

Given that we have hospital discharge data and those that die in hospital, we start by considering (3.5), where we describe the number of individuals moving from the hospital compartment H to the hospital recovered R_H and then we describe the number of individuals moving from the hospital compartment to the death in hospital compartment D_H. We then seek to find a linear relationship between the two descriptions and use linear regression analysis to estimate the relationship between them, giving an estimate of the parameters involved to be used in Section 2.

We begin by considering the hospital discharged data which is available for 84 consecutive days starting March 24th, 2020, apart from one missing data point on May 15th. We denote the set of data points by C_H, with each data point denoted as $C_{H,i}$, for $i = 1, \ldots, 84$. By considering (3.5) and (3.7), the rate of individuals being discharged in a given day from hospital is given by $\gamma_H H$. This allows us to compute the number of discharges per day in the following way

$$c_H(t) = \gamma_H \int_{t-1}^{t} H, \quad t \geq 1, \tag{.1}$$

where t is a given day. We assume that the data is perturbed with Gaussian noise of unknown mean, denoted m_H, and unknown standard deviation, denoted σ_D, which leads to the following relationship

$$C_{H,i} = c_H(i) + m_H + \xi_{H,i}, \tag{.2}$$

where i denotes a time point of the data and $\xi_{H,i} \sim \mathcal{N}(0, \sigma_H^2)$.

We now consider the data for those who die in hospitals. The data for deaths in hospitals is available on a weekly basis, starting on week 14 (week ending April 3rd, 2020) up to week 23 (week ending June 5th, 2020). We denote the set of data points by $C^w_{D_H}$, with each data point denoted as $C^w_{D_H,j}$, for $j = 14, \ldots, 23$. By considering (3.5) and (3.9), the rate of individuals dying in hospital in a given day is given by $\mu_H H$. In a similar fashion, we can calculate the number of deaths per day in the following manner

$$c_{D_H}(t) = \mu_H \int_{t-1}^{t} H, \quad t \geq 1. \tag{.3}$$

Since the data is weekly rather than daily, the number of deaths per week is calculated by

$$c^w_{D_H}(\tau) = \mu_H \int_{\tau-7}^{\tau} H, \quad \tau \geq 7(\text{weekly}). \tag{.4}$$

We again assume that the data is perturbed with Gaussian noise of unknown mean, denoted by m_D, and unknown standard deviation, denoted by σ_D, which leads to the following relationship

$$C^w_{D_H,j} = c^w_{D_H}(t_j) + m_D + \xi_{D,j}, \tag{.5}$$

where t_j is the last day of week j and $\xi_{D,j} \sim \mathcal{N}(0, \sigma_D^2)$.

Regression analysis

From (.1) and (.3) we find

$$c_H(t) = \frac{\gamma_H}{\mu_H} c_{D_H}(t), \tag{.6}$$

which is a consequence of equations (3.5), (3.7) and (3.9), and gives light to one of the parameters, $\gamma_H \mu_H^{-1}$, to be estimated. We are going to treat the daily deaths as unknowns in our model, by solving

the least squares problem defined by (.2) and (.5), together with $c^w_{D_H}(\tau) = \sum_{k=0}^{6} c_{D_H}(\tau - k)$. For simplicity, we assume that the covariance for the weekly mortality data is seven times larger than the covariance for the daily discharged data, i.e. $\sigma_D^2 = 7\sigma_H^2$. The rationale of this assumption is that the relationship assumes that on the same time scale, mortality and discharged data would have the same noise levels, and that mortality data is composed of independent daily measurements. We note that this is a conservative assumption, since data collection methods are diverse.

Let $\eta = \gamma_H \mu_H^{-1}$. The corresponding negative log-likelihood for the model given by (.2) and (.5) is

$$F(\eta, m_H, m_D, c_{D_H}; C_H, C^w_{D_H}) := \sum_{i=1, i\neq 53}^{84} (\eta\, c_{D_H}(i) + m_H - C_{H,i})^2 + \frac{1}{7}\sum_{j=14}^{23}\left(\sum_{k=0}^{6} c_{D_H}(t_j - k) + m_D - C^w_{D_H,j}\right)^2. \tag{.7}$$

We minimise F under the constraints $\eta > 0$ and $c_{D_H}(i) > 0$ using the Constrained Optimisation BY Linear Approximation (COBYLA) algorithm. The resulting parameter estimations are presented in Table 3. The linear regression relationship between the discharged data and hospital deaths is shown in Figure 6. Moreover we present the fit for the daily discharged individuals in Figure 7 and the fit for the weekly deaths in hospital in Figure 8. We note here as a reminder that the death data was only available weekly and so the daily death data for Figure 6 was inferred. Moreover we again note that day 0 corresponds to March 24th, 2020, and week 14 corresponds to the week ending April 3rd, 2020.

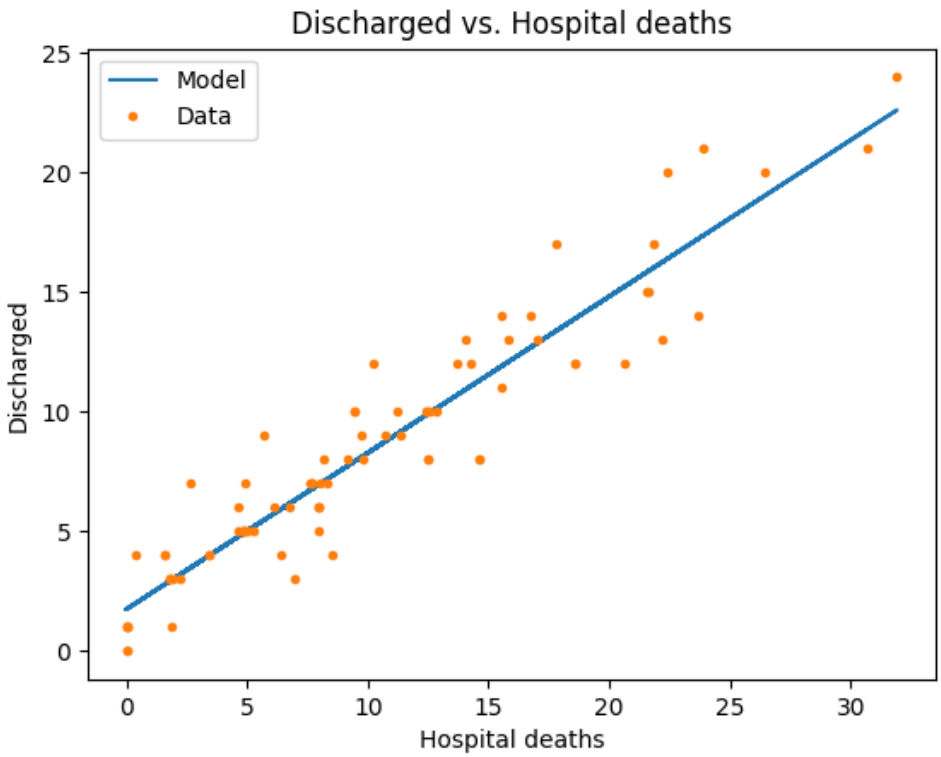

Figure 6: Model and data for the daily discharged patients and daily deaths. Note that only discharged data is available daily; the corresponding daily deaths were inferred (see Figure 8).

.1 Inference for hospital model

In this section we present the inference equations we derived in order to infer the parameters of the model. The concept is as follows: we define variables that describe the data we have, in terms of the model compartment and parameters, and reduce the model to only be in terms of these new variables

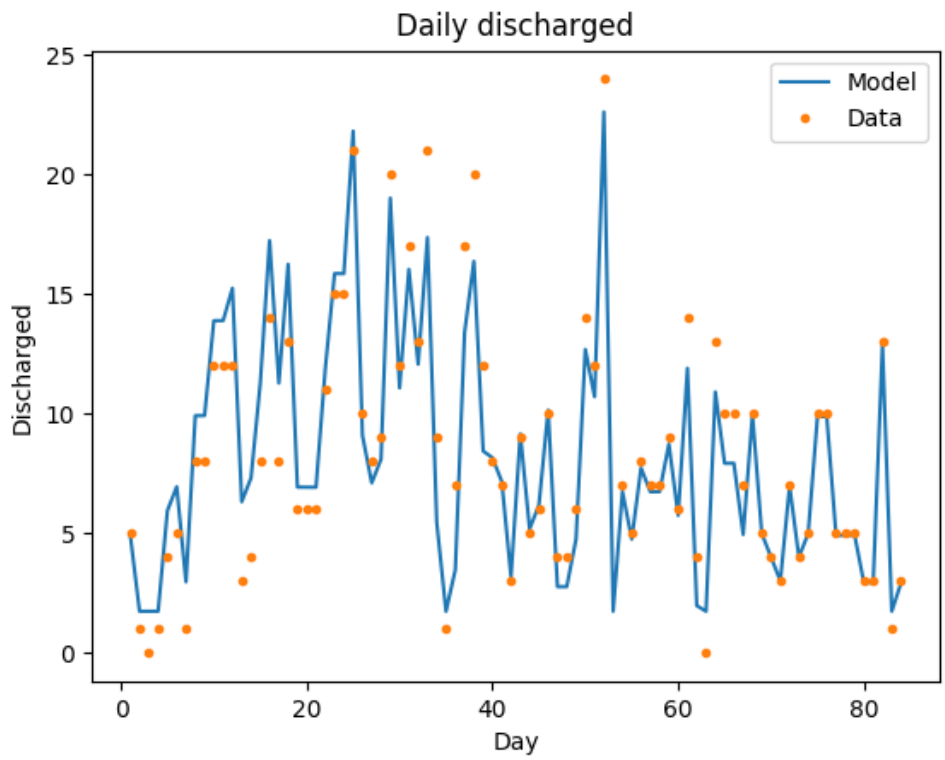

Figure 7: Model and data for the daily discharged patients. Days are counted starting on March 24th, 2020.

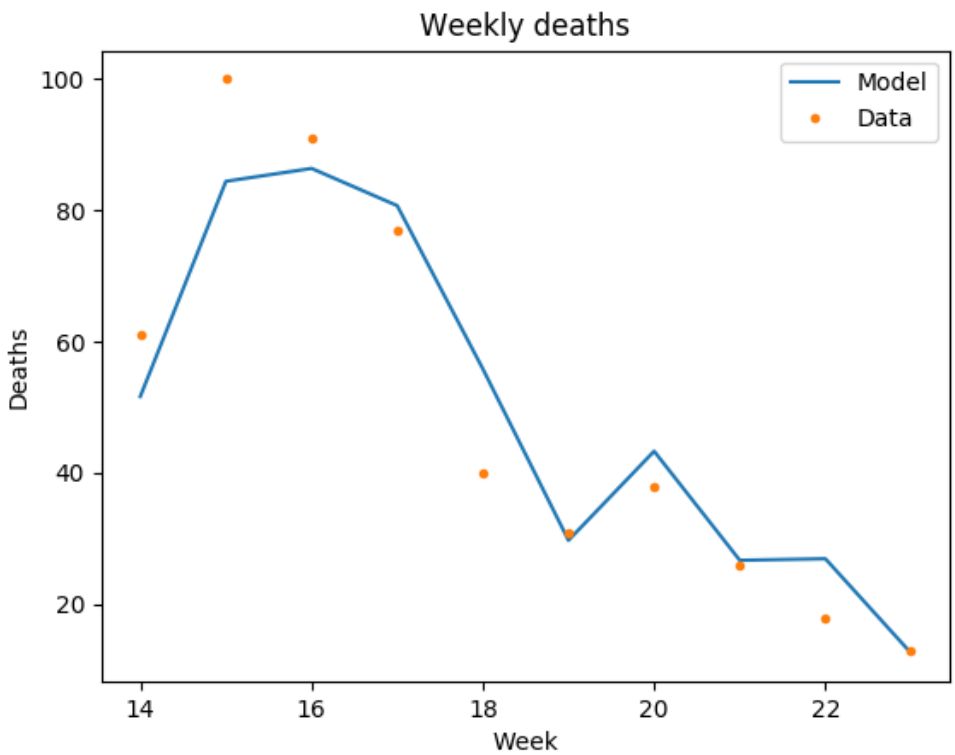

Figure 8: Model and data for the weekly deaths. Weeks are counted starting on the week ending on April 3rd, 2020.

Parameter	Value	Std.
$\frac{\gamma_H}{\mu_H}$	0.653902	0.004186276
m_H	1.748516	0.02085394
m_D	-26.825357	3.183021

Table 3: Results of fitting the linear model defined by (.7) to obtain the best values for $\eta = \gamma_H \mu_H^{-1}$, the means m_H and m_D, respectively.

and the model parameters. This establishes what parameters we can infer from the data as well as reduce the computational power needed to fit the model since we do not need to solve the whole system.

.1.1 Hospital and non-hospital datasets

The data we have to utilise is daily hospital admissions, daily hospital discharges and weekly non-hospital deaths. Note, we do not consider hospital deaths as we have already used this data to find the relationship on $\gamma_H \mu_H^{-1}$. In terms of the hospital data, considering (3.5), we know that $\dot{H} = H_{in} - H_{out} := admitted - (discharged + died)$ and thus

$$C_{Ud}(t) := \gamma_I \int_{t-1}^{t} I(s)\, ds, \tag{.8}$$

describes the daily hospital admissions, and

$$c_{Dis}(t) := \gamma_H \int_{t-1}^{t} H(s)\, ds \tag{.9}$$

describes the daily hospital discharges. Similarly, considering (3.6), we have that

$$c_{D_U}(t) := m_U \gamma_U \int_{t-1}^{t} U(s)\, ds \tag{.10}$$

describes the daily deaths outside of hospital. Note, for ease of notation we consider daily death data rather than weekly, and we have denoted (.9) this way so as not to confuse with (.1), since although it is the same data, the variables are being used in different settings.

Before we derive the reduced system we re-introduce the model for ease of exposition, where we have omitted the deaths and recovered compartments (equations 3.6-3.9), namely

$$\dot{S} = -\tilde{\beta} S(U + I), \tag{.11}$$
$$\dot{E} = \tilde{\beta} S(U + I) - \gamma_E E, \tag{.12}$$
$$\dot{U} = p\gamma_E E - \gamma_U U, \tag{.13}$$
$$\dot{I} = (1 - p)\gamma_E E - \gamma_I I, \tag{.14}$$
$$\dot{H} = \gamma_I I - \tilde{\mu}_H \gamma_H H. \tag{.15}$$

Here, we have only introduced the compartments which are necessary for the derivation of the reduced model. Here we have also denoted $\tilde{\beta} = \beta N^{-1}$ and $\tilde{\mu}_H = 1 + \mu_H \gamma_H^{-1}$ for ease of notation.

.1.2 Reduced system

We first look to manipulate equations (.11)–(.15) to only be in terms of the compartments $U(t)$, $I(t)$ and $H(t)$ and their derivatives so that an easy application of calculus gives us the reduced equations in terms of c_{D_U}, C_{Ud} and c_{Dis} that we require. Since we have three variables, we look to derive three equations in terms of U, I and H. To this extent we will use (.15) and (.13), once we have the

relationship for E, we then use the remaining equations to get the final equation. This means that we need to calculate S and E in terms of U, I and H.

We first note that, using (.14), we have

$$E = \frac{\dot{I} + \gamma_I I}{(1-p)\gamma_E}, \tag{.16}$$

which provides us with a way to describe E in terms of I. What remains is to find the relationship for S. Considering (.11) and (.12) one can see that $\dot{S} = -\dot{E} - \gamma_E E$, and subsequently $\ddot{S} = -\ddot{E} - \gamma_E \dot{E}$. Thus, if we can use the expressions describing $\ddot{S}$, $\dot{S}$, and equation (.16) to gain the final equation we need for the reduced system. This bypasses the need to explicitly solve (.11) to find S and keeps the calculations easier to follow. Indeed, by taking the derivative of (.11) we have

$$\ddot{S} = \frac{\dot{S}}{U+I}\left((\dot{U} + \dot{I}) - \tilde{\beta}(U+I)^2\right). \tag{.17}$$

Now straight substitutions of $\ddot{S}$ and $\dot{S}$ into (.17) yields

$$-\ddot{E} - \gamma_E \dot{E} = -(\dot{E} + \gamma_E E)\frac{(\dot{U} + \dot{I}) - \tilde{\beta}(U+I)^2}{U+I}. \tag{.18}$$

Hence, noting that (.16) gives us

$$\dot{E} = \frac{\ddot{I} + \gamma_I \dot{I}}{(1-p)\gamma_E}, \quad \text{and} \quad \ddot{E} = \frac{\dddot{I} + \gamma_I \ddot{I}}{(1-p)\gamma_E}, \tag{.19}$$

and so, using (.19) in (.18) and using (.16) in (.13), yields the following reduced system

$$\frac{\dddot{I} + \gamma_I \ddot{I}}{(1-p)\gamma_E} + \frac{\ddot{I} + \gamma_I \dot{I}}{1-p} = \left(\frac{\ddot{I} + \gamma_I \dot{I}}{(1-p)\gamma_E} + \frac{\dot{I} + \gamma_I I}{1-p}\right)\frac{\dot{U} + \dot{I} - \tilde{\beta}(U+I)^2}{U+I}, \tag{.20}$$

$$\dot{U} = \frac{p}{1-p}(\dot{I} + \gamma_I I) - \gamma_U U, \tag{.21}$$

$$\dot{H} = \gamma_I I - \tilde{\mu}_H \gamma_H H. \tag{.22}$$

For ease of notation, let us denote by $x := \dot{c}_{Ad}$, $y := \dot{c}_{D_U}$ and $z := \dot{c}_{Dis}$. We can now express $U(t)$, $I(t)$, and $H(t)$ in terms of $x(t)$, $y(t)$ and $z(t)$, respectively, through the change of variables above, since

$$I^{(i)} = \frac{x^{(i)}}{\gamma_I}, \tag{.23}$$

$$U^{(i)} = \frac{y^{(i)}}{m_U \gamma_U}, \tag{.24}$$

$$H^{(i)} = \frac{z^{(i)}}{\gamma_H}, \tag{.25}$$

where $f^{(i)}$ denotes the $i-$th derivative of f. Thus, using (.23)–(.25) in (.20)–(.22), we have

$$\begin{aligned}\dddot{x} &= (\ddot{x} + (\gamma_I + \gamma_E)\dot{x} + \gamma_I\,\gamma_E\,x)\left(\frac{x}{\gamma_I} + \frac{y}{m_U\gamma_U}\right)^{-1} \\ &\quad \times \left(\frac{1}{1-p}\frac{\dot{x}}{\gamma_I} + \frac{p}{1-p}x - \frac{y}{m_U} - \tilde{\beta}\left(\frac{x}{\gamma_I} + \frac{y}{m_U\gamma_U}\right)^2\right) \\ &\quad - (\gamma_I + \gamma_E)\ddot{x} - \gamma_I\,\gamma_E\,\dot{x},\end{aligned} \tag{.26}$$

$$\dot{y} = \frac{m_U\gamma_U\,p}{1-p}\left(\frac{\dot{x}}{\gamma_I} + x\right) - \gamma_U y, \tag{.27}$$

$$\dot{z} = \gamma_H x - \gamma_H \tilde{\mu}_H z. \tag{.28}$$

In the above, we have used (.21) in (.20) to replace the $\dot{U}$ term so we can transform this system into a system of first order differential equations. Indeed, to solve (.26)–(.28), we also need to prescribe initial conditions $x(0) = x_0$, $\dot{x}(0) = \dot{x}_0$, $\ddot{x}(0) = \ddot{x}_0$, $y(0) = y_0$ and $z(0) = z_0$.

.1.3 Inference

Now that we have the reduced system of (.11)–(.15) in terms of the data, we can proceed to fit the data and gain the inferred parameter values by minimising a log-likelihood function. In a similar manner to Section 5 we denote the set of daily hospital admissions data as C_{Ud}, with each data point denoted as $C_{Ud,i}$, for $i = 1, \ldots, 84$. Similarly, we denote the set of daily hospital discharge data as C_{Dis}, with each data point denoted as $C_{Dis,i}$, for $i = 1, \ldots, 84$. Finally, we denote the set of weekly death data outside of hospital as $C^w_{D_U}$, with each data point denoted as $C^w_{D_U,j}$, for $j = 14, \ldots, 23$. Assuming that the data is collected subject to centered Gaussian errors, scaling the mortality data error to account for weekly data, we have the following negative log-likelihood function to minimise

$$F(C_{Ud}, c_{Dis}, c_{D_U}; C_{Ud}, C_{Dis}, C_{D_U}) := \sum_{i=1, i\neq 53}^{84} (C_{Ud}(i) - C_{Ud,i})^2 + \sum_{i=1, i\neq 53}^{84} (c_{Dis}(i) - C_{Dis,i})^2 + \frac{1}{7} \sum_{j=14}^{23} \left(\sum_{k=0}^{6} c_{D_U}(t_j - k) - C^w_{D_U,j} \right)^2, \tag{.29}$$

where we note that day 53 is missing in the hospital data. Without any constraints, (.29) has multiple minimisers. We therefore impose two constraints (which are fully justified physically) based on the following assumptions. First, we assume that the population within the model excluding the recovered and death compartments is between 90% and 100% of the total population N that we consider. This reflects the fact that an unknown fraction of the population is already immune or are not susceptible, due to the shielding programmes in place, for example. Second, we assume that the effective reproductive number $\mathcal{R}_{eff} = \frac{S_0}{N}\mathcal{R}_0$ at the beginning on the simulation is less than one. This reflects the effect of the lockdown on the population dynamics and avoids non-feasible parameters that involve very high infection rates leading to close to 100% of infections in a short period of time, which is unrealistic in the current climate. In order to include these constraints into the log-likelihood we need to describe them in terms of the data and parameters. We first look at the initial conditions.

Considering equations (.23)–(.25) one easily sees that

$$I_0 = \frac{x_0}{\gamma_I}, \tag{.30}$$

$$U_0 = \frac{y_0}{m_U \gamma_U}, \tag{.31}$$

$$H_0 = \frac{z_0}{\gamma_H}. \tag{.32}$$

Using (.16) and (.30) we see that

$$E_0 = \frac{1}{(1-p)\gamma_E} \left(\frac{\dot{x}_0}{\gamma_I} + x_0 \right), \tag{.33}$$

and similarly, (.30) and (.33) leads directly to

$$\dot{E}_0 = \frac{1}{(1-p)\gamma_E} \left(\frac{\ddot{x}_0}{\gamma_I} + \dot{x}_0 \right). \tag{.34}$$

To find S_0, we rearrange (.12) and use (.30), (.31), (.33) and (.34) to yield

$$S_0 = \frac{1}{1-p} \frac{\tilde{\beta}^{-1}}{\frac{x_0}{\gamma_I} + \frac{y_0}{m_U \gamma_U}} \left(\frac{\ddot{x}_0}{\gamma_E \gamma_I} + \left(\frac{1}{\gamma_I} + \frac{1}{\gamma_E} \right) \dot{x}_0 + x_0 \right). \tag{.35}$$

We now turn our attention to the effective reproductive number $\mathcal{R}_{eff}$. In order to calculate the effective reproduction number, we use the method of next-generation matrices derived in [8] to obtain the following expression for $\mathcal{R}_0$:

$$\mathcal{R}_0 := \beta \left(\frac{p}{\gamma_U} + \frac{1-p}{\gamma_I} \right). \tag{.36}$$

Including the constraints, the negative log-likelihood then reads

$$
\begin{aligned}
&G(C_{Ud}, c_{Dis}, c_{D_U}, S_0, E_0, U_0, I_0, H_0; C_{Ud}, C_{Dis}, C_{D_U}, N) \\
&= \sum_{i=1, i\neq 53}^{84} (C_{Ud}(i) - C_{Ud,i})^2 + \sum_{i=1, i\neq 53}^{84} (c_{Dis}(i) - C_{Dis,i})^2 \\
&\quad + \frac{1}{7} \sum_{j=14}^{23} \left(\sum_{k=0}^{6} c_{D_U}(t_j - k) - C_{D_U,j} \right)^2 - w_1(1 - \mathcal{R}_{eff}) \\
&\quad - w_2(N - N_0)(N_0 - 0.9N), \qquad (.37)
\end{aligned}
$$

where, for ease of notation, we have left the initial conditions and reproductive number in their original notation. To aid clarity in notation, we have introduced $N_0 = S_0 + E_0 + U_0 + I_0 + H_0$. We have also included a weighting to ensure positivity of solutions. For the simulation, we took the weights $w_1 = 10^2$ and $w_2 = 10^{-9}$. We minimised (.37) using the Scipy implementation of the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B) [14, 20]. The box constraints were used to ensure positivity of the relevant parameters as well as positivity of the initial conditions.