Understanding the effect of sampling effort on covid-19 case numbers

Candidate Number: LDBD3

Supervisor: Dr. Manolopoulou Ioanna

Department of Statistical Science
University College London

Word count: 10669

August 29, 2021

I, Candidate Number: LDBD3, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In the end of 2019, the new coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), outbroke in China and soon this syndrome was found in the whole world. The general underestimation of coronavirus disease 2019 (COVID-19) in early periods made the UK suffer from the large epidemics. In addition, since no effective vaccine or other pharmaceutical approaches are proposed to contain the spread of the epidemic, the government had to implement non-pharmaceutical interventions in order to prevent it from further developing. This paper mainly studies the sampling effects across the UK for the period from the start of the COVID-19 epidemics in 31^{st} January 2020 until the time that the mutation of SARS-CoV-2, Alpha, was first found in the UK after 1^{st} October 2020. Specifically, combining techniques of epidemic modelling, bayesian inferring and MCMC simulating, the model estimates the transmissions based on either deaths or the increasements of the number of people in Mechanical Ventilation Beds, which were observed in the following weeks. By matching the estimated infection time-series data to the timeline of government policies, together with the searching index in Google for 'Protest', it makes it possible to interpret the effect of the non-pharmaceutical interventions. In the meantime, it is noticeable that the differences between the recorded and 'true' number (estimated by the models) of infections narrows as time flows, which also provides strong evidence that the COVID-19 detection capability of the UK government is increasing over time. Furthermore, the major non-pharmaceutical interventions (NPI), especially lockdowns, are proved to have a large effect on reducing transmission. Moreover, the model further shows that the transmission is potential to increase again after easing the restrictions, indicating that a series of long-term non-pharmaceutical interventions are necessary to keep transmission of SARS-CoV-2 under control.

Acknowledgements

This project would not have been possible without the support of many people. Special thanks to my supervisor dr. Manolopoulou Ioanna, who have read numerous versions of this manuscript and patiently helped me with the perfection of every detail time after time. Thanks to my friends and fellow students, who have always been so supportive and encouraging. Thanks to my family, for the unconditional love and support throughout the entire thesis process and each day in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The global impact of COVID-19 is far-reaching, and it has been considered to be the most serious of respiratory viruses since the H1N1 influenza pandemic in 1918. Evidence shows that SARS-CoV-2 will continue to spread from person to person (Riou and Althaus, 2020). It can cause serious illness (Huang et al., 2020), and the elderly are prone to bear higher risk of severe consequences and even death (Surveillances, 2020). The first two COVID-19 cases in the UK were confirmed on January 31st, 2020. Although the implementation of testing, quarantine and contact tracing may slow down early transmission (Hellewell et al., 2020) this is not enough to contain the outbreak in the UK (Davies et al., 2020). Aside from medical aids, *Non-Pharmaceutical Intervention* (NPI) has played a key role in reducing the basic reproduction number R_t and the impact of COVID-19 in the United Kingdom. R_t is a basic epidemiological number, which represents the average number of infections produced by each infected case during the course of infection (Flaxman et al., 2020). If $R_t = 0.5$ then, on average, for each 2 infected people only 1 person would be newly infected. R_t implies the direction of the epidemic. $R_t > 1$ indicates that the epidemic is growing while $R_t < 1$ indicates the epidemic is shrinking. However, the single R_t does not imply how quickly an epidemic is changing, while an other term: serial interval indicating the average time each transmission to take place (John, 2001). Combing R_t and the serial interval, the direction and the speed of the epidemic change could be ensured. NPI will continue to be the main public health tool against SARS-CoV-2 until all people at risk of COVID-19 can obtain an effective vaccine.

The NPI policy response to COVID-19 can be complex. These policies can be at a country level or a local level. Different countries resort to different methods to manage the pandemic, and it also evolves over time according to the developing situation of infection and economy. In Europe, Italy was the first to carry out NPI policies, and other European countries followed soon afterwards. NPI usually con-

tains closings of schools and universities, closings of workplaces, canceling public events, limits on gatherings, closing of public transport, stay at home requirements, restrictions on internal movement between cities/regions, and restrictions on international travel. However, NPI can have a severe negative impact on people's overall well-being, social operations, and the economy. Therefore, these interventions should be carefully used and guided by data so as to protect the most vulnerable individuals in society. The timeline of NPI policies in the UK taken from Institute for government (2021) is shown in Figure 1.1.

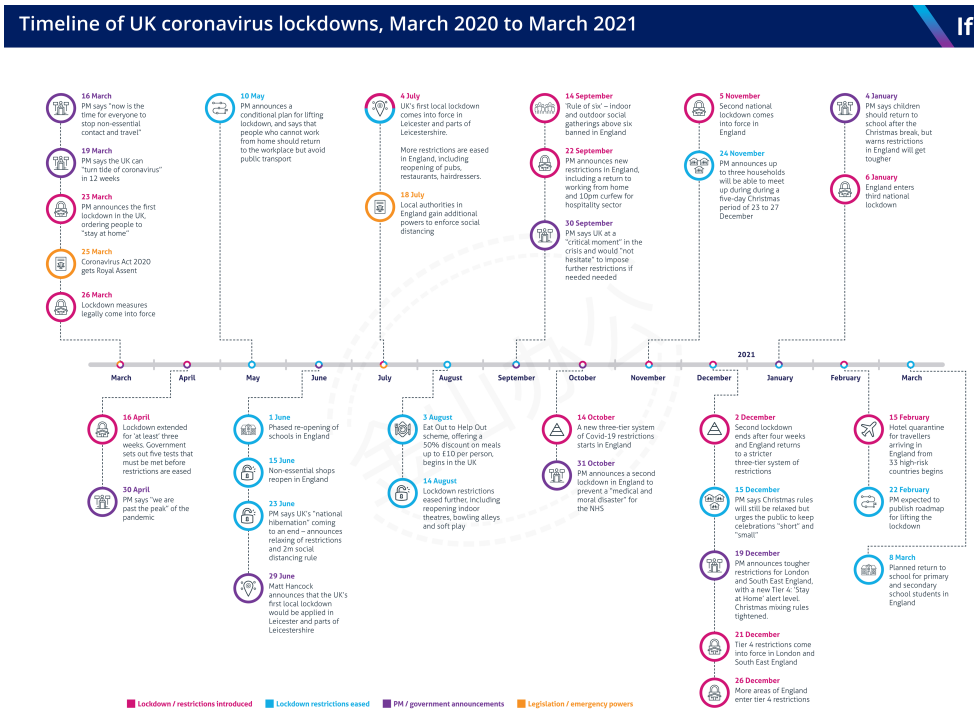

Figure 1.1: The timeline of NPI policies in the UK.

This figure shows that the NPI in the UK from March 2020 to March 2021 consists of 4 aspects (marked by four colors). This thesis does not consider the effect of the PM announcement here, because this is difficult to consider and the subsequent policies will also reflect the role of the PM announcement. It can be primarily deduced here that the lockdown in the UK will gradually ease from May to August.

The interventions can be of different levels and extents, according to the situation. For example, the UK government recommended closing schools in March 2020. Schools can open with alterations resulting in significant differences compared to non-Covid-19 operations. After September 2020, most international students take online classes, and some students take mixed classes online and offline. In May 2020, the UK government required closing only some levels or categories.

The closing policies for primary schools and high schools are canceled. In some other countries, when the infection rate is high, the government can even require closing all levels of schools in a certain region. For another example, in the international travel controls, the levels can be (from mild to severe) screening arrivals, quarantine arrivals from some or all regions, ban arrivals from some regions, and ban on all regions or border closer. Moreover, the lock-down effect also spread to other countries and regions. For example, if a city is closed, people in other cities will feel that the entire country is also very dangerous and therefore behave more cautiously (e.g. stay at home, keep social distance, etc.).

The main purpose of these interventions is to calculate sampling effort. Based on the observed cases, this thesis aims at simulating the real number of infections. To do this, it is required to use R_t and observed data such as fatalities data to simulate the real number of cases, and thus detect the sampling effort. The daily sampling effort can be simply achieved by dividing the number of new cases announced every day by the number of simulated cases. This thesis will estimate effects of NPIs on Covid-19 in the UK. To be specific, the effect of major interventions will be studied across the UK from period from the start of the Covid-19 epidemics in 31st January 2020 until 1st October 2020, when the majority of universities reopened with an upturn of the epidemics. Before 1 October 2020, The virus has not mutated, especially the delta strain has not yet become a pandemic in the UK (Lopez Bernal et al., 2021), so under the same medical condition and NPI policy, it is reasonable to assume that SARS-CoV-2 during this period is caused by the same strain, leading to essentially invariant fatality rate, hospitalization rate, serial interval, etc. Moreover, the UK government began popularizing vaccination at December 8th, 2020. Since vaccination has a very important impact on reducing R_t, the data after December is not considered. Also, 1st October is also the very beginning of the second wave of the epidemic in the UK, which can also be use to verify the trend of the simulated cases. On the other hand, this thesis assumes that during this period the cumulative number of people infected is not very large. In other words, there are not many people with antibodies, and the existence antibodies in a small proportion of people will not result in a great impact on R_t. It is known that a person can be infected twice with the Covid-19, but the probability of a second infection will be much less than the first time because of the appearance of antibodies in his/her body. However, in the time span of this research, the concentration of antibodies in the population was not high, so the adverse effects of antibodies in the time span of the rearch are ignored.

The data is downloaded from Office for National Statistics (ONS), and combined with the previous NPI data. It evaluates the trend of R_t the daily increasing

cases, the cumulative number of detected cases, and the daily samping effects. The data is used to study both the individual and shared effects on the time-varing reproduction number R_t of the effect of public isolation policies including lock down, self isolation, public events forbidden, social distance, and reopens. The goal is to simulate the actual number of Covid-19 daily infections in the UK in order to assess the daily sampling effect. These simulated results have a potential to verify if the interventions have so far been successful at reducing R_t to values below 1.

To calculate the sampling effect, daily Rt should firstly be simulated. However, it is quite challenging to estimate the R_t of Covid-19 for two reasons. Firstly, the high proportion of undetected infections in the health system (Jombart et al., 2020, Li et al., 2020, Zhao et al., 2020). Given the high incidence of non-specific and mild symptoms, the COVID-19 pandemic may be ignored in the new location until the first severe case or fatality is reported. Also, regular changes in testing policies that cause the proportion of detected infections to vary from time to time and from region to region. From the very beginning of the Covid-19 epidemics, most countries and regions can only detect a small number of suspected cases, and retain the ability to test critically ill patients or high-risk groups. Secondly, there are many precise experiment conducted by many universities such as Imperial College London and Oxford University. However, they only focus on a small region, ying the Geographic bias when representing R_t for the whole UK.

Figure 1.2 shows the number of detected cases (green line), fatality (blue line), and people in mechanical ventilation beds (red) observed from Jan 2020 to August 2021. It can be observed that in the three waves of the epidemic in the UK approximately on April 2020, January 2021, July 2021 respectively, the fatality data and the number of people in ventilation beds are similar, but the detected cases in the first round were significantly fewer. One of the major problem seen from this figure is that because there are mutated strains such as alpha, beta, gamma, and delta in the second wave (the part before 8th Dec), the fatality rate and people in ventilation beds can not be much lower than that in the first round (as is shown in Figure 1.2). Therefore, there is reason to believe that the reported data of detected infections is often biased downwards. Also, he number of deaths and the number of people in the ventilation bed may not be accurate. For example, the number of fatality may be overestimated because many people are already occupied in the hospital and there is a shortage of medical resources. Thus the fatality may be overestimated. On the other hand, The number of people in the ventilation bed may be underestimated, because the ventilation machine is limited, and many people cannot use it even until fatality. This thesis is going to simulate the number of cases in the first round and correspondingly resort to alternative ways to estimate R_t and thus to estimate the

sampling effect.

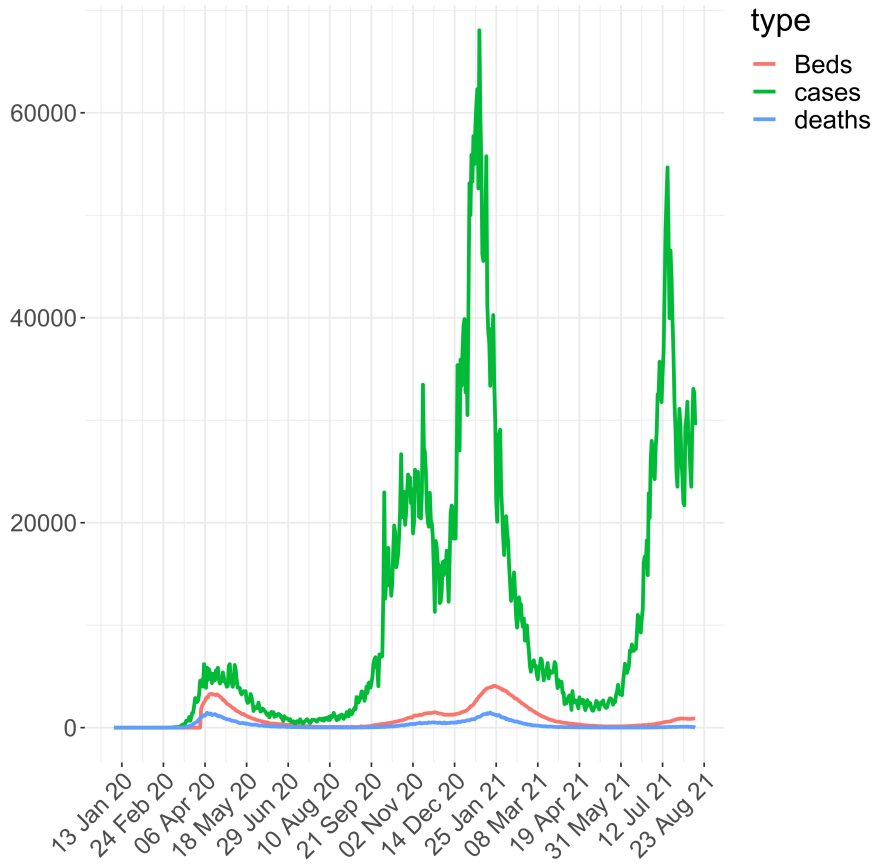

Figure 1.2: Number of detected cases, fatality, and people in mechanical ventilation beds.

The number of detected cases (green line), fatality (blue line), and people in mechanical ventilation beds (red) observed from Jan 2020 to August 2021. In the first two waves of the epidemic in the UK, the number of fatality and people in ventilation beds are similar, but the detected cases in the first wave were significantly fewer.

One of the alternative methods is to estimate the course of an epidemic is to calculate backwards from the number of fatality observed to the number of infections (Flaxman et al., 2020). I will use a Bayesian regression model to link the infection cycle to the observed fatality and infer the total infected population (attack rate) and R_t. The key assumption here is that the distribution of the day from infection to fatality remains constant within a certain time range. Compared to the reported case data, reported fatality is likely to be far more reliable. Nonetheless, there are several limitations of the statistics of fatality, e.g. at the start of the Covid-19 epidemics, some cases may be attributed to other diseases. In this thesis, I separately assume that the fatality case confirmed by the UK government is accurate or inaccurate.

Another alternative method is to estimate the reproducing number through the

number of people in mechanical ventilation beds. The key assumption are that 1) the mechanical ventilation beds data is assumed to be inaccurate. It is affected by the limitations of medical resources when the epidemic situation is severe. 2) The probability that an infected patient in mechanical ventilated bed is constant. 3) The day distribution of a person from infection to being in the b~~s fixed, comforting a certain distribution~~. The last two assumptions correspond to the settings of the previous researches. In this model, a Bayesian regression model is also used to infer R_t. The model with the best simulation result is shown in Section 5, and the others will be presented in the appendix.

The model relies on fixed estimates of some epidemiological parameters, such as the onset-to-death distribution, the infection fatality rate, and the generation distribution. In the model, I assume that only non-pharmaceutical interventions would impact the reproduction number (R_t). In practice, a continuous variable representing the degree of lock-down is used. On the other hand, not all people will obey the lockdown rules, they want to protest on the street at their own risk, which increases the number of reproduction R_t. Thus, a protest index of citizens against the lockdown should be added to the model, and the simplest way to measure it is the search index in Google. Furthermore, the serial interval of Covid-19 and the number of days of seed infection are assumed to be constant. This assumption originates from the built-in property of covid-19 and will not be affected by the NPIs. It is also implicitly assumed in the model that changes in R_t are an immediate response to interventions rather than gradual changes in behaviour and that individual interventions have a similar effect in different regions of the UK. The output of the model is the estimated value of R_t over time. Besides, the 95%, 60%, and 30% confidence intervals will also be reported.

The main experimental results are briefly shown as follows. When modeling through people in mechanical ventilation beds, the peak of simulated daily increasing cases is nearly 70K (46k, 107k) on March 20th. By dividing the number of detected cases by the simulation result, the daily sampling effect rises with fluctuations over time and reaches 23.77% with CI $(8.81\%, 66.64\%)$ on the last day of time span of this thesis, while remains under 100% through out the entire time span. This indicates that the UK government still has to work hard to enhance the sampling effort. Matching the timeline of government policies with the estimated transmissions in time-series format, the effect of the non-pharmaceutical interventions (NPI) can be interpreted. The results show that major non-pharmaceutical interventions, especially lockdowns, have had a large effect on reducing transmission.

Chapter 2

Statistical Background

2.1 Bayesian Statistics

Bayesian statistics was firstly introduced by Thomas Bayes, where a specific case of Bayes' theorem was published in a paper in 1763. Throughout the 20th century, Bayesian methods were underestimated by many statisticians due to philosophical and practical considerations. Many Bayesian methods heavily rely on computational power to implement, while the most widely used Bayesian methods back in that time were based on the frequentist interpretation. However, as the rapid development of computers, the computational power enhances greatly. With new algorithms like Markov chain Monte Carlo, Bayesian methods prevails widely over all walks of statistics in the 21st century (Gelman et al., 1995).

The core of Bayesian statistical methods is Bayes' theorem. In Bayesian statistics, the probabilities are computed and updated each time a new data point is fed into the model. Bayes' theorem introduced by Thomas Bayes is given by

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}, \tag{2.1}$$

where θ is the vector of parameters, X is the data fitted to the model, and $p(X) \neq 0$. Here, $p(X|\theta)$ is called the likelihood, $p(\theta)$ is the prior probability, $p(X)$ is the evidence or the so-called normalizing constant, and $p(\theta|X)$ is the posterior probability.

This theorem reveals the relationship between the posterior probability of the parameters and the likelihood of data under the parameters, with the prior knowledge about the distribution of parameters. In Bayesian inference, Bayes' theorem is capable of estimating the parameters of both probability distributions and statistical models.

It can be observed that to implement Bayesian inference, one need to synthesize two sources of information about unknown parameters of interest. First, given specific observation data, the likelihood function defines the possible values of the

model parameter. The second is the distribution of prior beliefs, which is used to express the confidence of model parameters based on past experience. Then the product of the two is scaled and integrated into one within a reasonable range of parameter values. The result is the posterior ~~confidence~~ distribution of a given ~~data~~ parameter, which represents the understanding of the parameter based on the prior information and the input data.

Generally speaking, Bayesian statistics can solve more complex problems and provide more reasonable and intuitive inferences. In addition, through Bayesian methods, direct probabilistic statements about the parameters of interest can be made.

2.2 Bayesian Linear Regression

In statistics, Bayesian linear regression is a linear regression method in which statistical analysis is performed in the context of Bayesian inference. When the error of the regression model follows a normal distribution, and a specific form of the prior distribution is assumed, the posterior probability distribution of the model parameters ca~~n get an explicit result.~~

The Bayesian linear regression model starts with the same model as the frequentist linear regression, i.e. given a predictor vector x_i, the response variable y_i is

$$y_i = \alpha + \beta x_i + \varepsilon_i, \text{ for } i = 1, \ldots, n, \tag{2.2}$$

where errors ε_i are assumed to be independent and identically drawn from the a normal distribution with zero mean and constant variance σ^2, denoted as $\varepsilon_i \sim N(0, \sigma^2)$. Note that this assumption is the same as that in the frequentist linear regression for testing and constructing confidence intervals for the parameters α and β.

Under this assumption of ε_i, the random variable of each response Y_i conditioning on the observed data x_i and the parameters α, β, and σ^2, turns out to be normally distributed, i.e.

$$Y_i \,|\, x_i, \alpha, \beta, \sigma^2 \sim N(\alpha + \beta_i x_i, \sigma^2), \text{ for } i = 1, \ldots, n. \tag{2.3}$$

Thus, the likelihood of Y_i is given by

$$p(y_i \,|\, x_i, \alpha, \beta . \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2} \right). \tag{2.4}$$

By independence, the likelihood of $Y_1, \ldots, Y_n$ is given by the the product of each

$p(y_i \mid x_i, \alpha, \beta, \sigma^2)$.

When considering the reference prior, the posterior distributions of α, β, and σ^2 is analogue to the frequentist results. Assume that the joint prior distribtuion of α, β, and σ^2 to be proportional to the inverse of σ^2, i.e.

$$p(\alpha, \beta, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{2.5}$$

Using the hierachical model framework, this is equivalent to assuming

$$p(\alpha, \beta \mid \sigma^2) \propto 1 \text{ and } p(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{2.6}$$

Then the marginal posterior distribution of β is the Student's t-distribution

$$\beta \mid y_1, \ldots, y_n \sim t(n-2, \hat{\beta}, \frac{\hat{\sigma}^2}{S_{xx}}) = t(n-2, \hat{\beta}, (se_\beta)^2), \tag{2.7}$$

with degrees of freedon $n-2$ centered at $\hat{\beta}$, with the scale parameter $\frac{\hat{\sigma}^2}{S_{xx}}$, which is the square of the standard error of $\hat{\beta}$ under the frequentist OLS model.

Similarly, α also follows the Student's t-distribution

$$\alpha \mid y_1, \ldots, y_n \sim t\left(n-2, \hat{\alpha}, \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) = t(n-2, \hat{\alpha}, (se_\alpha)^2). \tag{2.8}$$

The following results will be used to calculate the confidence interval of the response variable Y.

The mean of the response variable Y, μ_Y, at a point x_i is

$$\mu_Y \mid x_i = E[Y|x_i] = \alpha + \beta x_i. \tag{2.9}$$

Under the reference prior, μ_Y has a posterior distribution

$$S^2_{Y|x_i} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right). \tag{2.10}$$

Then any new prediction y_{n+1} at a point x_{n+1} also follows the Student's t-distribution

$$y_{n+1} \mid data, x_{n+1} \sim t(n-2, \hat{\alpha} + \hat{\beta} x_{n+1}, S^2_{Y|X_{n+1}}), \tag{2.11}$$

where

$$S^2_{Y|X_{n+1}} = \hat{\sigma}^2 + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{S_{xx}}\right) = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{S_{xx}}\right). \quad (2.12)$$

2.3 Markov chain Monte Carlo (MCMC)

In statistics, the Markov Chain Monte Carlo (MCMC) method includes a class of algorithms for sampling from probability distributions. By constructing a Markov chain with the desired distribution as its equilibrium distribution, a sample of the desired distribution can be obtained by recording the state in the chain. It is primarily used to calculate the numerical approximation of multi-dimensional integrals, such as Bayesian statistics, computational physics, computational biology and computational linguistics. In Bayesian statistics, the latest development of the MCMC method makes it possible to calculate large-scale hierarchical models that need to integrate hundreds to thousands of unknown parameters.

The realization of Monte Carlo simulation can be summarized into the following three steps:

- Convert the problem to be solved into a probabilistic process.
- Sampling from a known distribution.
- Calculate various statistics through samples, and such statistics are the solutions to the desired problem.

For common probability distributions, whether it is a discrete distribution or a continuous distribution, their samples can be obtained with the aid of a uniform distribution in $(0,1)$, denoted as $Unif(0,1)$. For example, to sample from a two-dimensional normal distribution (Z_1, Z_2). One can independently sample (X_1, X_2) from $Unif(0,1)$ and then obtain (Z_1, Z_2) by

$$Z_1 = \sqrt{-2\ln X_1}\cos(2\pi X_2)$$
$$Z_2 = \sqrt{-2\ln X_1}\sin(2\pi X_2).$$

Other common continuous distributions, such as t-distribution, F-distribution, Beta distribution, Gamma distribution, etc., can be transformed from sample samples obtained by $Unif(0,1)$ in a similar way.

For a distribution that is not a common probability distribution, a feasible method is to use accept-reject sampling to get a sample of the distribution. When $p(x)$ is too complicated to sample directly in the program, one can set a distribution that can be sampled by the program $q(x)$ such as Gaussian distribution, and then

reject some samples according to certain methods to get close to $p(x)$. Here $q(x)$ is called proposal distribution. To be specific, find a proposal distribution $q(x)$ and a constant k such that $p(x) \leq kq(x)$ for all x. First sample a value z_0 from $q(x)$, and then sample a value u from the uniform distribution $Unif(0, kq(z_0))$. If $u \leq p(z_0)$, accept z_0, and reject it otherwise. Repeat the above process, a series of acceptance and rejection decisions are used to achieve the purpose of simulating the probability distribution of $p(x)$ with $q(x)$.

Chapter 3

Epidemic Model

Epidemic modelling, a task focusing on learning the transmission pattern of epidemic disease, usually utilizes time-series data about populations, infected patients and the final observations to describe how the disease transmits and how powerful the transmission can be. Literally, it is believed that the result of a epidemic model is highly related to the training data. Consequently, a well trained epidemic model theoretically has no access to perfect interpretations to the underlying epidemic pattern, but can obtain the key components that influence the disease propagation.

As for the history of epidemic modelling, it is generally believed that the first series of mathematical epidemic models can date back to 1920s, when Kermack and McKendrick proposed their works in Kermack and McKendrick (1927). Basically, they assume people in consideration can only belong to 3 categories: susceptibles (S), infecteds (I) and removed (R), which leads to its alias "SIR models". Afterwards, various epidemic models have been proposed based on SIR models by changing the assumptions or the conditions, contributing to better interpretation to real-world diseases.

In this paper, epidemic modelling provides the basic analyzing structure, indicating that how the data will be used to infer the parameters or predict the transmission. In order to better understanding the model, classic epidemic models are introduced in this section.

3.1 SIR models

SIR models are one of the earliest epidemic models. Its basic assumption lies on that individuals from an invariant community are initially equally susceptible, and one will never re-catch the disease after recovering from the infection. As mentioned before, three distinct classes are included in the system: Susceptible individuals (S); Infected individuals (I); and individuals who recover from the infections and been Removed from the infecting system(R). Schematically, the only way an individual

can process is “healthy-infected-removed”.

Mathematically, let S_t, I_t, R_t denote the variables defined above at time t, respectively (Mentioned that R_t means removed individuals in this section). Assume that

- $S_t + I_t + R_t \equiv N$;
- two persons encounter at probability p_1 per unit time;
- each time a healthy person encounters with an infectious one, the healthy one catches the disease at probability p_2 without incubation period;

According to the function, the tuple of variables $(S_t,\ I_t)$ is a Markov Chain given the initialization S and I, where the transition probability can be demonstrated as

$$P_t\left((S-1, I+1)|(S,I)\right) := p_1 p_2 SI$$

With a dynamical system defined as:

$$\begin{cases} \frac{dS_t}{dt} = -p_1 p_2 S_t I_t \\ \frac{dI_t}{dt} = p_1 p_2 S_t I_t - \rho I_t. \end{cases}$$

where ρ is a probability related to the “infected-removed” process.

The key purpose of establishing the dynamic system for epidemic propagation is to find the steady state or the end state of the disease, asking questions like “Will the disease disappear naturally?”; “How fast will be the disappearance?”, etc. Subsequently, suppose the initialization of the system is non-negative healthy and infected people, together with removed people equals to 0.

To answer the questions, it is worthy looking back upon the dynamic system which defines the change of the infected people. By combining the same variable, it can be written as

$$displaystyle\frac{dI_t}{dt} = I_t(p_1 p_2 S_t -)$$

here it is obvious that the value of $\frac{p_1 p_2 S_t}{\rho}$ totally determines the changing trend that whether the number of infected people increases or decreases at time t. As a result, the important parameter $R_{reproduction} = \frac{p_1 p_2 S_0}{\rho}$, usually denoted as *basic reproduction number*, can be directly used to distinguish if a disease will develop

into an epidemic:

$$R_{reproduction} = 1 \text{ need further study}$$
$$R_{reproduction} < 1 \text{ no epidemic}$$
$$R_{reproduction} = 1 \text{ possible epidemic}$$

3.2 SIS models

When disease that the patients will gain completely immunity after recovering can be solved by SIR models, the majority of the diseases can infect an individuals again and again. To study this kinds of diseases, researchers have proposed SIS models with re-infection setting. Generally, SIS models are similar to SIR models, but assume that recovered individuals can get infected again. As a result, SIS model is a binary system where individuals changing between the state of "healthy" and "infected". Similarly, the dynamic system of the SIS model can be written as

$$\frac{dS_t}{dt} = -\beta S_t I_t + \alpha I_t,$$
$$\frac{dI_t}{dt} = \beta S_t I_t - \alpha I_t.$$

where

- the parameter β is highly related to the probability of being infected when healthy people encounter with infected ones.
- the parameter α is highly related to the probability of the recovery by infected people.

To study the long-term situation of SIS epidemic model, the changes of infected people is again in considerations:

$$\frac{dI_t}{dt} = (\beta N - \alpha) I_t - \beta I_t^2.$$

Different from SIR model, the dynamic equation of the infected people is a 2-order equation with solution $I_t = 0$ and $I_t = N - \alpha/\beta$. Since $I_t = 0$ represents a special situaition that the disease is gone (no infected people) in the system, the basic reproduction number of SIS model is usually defined as

$$R_{reproduction} := \frac{\beta N}{\alpha},$$

with determination that

$$R_{reproduction} < 1 \text{ no epidemic}$$
$$R_{reproduction} > 1 \text{ stable epidemic with} I = N - \alpha/\beta$$

3.3 SIR models with vital dynamics

In SIR models and SIS models, the population is assumed to be a constant. Nevertheless, when study the pattern of a disease in certain real-world area, the changes of the population must be taken into consideration. The changes of population can be caused by various reasons, for example, immigration, new births and deaths, etc. Consequently, to better simulate the real-world epidemics, SIR models with vital dynamics are proposed.

Mathematically, denoting μ and Γ as death and birth rates, respectively. The dynamic equations can be written as:

$$\begin{aligned}
\frac{dS_t}{dt} &= \Gamma - \mu S_t - \frac{\beta I_t S_t}{N} \\
\frac{dI_t}{dt} &= \frac{\beta I_t S_t}{N} - \gamma I_t - \mu I_t \\
\frac{dR_t}{dt} &= \gamma I_t - \mu R_t
\end{aligned}$$

where the equilibrium lies at

$$(S_t, I_t, R_t) = \left(\frac{\Gamma}{\mu}, 0, 0\right)$$

with basic reproduction number equals to $R_0 = \frac{\beta}{\mu+\gamma}$. and it can be shown that:

- $R_0 < 1 \Rightarrow (S_t, I_t, R_t) \to \left(\frac{\Gamma}{\mu}, 0, 0\right)$ is stable,
- $R_0 > 1 \Rightarrow (S_t, I_t, R_t) \to \left(\frac{\gamma+\mu}{\beta}, \frac{\mu}{\beta}(R_0 - 1), \frac{\gamma}{\beta}(R_0 - 1)\right)$, representing that the disease is not totally eradicated and remains in the population.

Chapter 4

Methodology

Aiming at estimating the number of infections,the model is established based on epidemic modelling, bayesian regression and MCMC simulation. Guided by Flaxman et al. (2020), the epidemic model provide mathematical relationship that establishes the mathematical relationship by modelling the the observed data from the infections. Nevertheless, the relationship in epidemic model is an inverse version of the aim. Compared with the basic models like SIR, SIS or SIR with vital dynamics, the epidemic model used in this paper does not focus on solving the dynamic equations and calculate the stable state of the system. Instead, it attaches more importance to mode realistic problems: how does the disease system work. For instance, how the healthy individuals are infected? And what is the differences between the infections and the recorded data, etc. Besides, the reproduction number is introduced in the epidemic model for simplifying the process.

On the other hand, in order to inferring infections from the observed data, the bayesian regression is introduced, considering all factors as variables and parameters with underlying unknown distributions. Last but not least, the further details about the model are demonstrated, including the prior distributions about the parameters and the median and confidence interval estimation through MCMC method.

Fig. 4.1 summarizes the overall procedures. In the beginning, the probability of a healthy person to be infected is determined by the reproduction number, and the time when the person is infected is affected by an empirical distribution of the series interval. Besides, the calculation of the reproduction number R_t is based on the results of REACT group and other control variables including the state of non-pharmaceutical indexes and search indexes of "protest" on google. It is worthy highlighting that the REACT (Real-time Assessment of Community Transmission) is an auxiliary dataset comprised of a series researches studying the the process of England's COVID-19 epidemic with home-testing records. It is well authorized and widely believed to be the most accurate regional records, yet the size of the dataset

is rather small. Consequently, employing REACT data can greatly help testifying and amending the estimated distribution of the reproduction numbers. Meanwhile, when the infection time interval keeps the same in a single day sampling from the empirical distribution, it is worthy highlighting that the number of infected individuals will also boost the transmission. After being infected, the state of a person will have many situations, including unrecorded, deaths, admit to ventilation beds, in hospital and so on. It is further assumed that a constant distribution can be employed to describe this process.

The structure of this section is as follows. the notations are firstly illustrated in Table 4.1. In Sec 4.1, the epidemic structure between the observed data and the infections is introduced. Then, the bayesian regression is utilized to explore the posterior distributions of R_t and other important parameters in Sec 4.1.3. In Sec 4.2, more details and explanations about the model are demonstrated. Finally, in Sec 4.3, the construction of the bayesian regression is given as a guidance for parameter estimating.

Before mathematically illustrating the model, the notations are firstly shown in this section. Then in this section, the capital letters without subscripts are restricted to vector form, i.e., $X = (X_1, \ldots, X_t)$, and an interval of time can be expressed with a colon ":", i.e. $0 : t = 0, 1, \ldots, t$ and $t : 0 = t, t-1 \ldots, 0$.

Table 4.1: Notations used in methodology

Notations	Meanings
t	Time stamp in days
y_t	expectation of response variable in day t
Y_t	response variables in day t (deaths or beds)
i_t	new infections in day t
R_t	reproduction number in day t
$p(\cdot)$	statistical distributions
π	distribution of time between infection to observation
ϕ, α, τ	parameters

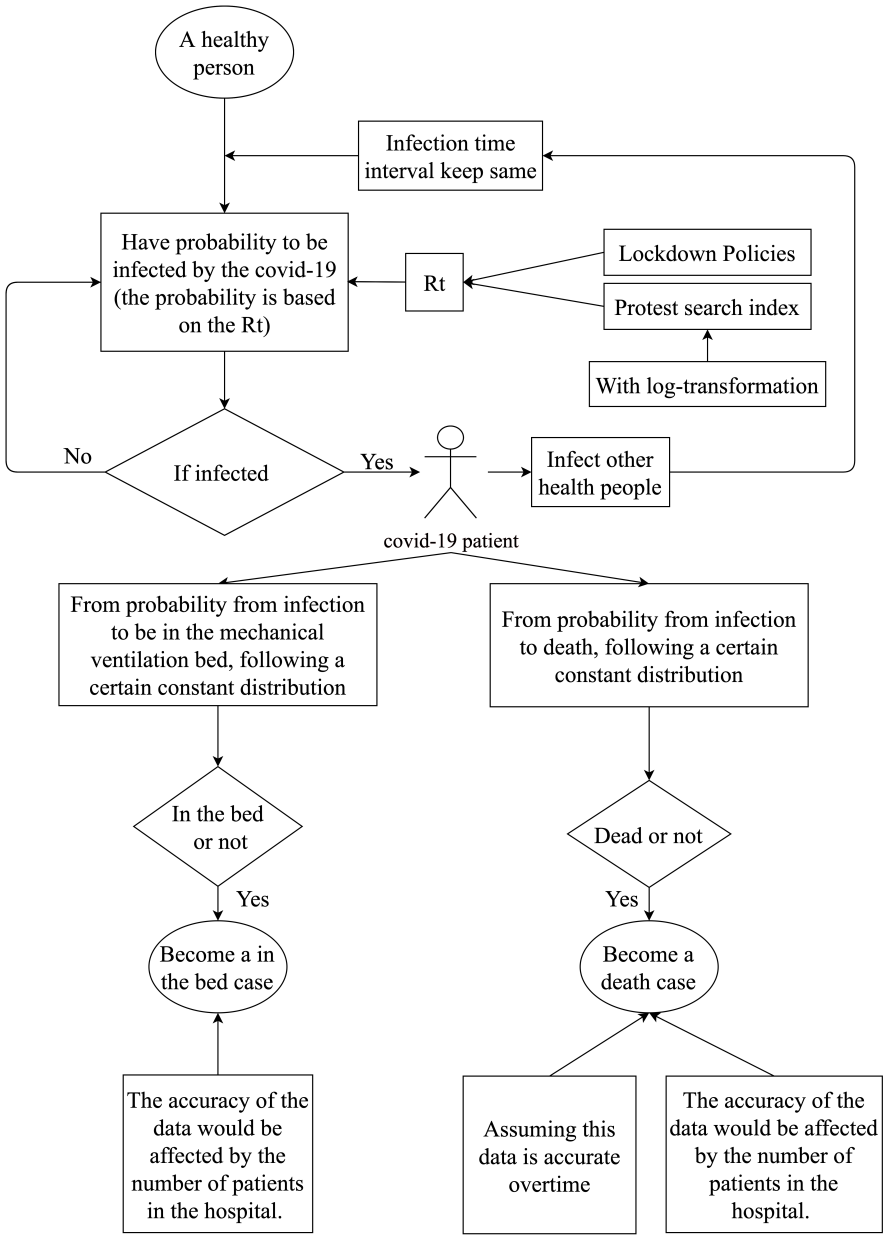

Figure 4.1: The flowchart of the model.

In this flowchart, three approaches from infection to observation are demonstrated: becoming a bed case; becoming a death case with assumption that the data is accurate; becoming a death case with assumption that the data is inaccurate. The evaluation of the three models will be illustrated in the following parts, where an example is detailed explained and the other two are depicted in the appendix.

4.1 Basic epidemic model

4.1.1 Observed-Infected Model

According to the variants of SIR model and the structure of bayesian analysis, the relationship between the possible "observed" and "infected" variables can be derived. Let $Y = (Y_1, \ldots, Y_n)$ denote the observed non-negative vector of data in n days. From a perspective of statistical modelling, infections are used in the past few days $i_s, s < t$ to model Y_t. Nevertheless, since Y_t is regarded as an observation of a distribution, the relationship between infection and the observed data must be bridged on the expectations. As a result, the model can be expressed as:

$$Y_t \sim p(y_t, \phi) \tag{4.1}$$

$$y_t = \alpha_t \sum_{s<t} i_s \pi_{t-s} \tag{4.2}$$

where $y_t = E(Y_t)$, $p(y_t, \phi)$ is the underlying distribution to generate Y_t, ϕ is the structural parameter of the distribution, α_t is the proportion of events at time $t$that are recorded in the data, and π denotes the time distribution from infection to observation, indicating the hysteresis.

4.1.2 Self-development of infections

According to the epidemic modelling Bellman and Harris (1952, 1948), Cauchemez et al. (2008), Cori et al. (2013), Fraser (2007), Nouvellet et al. (2018), infections i_t develop through a cumulative increasing pattern: the past infections will continuously contribute to the increasements of new infections (Bhatt et al., 2020). Then, the new infections i_t can be modeled through a renewal equation that controlled by both the reproduction number R_t and a degradation function g. Formally,

$$i_t = R_t \sum_{s<t} i_s g_{t-s} \tag{4.3}$$

Recall that the observed data is $Y = (Y_1, \ldots, Y_n)$. In other words, no access is available to the previous information, including the initialization of the epidemic. In order to model the reproduction of the infections, the unknown information about the previous infections is as parameters $i_{v:0}$. Subsequently, the equation implies that infections $i_t, t > 0$ determined by given the reproduction number R and the initialization infections $i_{v:0}$.

4.1.3 Bayesian inference for parameters

From the perspective of bayesian statistics, let all parameters are assigned priors:

$$i_{v:0}, R, \phi, \alpha \sim p(\cdot)$$

where $i_{v:0}$ can be used to initialize the renewal equation of the infections i_t, $R = (R_1, \ldots, R_n)$ denoting the reproduction numbers, $\alpha = (\alpha_1, \ldots, \alpha_n)$ denoting the vector of parameters.

Then, according to bayesian statistics, the posterior distribution of the parameters can be expressed as

$$p(i_{v:0}, R, \phi, \alpha | Y) \propto p(i_{v:0}) p(R) p(\phi) p(\alpha) \prod_t p(Y_t | y_t, \phi)$$

4.2 Details of the models

In this section, the exampl"regarding the number of people in Ventilation Beds as the observed data" is used to show how the epidemic models, the bayesian regression are combined to predict the infections.

4.2.1 Settings of observed-infection model

As mentioned in Sec 4.1.3, the bayesian regression asks prior distributions of the variables and parameters to give the inferences. For the number of people in Ventilation Beds, the data is assumed to follow a negative binomial distribution:

$$Y_t \sim \texttt{Negative Binomial}\left(y_t, y_t + \frac{y_t^2}{\phi}\right),$$

where $\phi \sim N^+(0, 0.5) := |N(0, 0.5)|$ is the scale parameter following a positive half normal distribution.

Moreover, the expected number y_t is mechanistically linked with the infections through the multiplier α and the infection-to-observation-time distribuiton π by 4.1. The multiplier α is actually a parameter of COVID-19 called infection-fatality-ratio (*ifr*) (Verity et al., 2020), which was derived in Ferguson et al. (2020) with assuming homogeneous attack rates across age-groups. In this model, the *ifr* is assumed to be a constant overtime $\alpha = 1$ (percentage), an additional noise is attached in this model

$$\alpha^* = \alpha \cdot N(0, 0.5)$$

As for π, according to the historical researches Ferguson et al. (2020), Verity

et al. (2020), it is assumed that π is the distribution of times between an individual gets infected and shows the syndrome. In this model, following Lauer et al. (2020), the infection-observed(Beds) distribution is defined as a log normal distribution with parameters as

$$pi \sim \text{LogNormal}(1.921, 0.428)$$

However, instead of using the distributional format of π, the calculations of y_t employ the discrete format with time t. In this model, define $\pi_t = \int_{t-0.5}^{t+0.5} \pi(\mu)d\mu$, for $t = 2,3,\ldots$ and $\pi_1 = \int_0^{1.5} \pi(\mu)d\mu$, where $\pi(\mu)$ is the density of π.

4.2.2 Settings of infection-renewal model

In the self-exciting renewal process of infections, function g denotes a generation distribution modelling the interval between the time a person get infection and the time the person successively infects another one. Nevertheless, the ground truth of generation distribution is never accessible. Consequently, an empirical approximation of the generation distribution is used from result in Bi et al. (2020), as

$$g \sim \texttt{Gamma}(6.5, 0.62)$$

Similar to π, the generation distribution g is also discretized as $g_t = \int_{t-0.5}^{t+0.5} g(\mu)d\mu$, for $t = 2,3,\ldots$ and $g_1 = \int_0^{1.5} g(\mu)d\mu$.

Meanwhile, to simplify the model, the functional form for the time-varying reproduction number is replaced by a piecewise constant function established from the start value R_0 and determined by the known major non-pharmaceutical interventions in various times. On the one hand, the model sets several starts points R_0^t using the data collected from Real-time Assessment of Community Transmission (REACT) group. On the other hand, this model specifies the lockdown, the search index of "protest" on Google, and the closure of schools and universities as the values dominating the changes of the reproduction nunmber. Moreover, by setting $k \in \{1,2,3\}$ the indexes of interventions and $I_{k,t}$ the levels of interventions at time t, it is possible to establish the target linear regression. To be specific, with the assumption that each of the intervention functions multiplicatively, the model defines

$$R_t = R_0 exp\left(-\sum_{k=1}^{3} \beta_k I_{k,t} - \lambda I_t^*\right)$$

where I_t^* is the specific value of the last intervention during the interval from the start of the epidemic to time t. The model employs exponential form to guarantee the

positivity of the reproduction number. In addition, β_k are parameters or coefficients of this generalized liner model and can be estimated from the data, where the prior distributions are chosen to be

$$\beta_k \sim \texttt{Gamma}(2, \frac{1}{3}) - \frac{log(1.05)}{6}$$

The reason that all effects of different interventions is all interventions are believed to be equally useful for influencing the reproduction number R_t. Besides, the term $\frac{(1.05)}{6}$ is set to shift the gamma distribution to negative value, making it possible for both increasements and decreasements of the reproduction number.

Finally, the initialization of infections and their prior distribution are set with the guidance in Flaxman et al. (2020). Let $v = -5$, and

$$\begin{aligned} i_k &\sim exp(\tau^{-1}) \quad k \in \{-4,-3,-2,-1,0\} \\ \tau &\sim exp(0.03) \end{aligned} \tag{4.4}$$

4.3 Construction of the Bayesian Regression

In this section, a list of linear-bayesian-based models are proposed with priors distributions defined in former sections. Specifically, those models employ Generalized Linear Model (GLM) technique to take the place of simple linear relationships between independent and dependent variables, making the model more flexible but more difficult to compute.

The general structure of the GLM-based inference model can be demonstrated as a transformed linear predictor consisting of fixed effects, random effects and autocorrelation terms with being a rational chosen link function. Mathematically, the model requires

$$Y = h^{-1}(\eta),$$

where h is the pre-defined link function with various options, i.e. the scaled logit link function

$$g^{-1}(x) = \frac{M}{1+e^{-x}}$$

where M is the upper-bound designed to restrict the dependent variable Y and η is the linear model:

$$\eta = \beta_0 + X\beta + Q\gamma$$

where, β_0 is the fixed effects; X is the data matrix with rows representing daily records of the information and columns representing different variables that determine the changes in Y; Q is the time-series-based binary matrix specifying the autocorrelation terms at time t.

For reproduction numbers, three fixed effects are included, representing the lockdown, the search index of "protest" on Google, and the closure of schools and universities. As for the link function, the scaled logit link function is utilized with the maximum limitation $M = 5.7$ for reproduction numbers:

$$R = h_R^{-1}(\beta_{R_0} + I_{lockdown}\beta_{R|lockdown} + I_{protest}\beta_{R|protest} + I_{school}\beta_{R|school})$$

$$h_R^{-1}(x) = \frac{5.7}{1+e^{-x}}$$

Chapter 5

Experiments

As shown in chapter 4, the model has been written by the team of Imperial College London as the R package 'epidemia' (Flaxman et al., 2020). This thesis calls the epidemia package and modifies the key parameters in it, so as to match the model settings in Chapter 4.

The data used in this section is published weekly by the ONS, and there is a lag in reporting of at least 11 days because the data are based on fatality registrations. The fatality data used in this thesis is the daily deaths with COVID-19 on the fatality certificate by date of death. Each hospital trust reports daily on the number of confirmed COVID-19 patients in hospital at 8am. The UK figure is the sum of the four nations' figures and can only be calculated when all nations' data are available. The data source of this thesis is the website `https://coronavirus.data.gov.uk/`.

5.1 Prior Reproduction Number

The lock-down of city and schools would significantly reduce the reproduction number of the coronavirus, while the pr[illegible] of the citizens would eliminate the efforts of lock-down. Hence, the shape of the prior reproduction number has been determined by the variation of these two conditions. To determine the number (location) of prior reproduction, the maximum reproduction number should be figured out at the early stage of epidemics(3.38, 95% confidence interval, 2.81 to 3.82) (Alimohamadi et al., 2020) and the experiment result of R_t conducted by the Imperical Colledge London (Riley et al., 2020a,b,c) and ONS (GOV.UK, 2021).

Recall that in figure 1.1, there are lots of lock-down and reopen policies for UK. An straightforward way to describe these policies is to use dummy variables. However, if all policies are add to the model as dummy variables, the Pareto k diagnostic value would come to the infinity, which means this model cannot fit well. Due to the multicollinearity problem, the variance of the coefficient of each NPI

policy on R_t will be very large in this model fitting, which leads to the unreliability of the final result. Moreover, it is inappropriate to simply add up all the dummies as a new variable or to use the PCA method to reduce the dimensionality of variables. To be specific, when adding up all dummies, all variables have inconsistent effects on the lockdown policy. At the same time, the effect of each policy will weaken over time The PCA method faces the same problem, even if a coefficient 1 or -1 is given. Therefore, this thesis combines all policy variables to a single continuous variable, describing the NPI policies by using manually subjective scoring. By this means, in the process of dimensionality reduction, the meaning of the NPI policy is retained to the greatest extent. Nonetheless, the disadvantages of this method are also very obvious. Since the scores are measured manually, this variable can be very subjective. Therefore, in the following simulations, it is necessary to verify whether this scoring principle complies with data in practice. The lock down variable and other variables to be used in the model are shown in Figure 5.1. It could be seen that the lockdown policy is strict when the infection is severe, and the lockdown eases when the infection decreases. Then lockdown policy and daily increasing cases turn out to be the cause of each other.

Table 5.1: Data description of REACT and ONS for reproduction number.

Group	Time span	Value	95% CI	Relationship with the prior	reference
REACT	May 1, 2020 - Jun 1, 2020	0.57	$(0.45, 0.72)$	within 60% CI of prior	Riley et al. (2020a)
	Jun 1, 2020 - Jul 1, 2020	0.89	$(0.86, 0.93)$	within 60% CI of prior	Riley et al. (2020b)
	Jul 24, 2020 - Aug 10, 2020	1.3	$(1.2, 1.4)$	within 30% CI of prior	Riley et al. (2021)
	Aug 20, 2020 - Sep 7, 2020	1.7	$(1.4, 2.0)$	within 30% CI of prior	
	Sep 18, 2020 - Sep 26, 2020	1.06	$(0.74, 1.46)$	within 30% CI of prior	Riley et al. (2020c)
ONS	March 30, 2020 - May 29, 2020		$(0.7, 0.9)$	within 95% CI of prior	GOV.UK (2021)
	May 30, 2020 - Jun 05, 2020		$(0.7, 0.9)$	within 60% CI of prior	
	Jun 06, 2020 - Jun 12, 2020		$(0.7, 0.9)$	within 60% CI of prior	
	Jun 13, 2020 - Jun 19, 2020		$(0.7, 0.9)$	within 30% CI of prior	
	Jun 20, 2020 - Jun 26, 2020		$(0.7, 0.9)$	within 60% CI of prior	
	Jun 27, 2020 - Jul 03, 2020		$(0.7, 0.9)$	within 60% CI of prior	
	Jul 04, 2020 - Jul 10, 2020		$(0.7, 0.9)$	within 60% CI of prior	
	Jul 11, 2020 - Jul 17, 2020		$(0.7, 0.9)$	within 60% CI of prior	
	Jul 18, 2020 - Jul 24, 2020		$(0.7, 0.9)$	within 30% CI of prior	
	Jul 25, 2020 - Jul 31, 2020		$(0.85, 0.9)$	within 60% CI of prior	
	Aug 01, 2020 - Aug 07, 2020		$(0.85, 1)$	within 60% CI of prior	
	Aug 08, 2020 - Aug 14, 2020		$(0.85, 1)$	within 60% CI of prior	
	Aug 15, 2020 - Aug 21, 2020		$(0.9, 1.1)$	within 60% CI of prior	
	Aug 22, 2020 - Aug 28, 2020		$(0.9, 1.1)$	within 60% CI of prior	
	Aug 29, 2020 - Sep 04, 2020		$(0.9, 1.1)$	within 60% CI of prior	
	Sep 05, 2020 - Sep 11, 2020		$(1, 1.2)$	within 60% CI of prior	
	Sep 12, 2020 - Sep 18, 2020		$(1.1, 1.4)$	within 60% CI of prior	
	Sep 19, 2020 - Sep 25, 2020		$(1.2, 1.5)$	within 60% CI of prior	
	Sep 26, 2020 - Oct 02, 2020		$(1.3, 1.6)$	within 60% CI of prior	

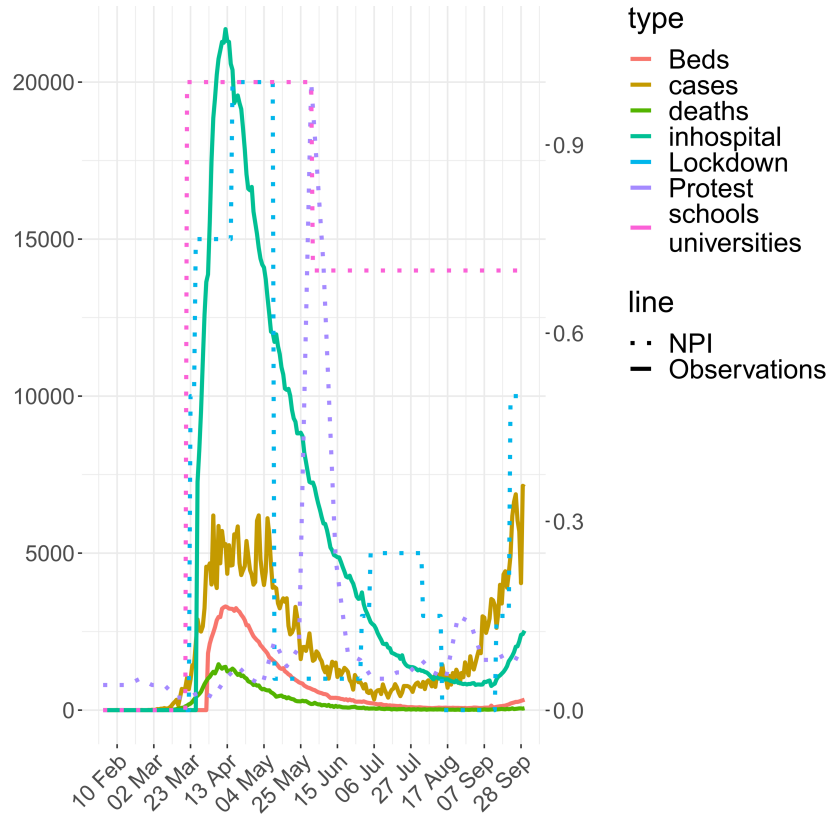

Figure 5.1: Cases, people in beds, fatality, people in hospital, and lockdown policy

This figure shows the raw data of observed cases, people in beds, fatality, and people in hospital. The dotted lines show the variables that describe the NPI policies, including lockdown, lockdown of schools and universities, and the protest index. Note that the values are shown on the left y-axis, and the index is shown on the right y-axis.

The extent of protesting the NPI policies is taken into consideration. The data used is the search volume of the protest in google, downloaded from google trend As can be observed from the raw data (Figure 5.1), the volume of the protest is usually low, but it is particularly high in June or July, so taking a logarithm transformation of the protest search index is considered here. A lockdown index for school and university is also considered. The environment of universities in the UK has very distinct characteristics from other places such as office buildings and neighborhoods. Universities are relatively closed environment with a very large flow of people, which facilitates the spread of the Covid-19 infection. Also, universities in the UK are highly internationalized, where international exchanges are more likely to cause the spread of viruses. Therefore, the lockdown of universities is separately considered in the model. All NPIs only provides a priori on the changing trend of R_t, but there is no a priori for the specific value. Therefore, the results of REACT and ONS should be resorted. According to the shape of the R_t prior simulated by

NPI, move the R_t prior up and down to make it consistent with the results of REACT and ONS. The REACT group has tested the monthly varied reproduction number. Each month, there are more than 150,000 people invited to take part. Individuals are randomly selected from across all 315 local authorities in England to ensure the sample representing the wider population. The description of the data is shown in Table 5.1 and Figure 5.2 shows the prior reproduction number (the blue area is the result of REACT group) The ONS result is shown in Table 5.1 and Figure 5.2 (the pink area is the statistical result of ONS). It describes the R_t simulated by ONS over the UK. To ensure high public value and quality, the statistics presented by ONS are in line with the Code of Practice for Statistics. These results are also used to determine the parameters of prior reproduction number.

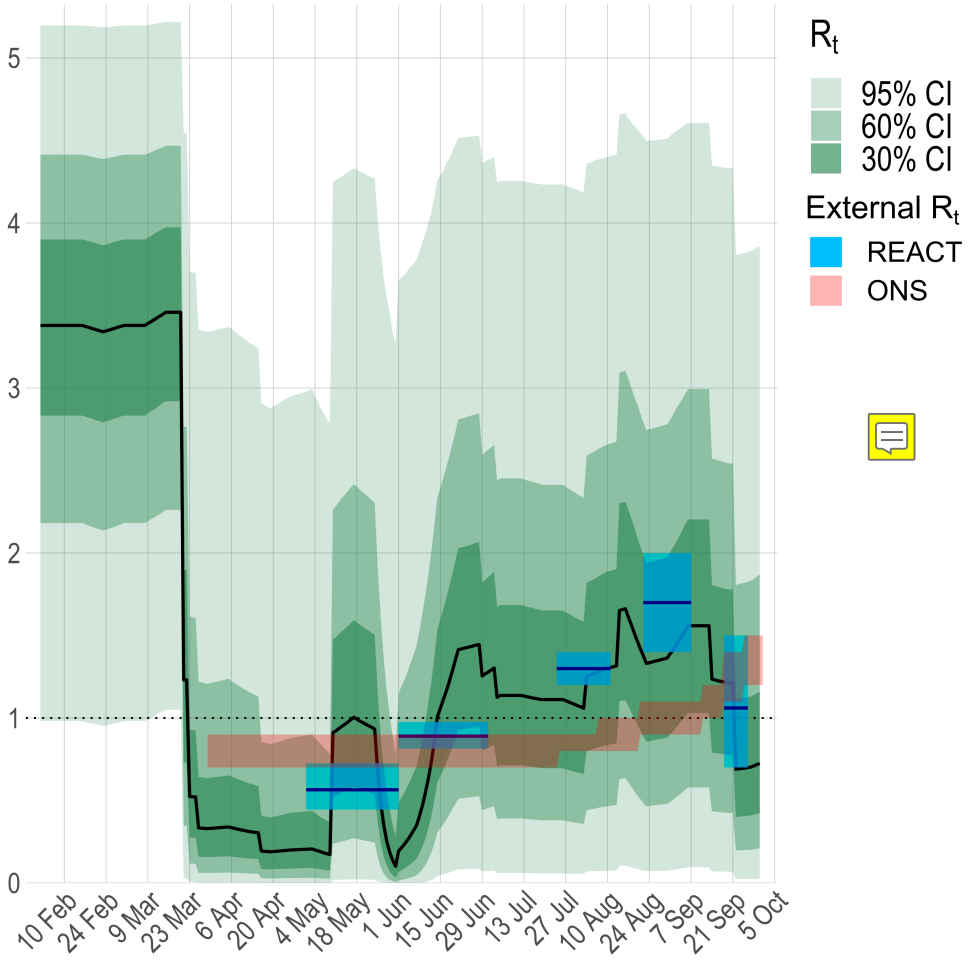

Figure 5.2: Prior reproduction number until Oct 1st, 2020

The median of prior production number is shown in the black line, with green areas indicating the 95%, 60%, and 30% confidence intervals. The red area is the ONS result, and the blue area is the REACT result. These two results locate approximately within the 60% CI of

the prior.

Typically, the more information considered, the more accurate prior. Here, by combining the lockdown policies, protest search index, and the REACT, the time series of prior reproduction number R_t could be achieved.

For the experimental implementations, assume the following assumptions hold.

1. During the simulation period, the mutation of the cornavirus did not occur.
2. The nature of the cornavirus will change o[illegible]if it mutates. The nature of the cornavirus means the serial interval, fatality rate, hospitalization rate etc.
3. The antibodies in the cured patients do not affect further infections.

The first assumption follows from previous works of estimating the serial interval for Covid-19. The second assumption ensures that the covid-19 is caused by the same strain, so the distributions of fatality rate, hospitalization rate, and other factors remain unchanged overtime. The third assumption assume that cured cases have no effect on R_t. These two assumption guarantees the reproduction number R_t to be predictable.

Based on these assumptions, if the time series data of reproduction number R_t is known, the daily increasing cases could be inferred. In the following section, R_t is estimated through the number of people in mechanical ventilation beds. Two alternative models of estimation through fatality data are also shown in the appendix (Section A.2).

5.2 Simu[illegible]tion through Daily Increasing People in Mechanical Ventilation Beds

In this section, the number of daily increasing cases is estimated through daily increasing people in mechanical ventilation beds data. Note that when reviewing the data, there is data missing in the mechanical ventilation beds and in hospital data, hence a quadratic imputation has been used to impute these data. There are three important assumptions for this part shown as follows.

1. The number of people in mechanical ventilation beds data is assumed to be inaccurate. This data is influenced by the daily people in hospital because of the limitation of mechanical ventilation beds. Since there are already many people occupied all mechanical ventilation beds, there are not enough beds available for the new coming patients who need to be treated in these beds.

From the Figure 5.3, it could be found that at the early stage of the epidemic, the daily increasing patient in the ventilation beds is rather underestimated. The simulated increasing number in the ventilation beds peaked on April 11, 12.9K (5.9K, 27.2K), while the increasing number in the ventilation beds reported on that day was merely $3,274$. It can also be seen here a very intuitive proportional relationship among the in hospital data, increasing in the ventilation beds, and simulated increasing in the ventilation beds. This fitness result from other aspect to checking the news that there is insufficient ventilation beds for patients in UK. This has also increased the fatality rate of the Covid-19 (also can be seen from the models in the appendix).

2. *The probability that an infected patient would in mechanical ventilated beds is constant.* The severe rate (13.8%) of covid-19 is used as the prior of this probability (Gomes, 2020).

3. *The time distribution of a person from infection to in the mechanical ventilation bed is fixed, comfort a certain distribution.* Firstly, the day distribution from infection to have symptom is known, comforting a log-normal distribution(Lauer et al., 2020). Meanwhile, as the WHO said, this mild patient would be severe quickly(Gomes, 2020). In experimental implementations, a shift of this distribution has been added.

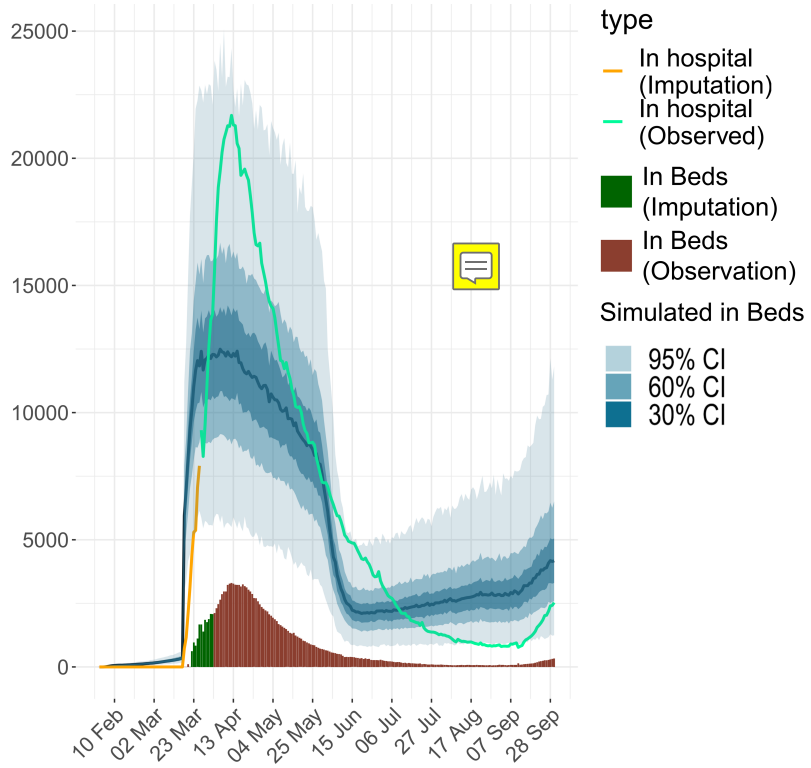

Figure 5.3: Simulation of daily increasing people in mechanical ventilation beds until Oct 1st, 2020

Since the daily increasing people in the mechanical ventilation beds are assumed to be inaccurate, the median of the simulation result of this data is shown in the blue line, with the blue areas representing the confidence intervals. The green and orange lines represent the imputation and observed people in hospital. The bars show the imputation and observed fatality data. This figure shows that the daily new increasing patient in the ventilation beds is underestimated at the early stage of the epidemic.

Combining the Assumptions 2 and 3 and prior reproduction number, the posterior reproduction number could be inferred as Figure 5.4. The green areas indicates the 30%, 60%, and 95% confidence intervals. The blue and red areas show REACT result and ONS result respectively. It can be observed that R_t reaches its minimum of 0.42 with the 95% CI $(0.39, 0.46)$ on May 31st. On June 17th, the 95% CI contains 10.98 $(0.95, 1.01)$. Afterwards, the 95% CI is always greater than or contains 1.

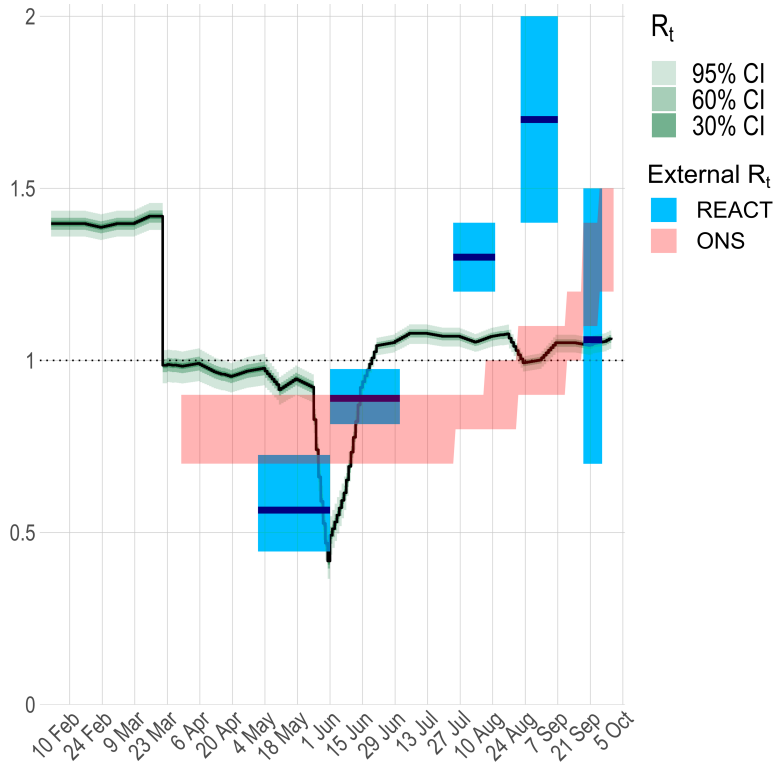

Figure 5.4: Simulation reproduction number from ventilated beds data until Oct 1st, 2020

The black line shows the median of posterior reproduction number. The green areas indicates the 30%, 60%, and 95% confidence intervals. The blue and red areas show REACT result and ONS result respectively.

Based on this result, the daily increasing cases could be inferred and the results are shown in Figure 5.5.

Figure 5.5 shows the cases, the observed people in hospital and in ventilation beds, and the corresponding imputation results. The simulation infection is described by the black line, with its 30%, 60%, and 95% confidence intervals marked as the blue areas. The peak of simulated daily increasing cases is nearly 70K (46k ,107k) on March 20th, which is approximately 19 days ahead of the peak of the daily increasing ventilation data (3301) on April 12th and the peak of the number of people in hospital (21,687) on April 12th. The shape of the simulated cases is different from the simulations through fatality data (see Section A.2). It can be observed that a high daily growth has been maintained for a period of time, while the infection cases later drops very quickly. Therefore the results simulated through mechanical ventilation beds are deemed to be better than those simulated through fatality. After June, as the lockdown relaxes, it can be clearly seen that the daily growth increases, which also indicates that it is inappropriate for the UK to ease the

lockdown policy at this time.

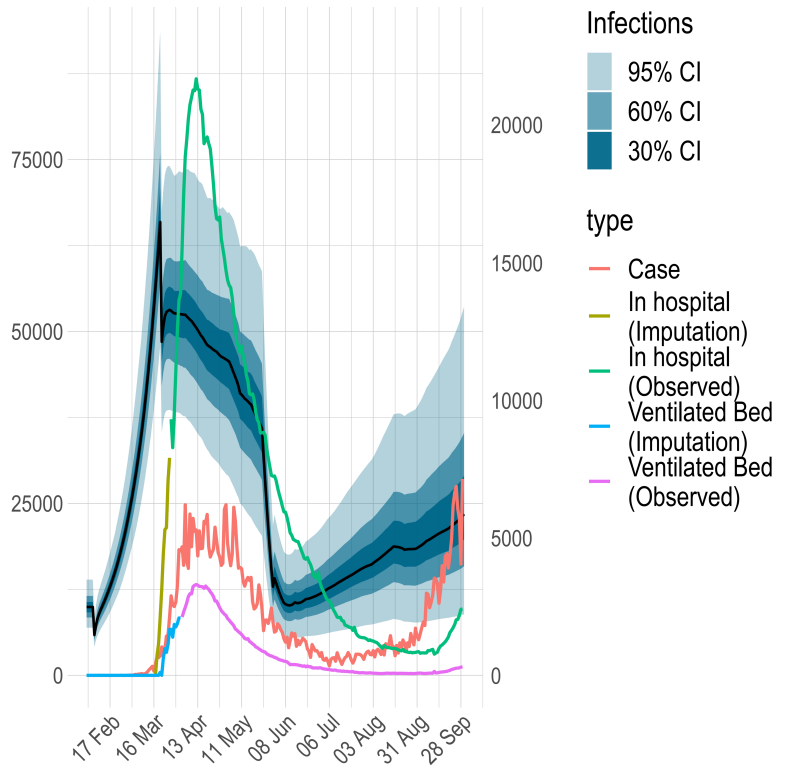

Figure 5.5: Simulation infection from ventilated beds data until Oct 1st, 2020

The black line shows the median of simulation infection from ventilated beds data (left y-axis). The blue areas indicates the 30%, 60%, and 95% confidence intervals. The red line represents the observed cases, the green lines are the imputation and observed number of people in hospital, and the blue and magenta lines are the imputation and observed number of people in ventilation bed. Note that only the infection data (in blue area) is shown on the left y-axis, and the observed data is shown on the right y-axis.

The cumulative cases simulated from the fatality data could been seen from figure 5.6. In the following, the simulation result is compared with the results of previous research (summarized in Table 5.2). The result of REACT group (Ward et al., 2021) is within the 95% CI of the simulation. The ONS results are within the 60%, 95%, and 60% CIs respectively. The research result in the Oxford area carried out by Oxford University (Lumley et al., 2020) also lies within the 60%CI of the simulation. Yet this is an experiment in a small region, so the credibility is not so high. In Great Glasgow region, the cumulative infection result (Thompson et al., 2020) lies outside the CIs of the simulation. Yet, Glasgow's sample size (470) is not large enough, and it is located in Scotland, so the credibility is low. Compared with the cumulative simulated cases from fatality (Figure A.4 and Figure A.8 in Section A.2), the simulation from ventilated beds show a better match with existing research results and thus is more reasonable.

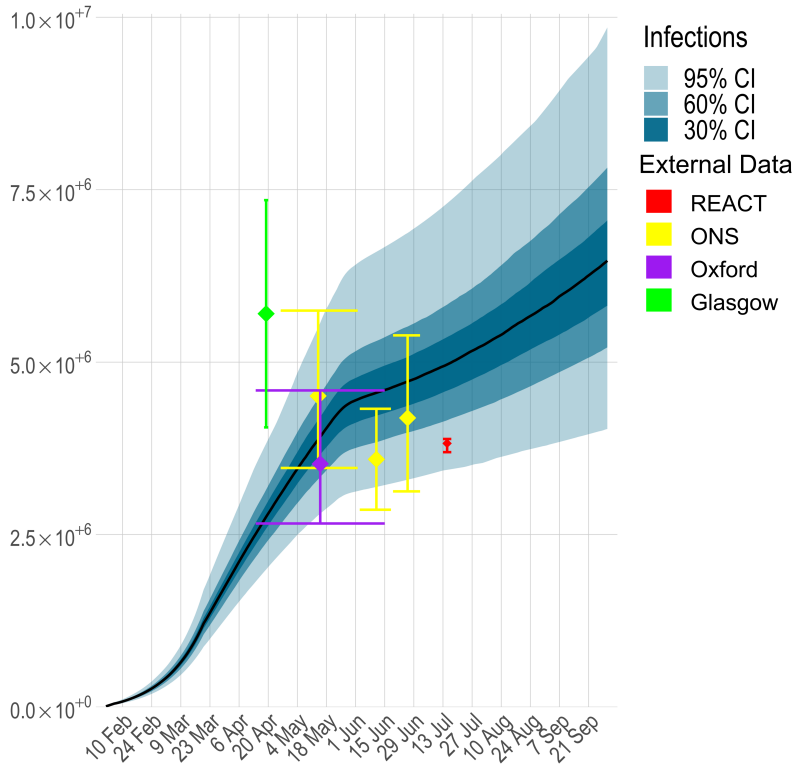

Figure 5.6: Cumulative simulated cases from ventilated beds data until Oct 1st, 2020

The black line shows the median of cumulative simulated cases from ventilated beds data. The blue areas indicates the 30%, 60%, and 95% confidence intervals. The result of REACT group (red line) is within the 95% CI of the simulation. The ONS results (yellow line) are within the 60%, 95%, and 60% CIs respectively. The research result in the Oxford area carried out by Oxford University (purple line) also lies within the 60%CI of the simulation. In Great Glasgow region, the cumulative infection result (green line) lies outside the CIs of the simulation.

Table 5.2: The cumulative cases estimated by previous research.

Color	Group	Area	Date	Estimation (million)	CI
red	REACT	UK	15 Jul	3.822	(3.695, 3.886)
yellow	ONS	UK	26 Apr - 2 Jun	4.510	(3.466, 5.747)
yellow	ONS	UK	19 Jun - 2 Jul	4.191	(3.126, 5.388)
yellow	ONS	UK	3 Jun - 18 Jun	3.592	(2.860, 4.324)
purple	Oxford	Oxford	15 May	3.526	(2.661, 4.590)
green	Oxford	Great Glasgow region	19 Apr	5.702	(4.054, 7.348)

Then divide the number of detected cases by the number of cases simulated through people in mechanical ventilation beds to stand for the sampling effect. Since the range of the result is large, a logarithm operation is performed and the result is shown in Figure 5.7. It can be observed that the sampling effect remains

under 100%. Since the detected cases are always underestimated due to the limitation of test ability ~~and the political consideration that the UK government is prone not to overestimate the cases~~, this result demonstrate the rationality of the simulation through people in mechanical ventilation beds. Compared with the results simulated through fatality (Figure A.5 and Figure A.9 in Section A.2), where the ratio of detected cases by the simulated cases eventually exceeds 100%, the simulation through ventilated beds turns out to be more reasonable. Observed from the shape of the figure, the simulation rises over time and reached 23.77% with CI $(8.81\%, 66.64\%)$ on the last day of observation. This indicates that although the effectiveness of testing is constantly improving, the percentage of detected cases is still relatively low. The British government still has to work hard to enhance the sampling effort.

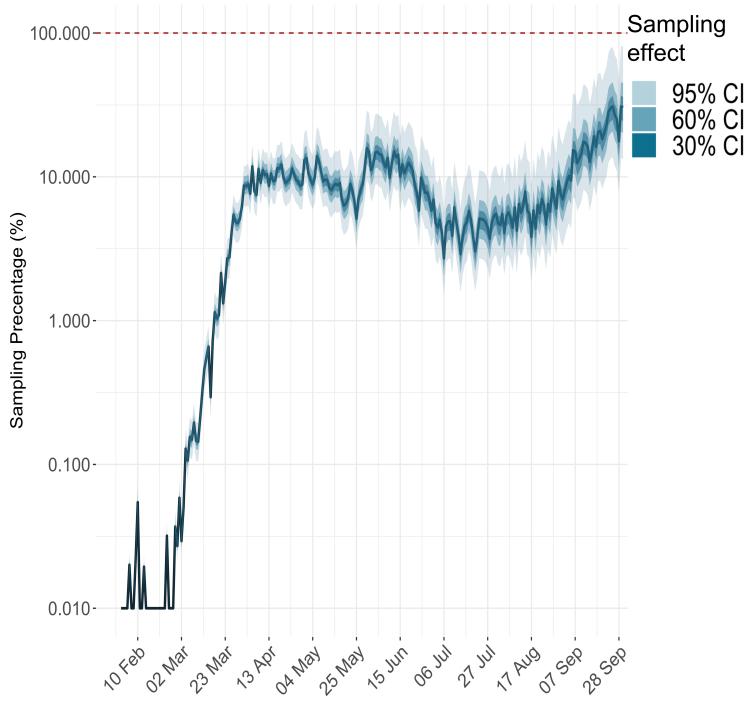

Figure 5.7: The number of detected cases by the number of cases simulated with people in ventilation beds until Oct 1st, 2020

The black line shows the median of the sampling effect. The blue areas indicates the 30%, 60%, and 95% CIs. It can be observed that the sampling effect reaches 23.77% with CI $(8.81\%, 66.64\%)$ and remains under 100%.

Chapter 6

Discussions

This thesis studies the sampling effect on the COVID-19 epidemics within the UK during the period from Jan 31^{st} 2020 to October 2020. The model comprises three parts: the epidemic modelling describes the underlying relationships between observed data and epidemiological parameters; the bayesian regression demonstrates how powerful the major interventions are against the COVID-19 based on historical researches and empirical experiments; the MCMC simulation broadens the distributional estimation obtained from the bayesian regression to a interpretable result with various confidence levels. Trained with the data provided by the official UK government website for data and insights on COVID-19, the model has the ability to justify the satisfactory performances of the major interventions in preventing the epidemics from becoming worse. Specifically, the overall procedures can be summarized as follows. Firstly, the public policies limiting the wide-range or heavy-crowded social activities are proved to be effective on reducing the level of the reproduction numbers of COVID-19. Then, small reproduction numbers slow the speed of the transmission from infected people to healthy ones. Finally, the observed data such as fatality or numbers of mechanical ventilation beds will decrease.

The experimental results suggest several patterns of the spread of the COVID-19. Firstly, the effectiveness of NPIs like lockdown inversely verify the hypothesis that population movements or any close-distance social activities play a major role in the infection process of COVID-19 within the UK. The lockdown and other interventions are implemented after the early stage of COVID-19, Consequently, the surge of fatality is not immediately stopped, but the NPIs did prevent the upward momentum of the epidemic. By matching the timeline of government policies with the estimated transmissions in time-series format, the effect of the non-pharmaceutical interventions (NPI) is able to be interpreted. The results show that major non-pharmaceutical interventions, especially lockdowns, have had a large effect on reducing transmission. Fore example, the lockdown policy implemented

in March significantly controls the simulation infection data, where the infections drops rapidly. This can be seen from results of all three models, i.e. Figures A.3, A.7, and 5.5. Next, the effects of different NPIs vary from each other. For example, the closures of universities or other educational facilities reduced contact among susceptible population, who spent long time together in in-door places. In addition, social distancing intervention seems to be a little move but has been especially helpful by ascertaining a large amount of daily face-to-face interactions being non-infectious. Thirdly, with evidences that the ease of NPIs can lead to the resume of the COVID-19, the government and people are supposed to follow part of the NPIs, which is the safeguard to ensure that the epidemic is always in control or to alleviate the influences brought by another wave of COVID-19 and its variants, for example, the Delta. Last but not least, the models reveal the sampling effects in the UK of the COVID-19 epidemics, indicating the gap between 'true' infections and the detected numbers. According to this index, the results of the experiments show that the UK government did well after the first outbreak around 28^{th} March, where the sampling effect keeps high afer then.

Nevertheless, the study has several limitations. Firstly, as an application of bayesian regression, the model is established on the epidemiology parameters with prior distributions that were provided in historical researches or empirical approximations for other countries, and may not be perfectly suitable for certain cases in the UK; As a result, if the epidemiology researches about the COVID-19 within the UK have been published, the model can be more accurate. Secondly, during the target period in this thesis, other factors such as movements of the population may affect the spread of COVID-19 as well. This model has potential to be improved with a more complex epidemic model with other epidemiology factors, where a trade-off of the computational cost and interpretability will be the new problem. Thirdly, although the bias in observed data has been taken into consideration, the ground truth is never accessible. Therefore, these model may demonstrate an incomplete procedures of the spread of the epidemics. Next, three models designed in this paper do not include the "daily additions of infections" published by the UK government as a variable. It is possible to utilize the recorded daily additions to modify the models, decreasing the bias and variance of the models. Finally, the contributions belonging to other interventions besides the target ones is not considered in this study, where their combined effects may be mingled into some of the major NPIs and exaggerate the effect of those studied in the model.

To sum up, COVID-19 has placed a heavy burden on health system and society all over the world, harming the agriculture, economic and many other aspects in human life. This study highlight that the government is supposed to implement poli-

cies about the early implementation of combination of different NPIs to minimize the loss to the country, and even to the world.

Appendix A

Further Explanation about Methodology

A.1 Settings of prior distribution for different observed data

Due to the fact that different kinds of observed data has different methods of the calculation of time interval between being infected and being observed, different settings of prior distributions are supposed to be demonstrated. In this section, two settings for "observation - deaths" and "observation - inaccurate deaths" are added. The only differences lie in the infection process, where the distribution of π and the link function can differ.

For "observation - deaths" and "observation - inaccurate deaths", the prior distribution of time interval π is calculated empirically from the data, compared to the log normal distribution used in "observation - People in Ventilation Beds". Meanwhile, when "observation - deaths" directly uses the data recorded by the government as Y_t, "observation - inaccurate deaths" and "observation - People in Ventilation Beds" build a generalized linear model with an auxiliary variable "inhospital", which means three problems share different assumptions toward the data, When "deaths" believes the data completely, the other two suspect the bias and introduce a new variable to enhance the data.

Specifically, "observation - inaccurate deaths" holds

$$\begin{aligned}\mu_{\text{deaths}} &= \beta_0 + \beta_1 \text{inhospital} + \varepsilon_{\text{deaths}} \\ \text{deaths} &= \frac{0.02}{1 + e^{-\mu_{\text{deaths}}}}\end{aligned}$$

where the noise term $\varepsilon_{\text{deaths}} \sim N(0, 0.3)$.

A.2 Simulation through Fatality

In this section, the R_t is estimated through fatality data. Two important assumptions are shown as follows.

1. The death probability of an infected patient is assumed to be constant, and the mean value is 0.66%. Please refer to Mahase (2020) for more details.

2. The distribution of the duration from infection to fatality is assumed to be fixed and follows from a certain distribution. According to Flaxman et al. (2020), the modelled deaths are informed by the infection-to-death distribution.

A.2.1 Inferring increasing cases using death data which is assumed to be accurate

In this section, the daily death number is shown to be accurate. Most of the dead cases experience severe symptoms, so most of the deaths will occur in the hospital. Therefore, in building the model, it is intuitive to try the assumption that fatalities are accurately recorded. Figure A.1 shows the fitness of observation from death data from February 1st, 2020 to October 1st, 2020. The black line represents the median of fatality estimations and the blue areas are the 30%, 60%, and 95% intervals. The brown bars represent the observed death. From Figure A.1, it could be found that this model fits the observed death data well, nearly all observed death data are in the 95% confident interval of the fitted model. Therefore, it is reasonable to consider the observed death data to be accurate overtime.

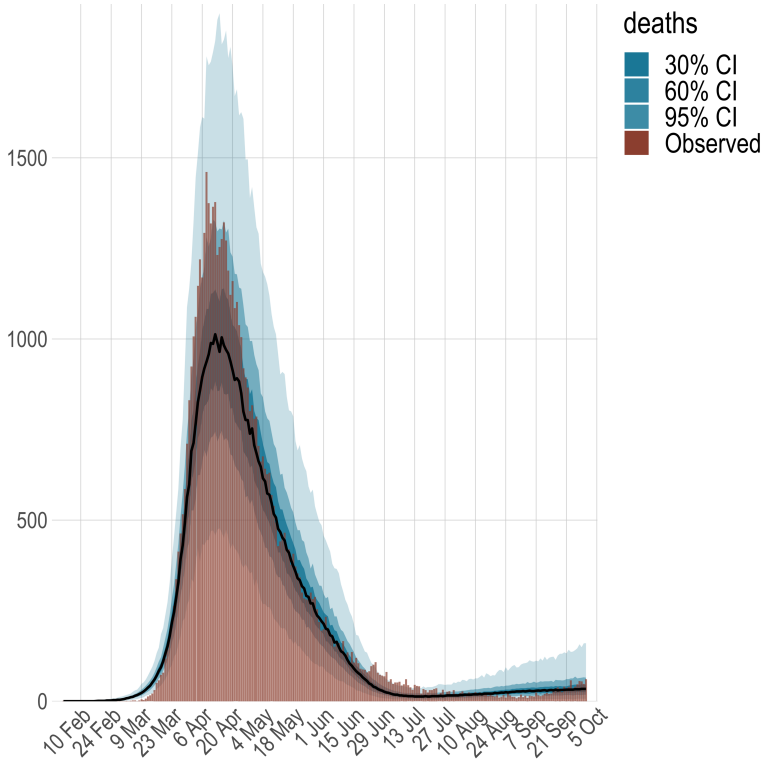

Figure A.1: Fitness of observation from death data until Oct 1st, 2020

The black line is the median of the fitted fatality data with confidence intervals shown in the blue areas. The brown bar is the observed death. This figure shows that the fatality data is relatively well fitted.

Combining Assumption 1 and 2 and prior reproduction number, the posterior reproduction number could be inferred as Figure A.2 and the variance of the reproduction number is significantly reduced. It can be seen that the prior doesn't vary too far away from the REACT result. Since the REACT group estimates R_t through survey, the result is highly likely to contain noisy information. Therefore, the prior result shown in Figure A.2 is considered to be acceptable.

Based on the simulation reproduction number, the daily increasing cases could be inferred. The result is shown in Figure A.3.

From the figure A.3, it could be found that the peak of simulated daily increasing cases is nearly 168K (120k,237k) on March 20th, which is approximately 19 days ahead of the peak of the daily data on April 8th (1461).

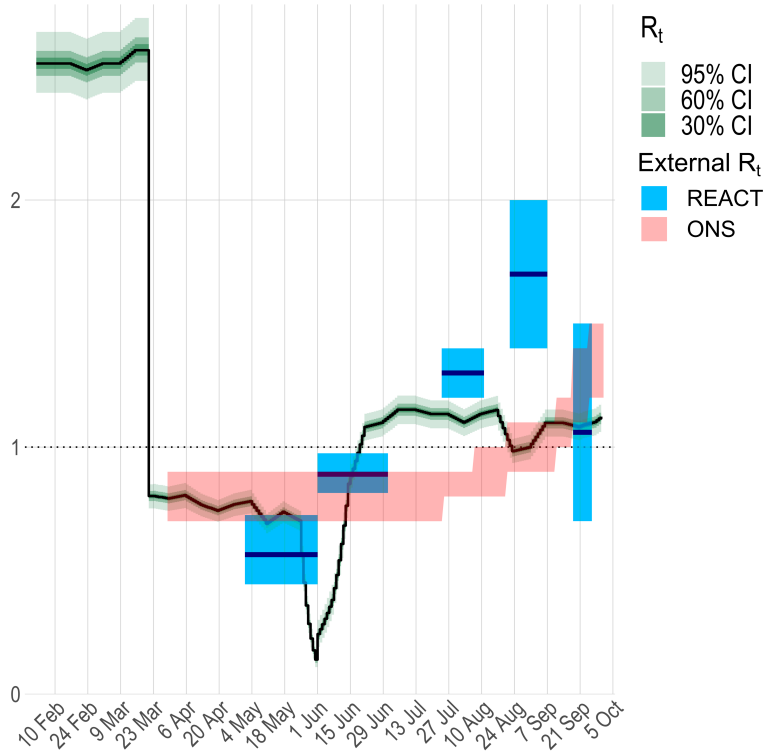

Figure A.2: Posterior reproduction number from death data until Oct 1st, 2020

The black line shows the median of simulation reproduction number. The green areas indicates the 30%, 60%, and 95% confidence intervals. The blue and red areas show REACT result and ONS result respectively.

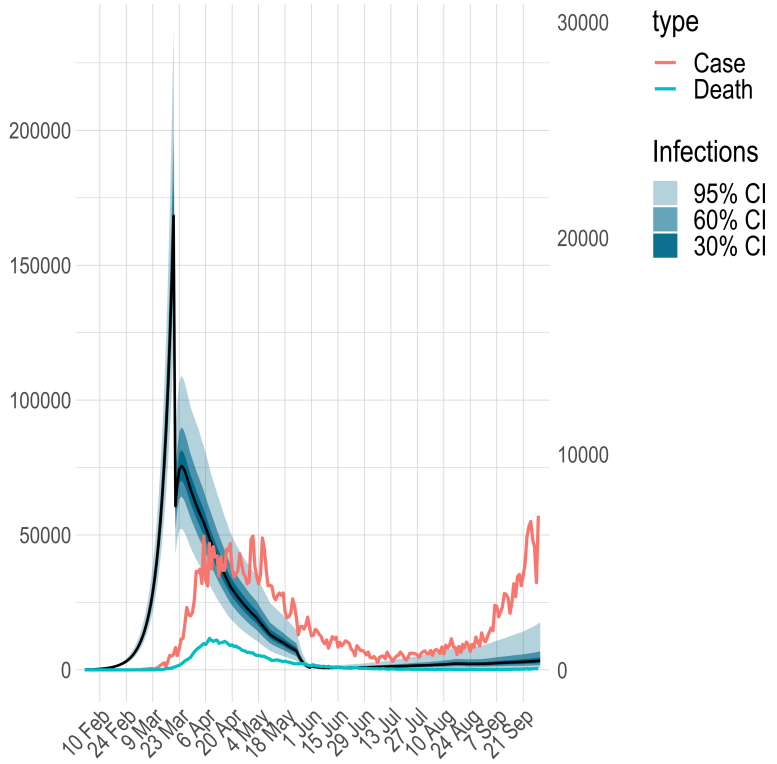

Figure A.3: Simulation infection from death data until Oct 1st, 2020

The black line shows the median of simulation (left y-axis). The blue areas indicates the 30%, 60%, and 95% confidence intervals. The red line represents the observed cases, and blue line shows the observed death. Note that the infection data is shown on the left y-axis, and the observed data is shown on the right y-axis.

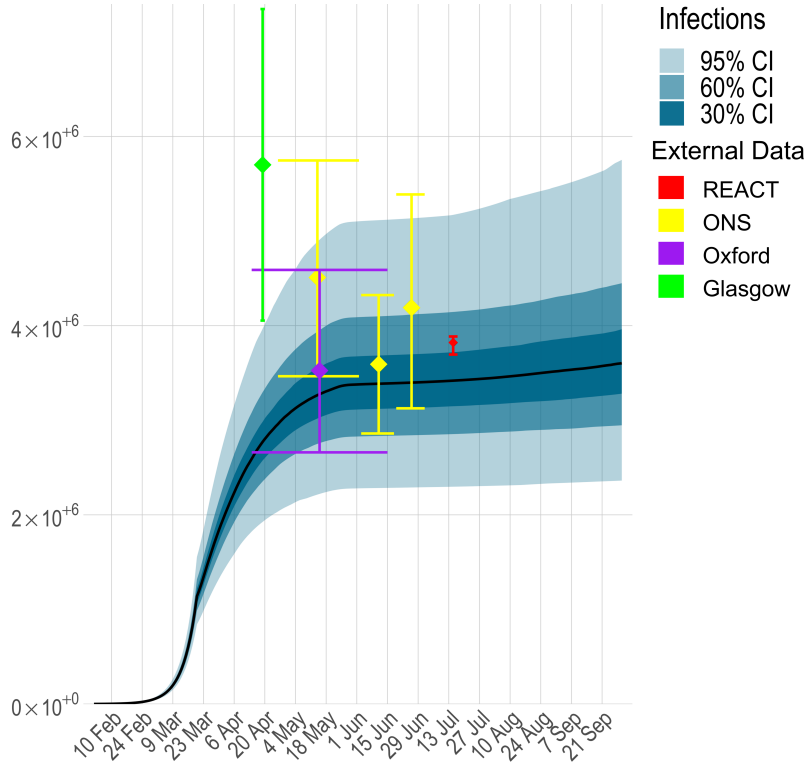

Figure A.4: Cumulative simulated cases from death data until Oct 1st, 2020

The black line shows the median of cumulative simulated cases. The blue areas indicates the 30%, 60%, and 95% confidence intervals. The result of REACT group (red line) is within the 95% CI of the simulation. The ONS results (yellow line) are within the 95%, 30%, and 95% CIs respectively. The research result in the Oxford area carried out by Oxford University (purple line) also lies within the 30%CI of the simulation. In Great Glasgow region, the cumulative infection result (green line) lies outside the CIs of the simulation.

The cumulative cases simulated from the death data could been seen from figure A.4. In the following, the simulation result is compared with the results of previous research (summarized in Table 5.2). The result of REACT group (Ward et al., 2021) is within 60% CI of the simulation. The ONS results are within 95%, 30%, and 95% CI respectively. The Oxford University also has another research in the Oxford area (Lumley et al., 2020). The result lies within 30% CI of the simulation. In Great Glasgow region, the cumulative infection result (Thompson et al., 2020) lies out of the CIs of the simulation.

Then divide the number of detected cases by the number of cases simulated with death data, and show the sampling effect in Figure A.5. It can be observed that the sampling effect exceeds 100% after June, which demonstrate that the simulation through death is not entirely accurate, since it is well acknowledged that the infected cases can only be under-detected. This indicates that the assumption that death data is accurate may not be plausible. Therefore, in the following section, the death data

is assumed to be inaccurate to conduct the simulation.

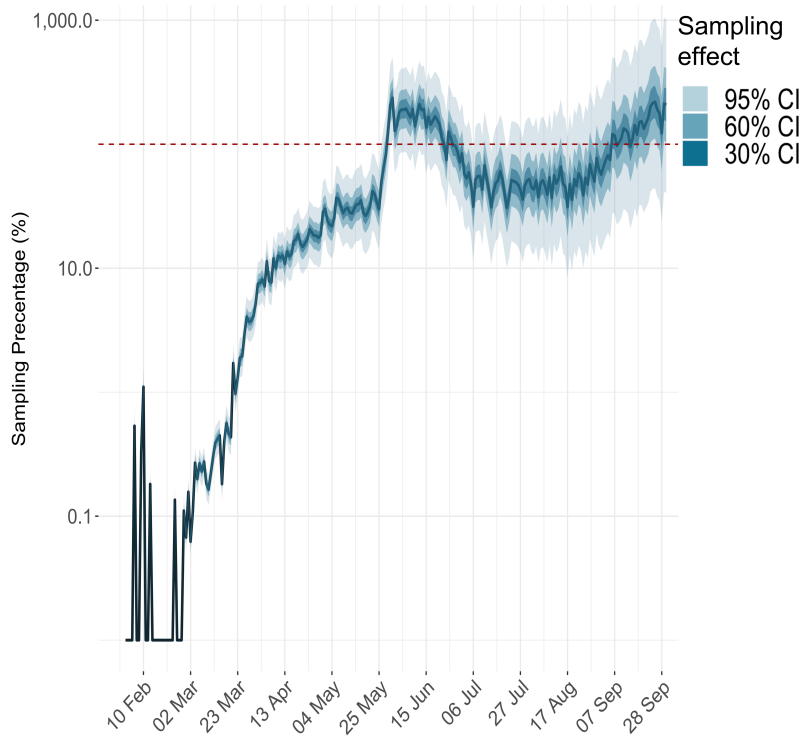

Figure A.5: the number of detected cases by the number of cases simulated with death data until Oct 1st, 2020

The black line shows the median of the sampling effect. The green areas indicates the 30%, 60%, and 95% CIs. It can be observed that the sampling effect approaches 100% and exceeds 100% over time.

A.2.2 Inferring increasing cases using death data which is assumed to be inaccurate

In this section, the number of increasing cases is estimated through death data while the fatality data is assumed to be inaccurate and that the death data changes as the number of people in hospital changes over time.

Then combining the assumptions 2 and 3 and prior reproduction number, the posterior reproduction number could be inferred as figure A.6 and the variance of the reproduction number is significantly reduced.

Based on the simulation reproduction number, the daily increasing cases could be inferred. The result is shown in Figure A.7.

From the figure A.7, it could be found that the peak of simulated daily increasing cases is nearly 271K with the 95% confidence interval (193k ,386k) on March 20th, which is approximately 19 days ahead of the peak of the daily data (1461) on

April 8th and 15 days ahead of the peak of the number of people in hospital (21,687) on April 12th.

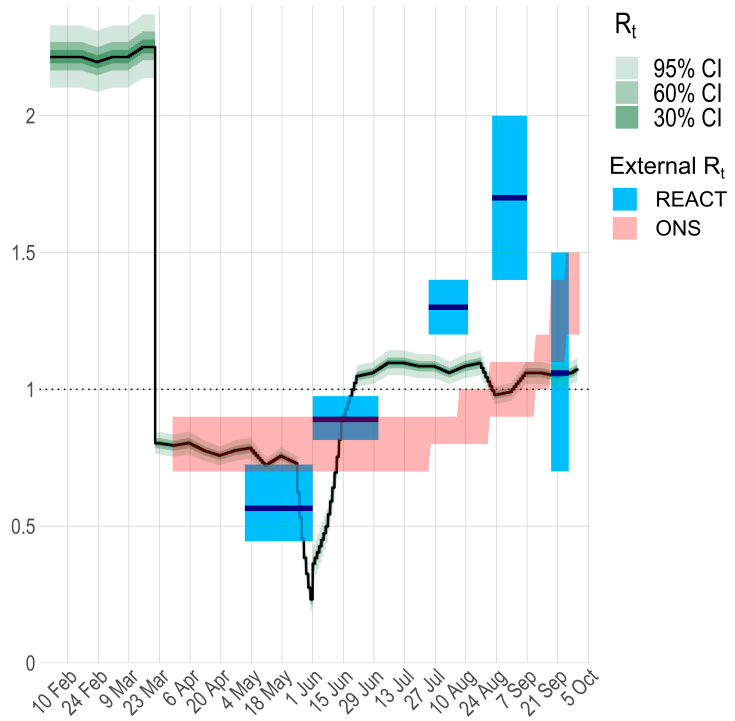

Figure A.6: Simulation reproduction number from death data until Oct 1st, 2020

The black line shows the median of simulation reproduction number. The green areas indicates the 30%, 60%, and 95% confidence intervals. The blue and red areas show REACT result and ONS result respectively.

The cumulative cases simulated from the death data could been seen from figure A.8. In the following, the simulation result is compared with the results of previous research (summarized in Table 5.2). The result of REACT group and the ONS results are without the CIs of the simulation. The research result in the Oxford area carried out by Oxford University also lies without the CIs of the simulation. In Great Glasgow region, the cumulative infection result lies within the 30% CI of the simulation.

Then divide the number of detected cases by the number of cases simulated with death data, and show the sampling effect in Figure A.9. It can be observed that the sampling effect (including the 30% CI) remains under 100% while approaches 100% over time, which demonstrate that the simulation through death is reasonable when assuming that death data is inaccurate.

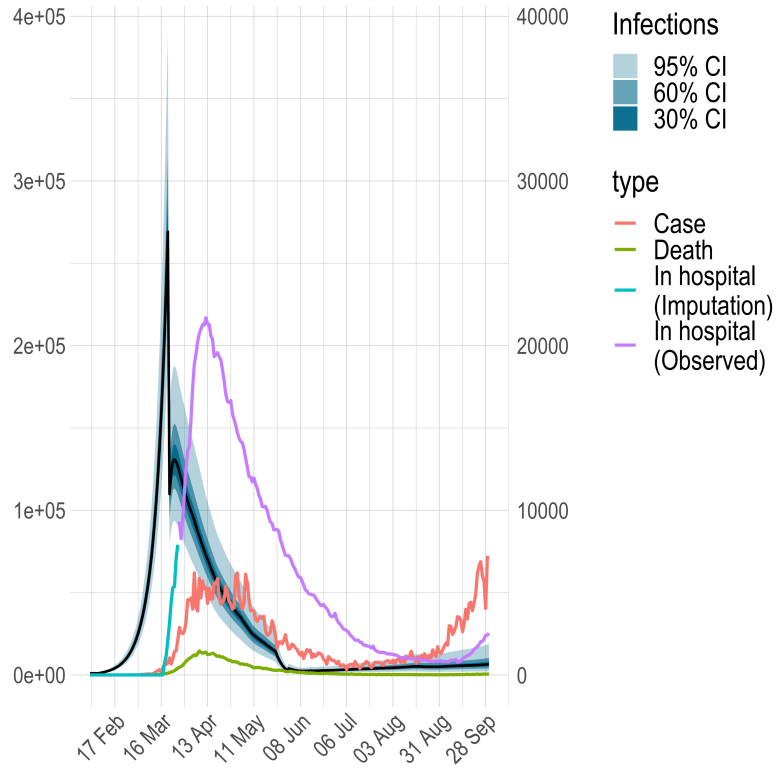

Figure A.7: Simulation infection from death data until Oct 1st, 2020

The black line shows the median of simulation (left y-axis). The blue areas indicates the 30%, 60%, and 95% confidence intervals. The red line represents the observed cases, and blue line shows the observed death. Note that the infection data is shown on the left y-axis, and the observed data is shown on the right y-axis.

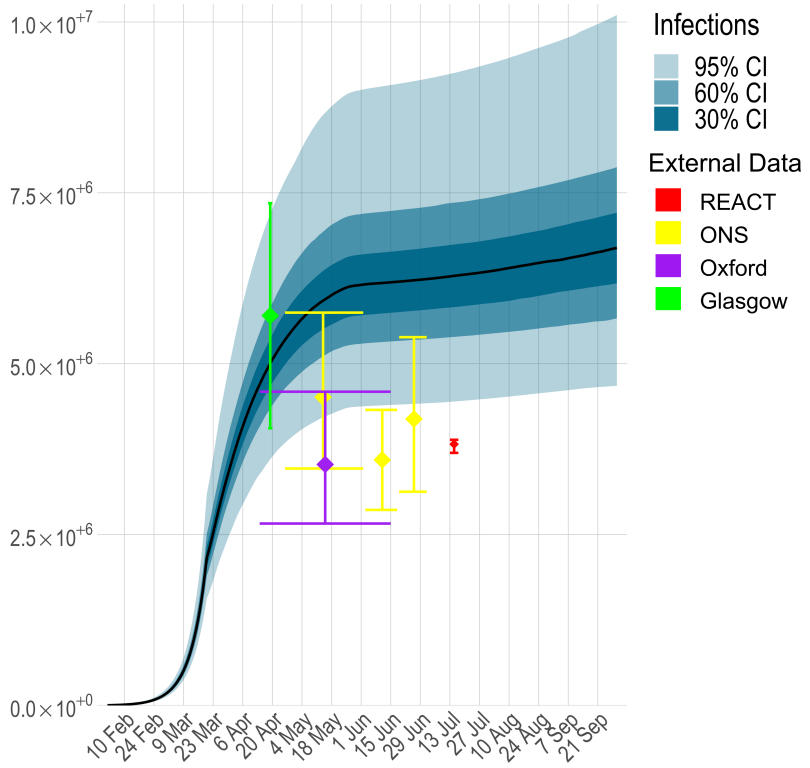

Figure A.8: Cumulative simulated cases from death data until Oct 1st, 2020

The black line shows the median of cumulative simulated cases. The blue areas indicates the 30%, 60%, and 95% confidence intervals. The result of REACT group (red line), the ONS results (yellow line), and result in the Oxford area carried out by Oxford University (purple line) are outside the CIs of the simulation. In Great Glasgow region, the cumulative infection result (green line) lies winthin the 30% CI of the simulation.

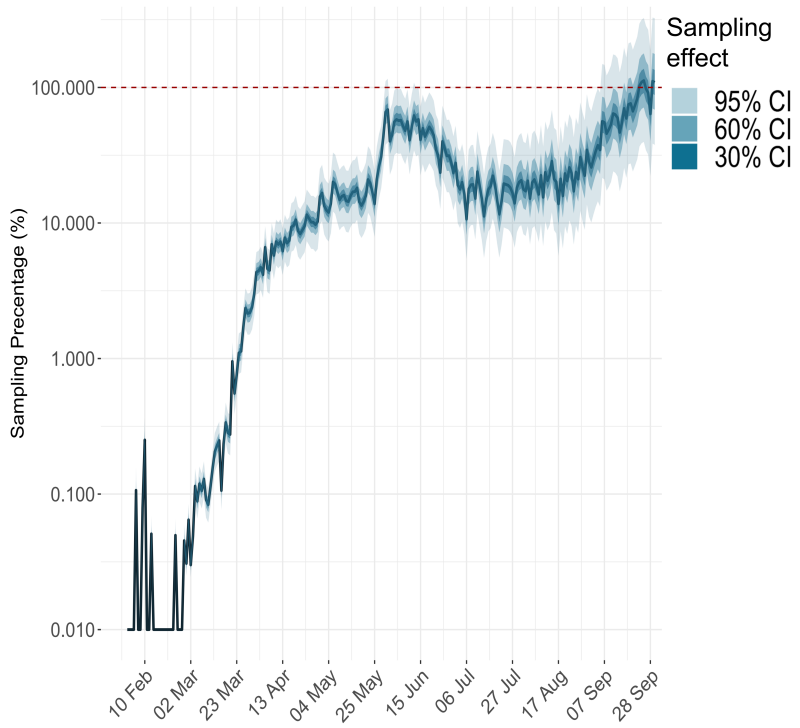

Figure A.9: the number of detected cases by the number of cases simulated with death data until Oct 1st, 2020

The black line shows the median of the sampling effect. The green areas indicates the 30%, 60%, and 95% CIs. It can be observed that the sampling effect approaches 100% and the 60% CI exceeds 100% over time.

Appendix B

Code

```
library(epidemia)
plot.cases <- function(method, data) {
  # reproduction number
  rt <- epirt(
    formula = R(Nation, date) ~ 1 + Protest +
      schools_universities + Lockdown,
    prior = shifted_gamma(
      shape=2, scale = 1/3, shift = log(1.05)/6
    ),
    prior_covariance = decov(
      shape = c(2, rep(0.5, 5)),scale=0.25
    ),
    prior_intercept = rstanarm::normal(0.5,1),
    link = scaled_logit(5.7)
  )

  # the infection model
  inf <- epiinf(
    gen = EuropeCovid2$si,
    seed_days = 5
  )
  if (method == 'deaths') {
    observation <- epiobs(
      formula = deaths ~ 1,
      i2o = EuropeCovid2$inf2death,
      prior_intercept = normal(0, 1),
      link = scaled_logit(0.02)
    )
  } else if (method == 'Beds') {
```

```
  observation <- epiobs(
    formula = Beds ~ 1 + inhospital,
    i2o = dlnorm(1:20, 1.921, 0.428),

    link = scaled_logit(0.276),
    prior_intercept = normal(0, 0.5),
    center = TRUE
  )
} else if (method == 'deaths_inacurate') {
  observation <- epiobs(
    formula = deaths ~ 1+inhospital,
    i2o = EuropeCovid2$inf2death,
    prior_intercept = normal(0, 0.3),
    link = scaled_logit(0.02)
  )
  method <- 'deaths'
}

args <- list(
  rt=rt, inf=inf, obs=observation,
  data=data, seed=12345, refresh=0
)
options(mc.cores = parallel::detectCores())
pr_args <- c(
  args,
  list(
    algorithm='sampling', iter=1e4,
    prior_PD=TRUE, control = list(adapt_delta = 0.99)
  )
)

# prior R_t
fm_prior <- do.call(epim, pr_args)
p1 <- plot_rt(fm_prior, levels = c(30, 60, 95))
args$algorithm <- 'fullrank'
args$iter <- 50000
args$tol_rel_obj <- 1e-8
fm <- do.call(epim, args)
p2 <- plot_rt(
  fm, step = T, levels = c(30, 60, 95)
```

```
  )
  p3 <- plot_obs(
    fm, type = method, step = T, levels = c(30, 60,
        95)
  )
  p4 <- plot_infections(
    fm, step = T, levels = c(30, 60, 95)
  )
  p5 <- plot_infections(
    fm, cumulative = TRUE, step = T, levels = c(30,
        60, 95)
  )
  return(
    list(
      prior = p1, rt = p2, obs = p3, infection = p4,
          cum = p5
    )
  )
}
```

Bibliography

Yousef Alimohamadi, Maryam Taghdir, and Mojtaba Sepandi. Estimate of the basic reproduction number for covid-19: a systematic review and meta-analysis. *Journal of Preventive Medicine and Public Health*, 53(3):151, 2020.

Richard Bellman and Theodore Harris. On age-dependent binary branching processes. *Annals of Mathematics*, pages 280–295, 1952.

Richard Bellman and Theodore E Harris. On the theory of age-dependent stochastic branching processes. *Proceedings of the National Academy of Sciences of the United States of America*, 34(12):601, 1948.

Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A Scott. Semi-mechanistic bayesian modeling of covid-19 with renewal processes. *arXiv preprint arXiv:2012.00394*, 2020.

Qifang Bi, Yongsheng Wu, Shujiang Mei, Chenfei Ye, Xuan Zou, Zhen Zhang, Xiaojian Liu, Lan Wei, Shaun A Truelove, Tong Zhang, et al. Epidemiology and transmission of covid-19 in shenzhen china: Analysis of 391 cases and 1,286 of their close contacts. *MedRxiv*, 2020.

Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boelle, Antoine Flahault, and Neil M Ferguson. Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750–754, 2008.

Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512, 2013.

Nicholas G Davies, Adam J Kucharski, Rosalind M Eggo, Amy Gimma, W John Edmunds, Thibaut Jombart, Kathleen O'Reilly, Akira Endo, Joel Hellewell, Emily S Nightingale, et al. Effects of non-pharmaceutical interventions on covid-19 cases, deaths, and demand for hospital services in the uk: a modelling study. *The Lancet Public Health*, 5(7):e375–e385, 2020.

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. 2020.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

Christophe Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS one*, 2(8):e758, 2007.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Carlos Gomes. Report of the who-china joint mission on coronavirus disease 2019 (covid-19). *Brazilian Journal of Implantology and Health Sciences*, 2(3), 2020.

GOV.UK. Gov.uk, 2021. URL `https://coronavirus.data.gov.uk/`.

Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8(4):e488–e496, 2020.

Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395 (10223):497–506, 2020.

Institute for government. Institute for government, 2021. URL `www.instituteforgovernment.org.uk`.

M John. *A dictionary of epidemiology*. Oxford university press Oxford, UK, 2001.

Thibaut Jombart, Kevin Van Zandvoort, Timothy W Russell, Christopher I Jarvis, Amy Gimma, Sam Abbott, Sam Clifford, Sebastian Funk, Hamish Gibbs, Yang Liu, et al. Inferring the number of covid-19 cases from recently reported deaths. *Wellcome Open Research*, 5, 2020.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172 (9):577–582, 2020.

Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493, 2020.

Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simmons, Simon Thelwall, Julia Stowe, Elise Tessier, Natalie Groves, Gavin Dabrera, et al. Effectiveness of covid-19 vaccines against the b. 1.617. 2 (delta) variant. *New England Journal of Medicine*, 2021.

Sheila F Lumley, David W Eyre, Anna L McNaughton, Alison Howarth, Sarah Hoosdally, Stephanie B Hatch, James Kavanagh, Kevin K Chau, Louise O Downs, Stuart Cox, et al. Sars-cov-2 antibody prevalence, titres and neutralising activity in an antenatal cohort, united kingdom, 14 april to 15 june 2020. *Eurosurveillance*, 25(42):2001721, 2020.

Elisabeth Mahase. Covid-19: death rate is 0.66% and increases with age, study estimates. *BMJ: British Medical Journal (Online)*, 369, 2020.

Pierre Nouvellet, Anne Cori, Tini Garske, Isobel M Blake, Ilaria Dorigatti, Wes Hinsley, Thibaut Jombart, Harriet L Mills, Gemma Nedjati-Gilani, Maria D Van Kerkhove, et al. A simple approach to measure transmissibility and forecast incidence. *Epidemics*, 22:29–35, 2018.

Steven Riley, Kylie EC Ainslie, Oliver Eales, Benjamin Jeffrey, Caroline E Walters, Christina J Atchison, Peter J Diggle, Deborah Ashby, Christl A Donnelly, Graham Cooke, et al. Community prevalence of sars-cov-2 virus in england during may 2020: React study. *medRxiv*, 2020a.

Steven Riley, Kylie EC Ainslie, Oliver Eales, Caroline E Walters, Haowei Wang, Christina J Atchison, Peter Diggle, Deborah Ashby, Christl A Donnelly, Gra-

ham Cooke, et al. Transient dynamics of sars-cov-2 as england exited national lockdown. *medRxiv*, 2020b.

Steven Riley, Kylie EC Ainslie, Oliver Eales, Caroline E Walters, Haowei Wang, Christina J Atchison, Claudio Fronterre, Peter J Diggle, Deborah Ashby, Christl A Donnelly, et al. High prevalence of sars-cov-2 swab positivity in england during september 2020: interim report of round 5 of react-1 study. *medRxiv*, 2020c.

Steven Riley, Kylie EC Ainslie, Oliver Eales, Caroline E Walters, Haowei Wang, Christina Atchison, Claudio Fronterre, Peter J Diggle, Deborah Ashby, Christl A Donnelly, et al. Resurgence of sars-cov-2: Detection by community viral surveillance. *Science*, 372(6545):990–995, 2021.

Julien Riou and Christian L Althaus. Pattern of early human-to-human transmission of wuhan 2019 novel coronavirus (2019-ncov), december 2019 to january 2020. *Eurosurveillance*, 25(4):2000058, 2020.

Vital Surveillances. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19)—china, 2020. *China CDC weekly*, 2(8): 113–122, 2020.

Craig P Thompson, Nicholas E Grayson, Robert S Paton, Jai S Bolton, José Lourenço, Bridget S Penman, Lian N Lee, Valerie Odon, Juthathip Mongkolsapaya, Senthil Chinnakannan, et al. Detection of neutralising antibodies to sars-cov-2 to determine population exposure in scottish blood donors between march and may 2020. *Eurosurveillance*, 25(42):2000685, 2020.

Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick Walker, Han Fu, et al. Estimates of the severity of covid-19 disease. *MedRxiv*, 2020.

Helen Ward, Christina Atchison, Matthew Whitaker, Kylie EC Ainslie, Joshua Elliott, Lucy Okell, Rozlyn Redd, Deborah Ashby, Christl A Donnelly, Wendy Barclay, et al. Sars-cov-2 antibody prevalence in england following the first peak of the pandemic. *Nature communications*, 12(1):1–8, 2021.

Juanjuan Zhao, Quan Yuan, Haiyan Wang, Wei Liu, Xuejiao Liao, Yingying Su, Xin Wang, Jing Yuan, Tingdong Li, Jinxiu Li, et al. Antibody responses to sars-cov-2 in patients with novel coronavirus disease 2019. *Clinical infectious diseases*, 71 (16):2027–2034, 2020.