# Reconstructing the first COVID-19 pandemic wave with minimal data

Siyu Chen[1], Lisa J White[1]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, United Kingdom

siyu.chen@ndm.ox.ac.uk & lisa.white@ndm.ox.ac.uk

## Abstract

Early pandemic diagnostic tool is key important to provide timely pieces of evidence for public health decision making in the early pandemic where data was sparse. In this paper, we demonstrate how a model can be used to reconstruct the first COVID-19 pandemic wave with minimal dataset.

## Introduction

The COVID-19 pandemic has been with us more than three years, inflicting devastating effects on global populations and economies [1, 2] and now still affecting countries in very different ways. Reviewing the COVID-19 pandemic evolution and evaluating previous responses is vital important for future pandemic preparedness [3-6].

In the early pandemic information related with COVID-19 transmission was sparse. Different types of data stream were available with the pandemic was evolving into different stages. Epidemiological data including confirmed cases and mortality was always the first one to be collected and reported mostly due to the syndrome surveillance systems [7, 8]. Although the reported case is an important metric to tract the pandemic, it is always underestimating exposure because of the very limited capacity of viral tests in the beginning of the pandemic. The rate of underestimating was estimated mostly by SEIR type compartment models [9-12] while a probabilistic convolution model linking the distribution of the time from infection to death and the probability of death given infection was developed [13]. The

rate varied along the pandemic and depended on the relation between exposure, testing capacity and testing behaviour.

Serological data from convenient samples, e.g., blood donors [14] and seroprevalence survey [15] became accessible after limited by the availability of antibody testing kits and feasibility of large-scale sampling. These data were collected to understand the exposure level in the population and its immunological implications. However, extra work of adjustment must be done to correct the bias introduced by the antibody decay [16].

Along with seroprevalence survey, viral survey was conducted at the same time or a bit later. The viral survey was designed to understand the transmission and prevalence so that various non-pharmaceutical interventions could be developed and implemented. UK Office for National Statistics conducted a national wide COVID-19 viral testing survey [17] that has been successfully tracking the trajectories of COVID-19 infections in the UK since April of 2020. It has provided a clearer picture of the pandemic evolution and valuable pieces of evidence for the future pandemic wave preparedness in the UK [17]. Because of its representative sampling across households this study is recognised to have a strong power to capture asymptomatic infections which might be missed out by symptomatic testing scheme in the early pandemic [18]. However, this study started collecting samples from April of 2020 and then reporting the estimates of incidence from May of 2020 while the first death due to COVID-19 disease in the UK was documented in February 2020 [19]. This might suggest that the transmission of COVID-19 in the community might be much earlier than the survey and the survey might not be able to capture the early pandemic.

We developed a dynamic model to link multiple datasets including mortality, seroprevalence and virus PCR testing positivity ratio, and estimated the total exposure level across different regions of England after accounting for the antibody decay [16]. Here we examined and evaluated the model in the context of reconstructing the first COVID-19 pandemic from three perspectives: validation from ONS Infection Survey, relationship between model performance and data abundance and time-varying case reporting rate.

**Result**

**Validation from ONS Infection Survey**

Comparing the incidence of SARS-CoV-2 in England estimated by our model with those inferred by ONS COVID-19 Infection Survey (Figure 1), we found that our model could reveal the first but unobserved epidemic wave of COVID-19 in England from March 2020 to June 2020 additionally, with the second wave validated by the estimates from ONS Infection Survey. Further, we found our model results were highly consistent with those using SEIRS type compartmental models with time-varying force of infection [20, 21].

**Relationship between model performance and data abundance**

We then examined the relationship between model performance and data abundance - how estimates of exposure from our model change with more serological data points being added into the fitting procedure one by one over time (Figure 2). We found a highly robust pattern of exposure across different regions of England was estimated in general. Specifically, the model could only start estimating the interested quantities: exposure and two parameters (infection fatality ratio and antibody decaying rate) when at least two serological measurements from April to June 2020 in each region were given as inputs. However, these estimates were already highly consistent with those when more serological measurements were added although the credible bands were wider. The wide credible bands suggested a bigger uncertainty around the estimates when little information was available. When three serological measurements in each of region were included the estimates of exposure level were gradually stable at the results when all serological measurements were used. This might be attributed to the timing of these third serological measurements since then the seroprevalence in most regions started decreasing. With more and more serological measureemnts being added, the credible bands of estimates of exposure were gradually narrowing down.

**Time-varying case reporting rate**

While comparing the reported cases with the incidence estimated by our model (Figure 3), we found the confirmed cases in England only accounted for 9.1% (95%CrI (8.7%,9.8%) of cumulative exposure by the end of October 2020. Further, the relative size of two infection waves in England in 2020 estimated by our model, Spring wave from February to June and

Autumn wave from September to November, were in completely opposite direction with what reported by the confirmed cases. The case reporting rate relative to the total exposure was also dramatically different in these two-epidemic waves. If separating the two waves from the first of August 2020, we found during January 2020 to August 2020 the case reporting rate was only 4.3% (95%CrI (4.1%, 4.6%)) which increased to 43.7% (95%CrI (40.7%, 47.3%)) during August 2020 to October 2020, highlighting the dominate effect of testing effort in shaping the case curve in the early stage of a pandemic. Limited by the capacity of tests in the early stage of a pandemic, reported cases are almost always underestimating the total underlying infections in the population.

**Discussion**

The comparison exercise with ONS Infection Survey suggests it is a valuable early pandemic diagnostic tool, a dynamic model with a concise structure linking minimal dataset e.g. mortality, seroprevalence and PCR test positivity ratio, that is sufficiently powerful to reconstruct the early epidemic infection wave that had happened before any prevalence survey and any massive testing programme. The simple model structure could also avoid unnecessary complexity and structure-based uncertainty in a full dynamic model, for example, SEIR type compartmental models. The accuracy of model results would be improved as more and more information was available. In this paper we relooked into a dynamic model and argued that it is an important tool to quickly reconstruct the early pandemic transmission wave when data was sparse and large-scale infection survey was not available and feasible, for example, in Low- or Middle-Income Country (LMIC).

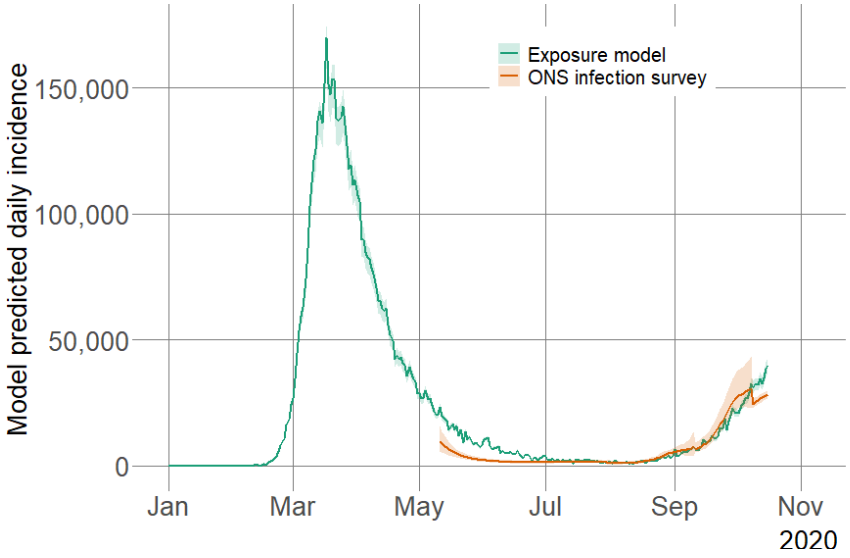


**Figure 1.** Comparison of model predicted daily incidence of SARS-CoV-2 in England. The green and orange lines show the predictions of median daily incidence by our model and ONS Infection Survey while the shaded areas correspond to the 95% CrI.


**Figure 2.** Comparison of estimates of exposure in seven regions of England when serological measurements are added one by one. The green and orange lines show the model

predictions of median exposure and seroprevalence, respectively, while the shaded areas correspond to the 95% CrI.

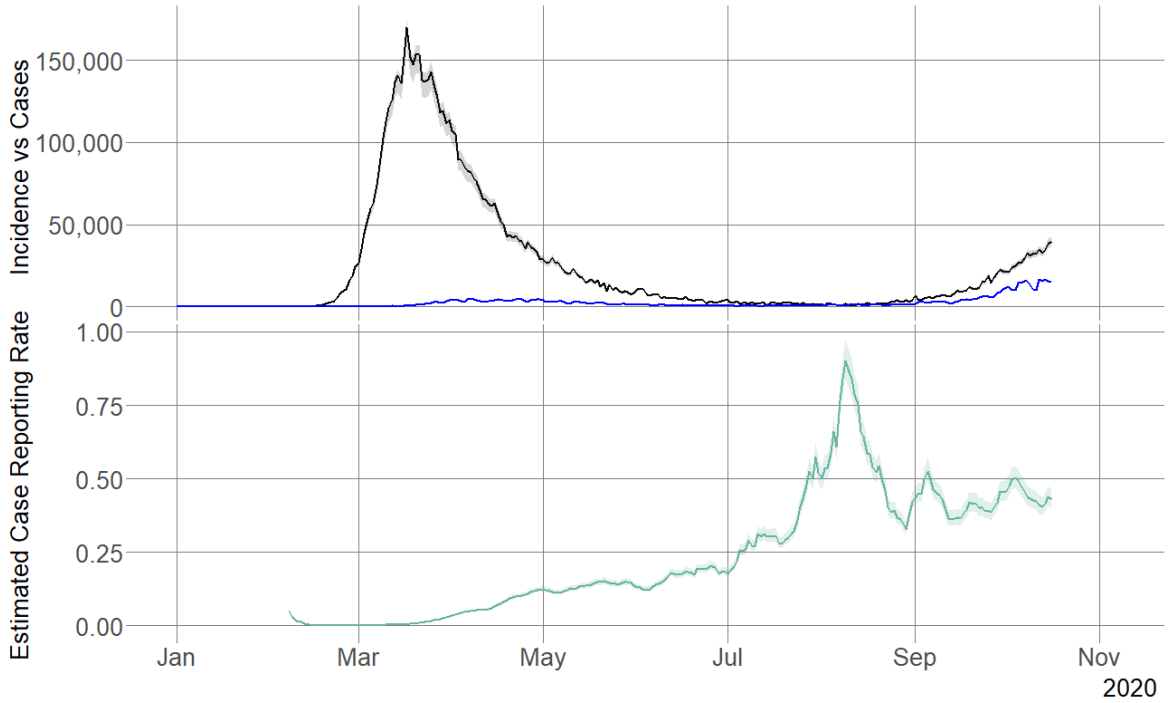


**Figure 3.** Comparison between estimates of daily incidence with reported cases of SARS-CoV-2 in England and estimated case reporting rate. In the top figure, the black lines show the predictions of median daily incidence by our model while the shaded areas correspond to the 95% CrI. The blue lines show the reported confirmed cases in England. In the bottom figure, the green lines show the estimates of median case reporting rate in England while the shared areas correspond to the 95% CrI.

## Materials and methods

### Data sources

We used publicly available epidemiological data to conduct the analysis, as described below

*ONS estimated incidence*

SARS-CoV-2 daily incidence in England in 2020 estimated by UK Coronavirus (COVID-19) Infection Survey were retrieved from the Office for National Statistics (ONS) [18] on March 17, 2023.

*Model estimated exposure*

Model estimated cumulative exposure to SARS-CoV-2 in seven regions of England were obtained from publication [16].

*7-day average of reported COVID-19 cases in England*

7-day average of reported COVID-19 daily cases in England in 2020 were retrieved from the UK government's official COVID-9 online dashboard [19] on March 17, 2023.

**Method**

We firstly calculated the incidence in England estimated by exposure model [16] by computing the difference of cumulative exposure in two successive days and adding together to the whole England:

$$I_i(t) = E_i(t+1) - E_i(t), t = 1,2, \dots, n, i = 1,2, \dots ,7$$

*Equation (1)*

$$I_{England}(t) = \sum_{i=1}^{7} I_i(t)$$

*Equation (2)*

Here, $E_i(t)$ is the daily exposure at region $i$ estimated by exposure model [16], $n$ is the total number of days, $i = 1, \dots 7$ represents London, Southwest, Southeast, Northeast, Northwest, East, Midland. $I_{England}(t)$ represents the daily incidence of England.

The estimated reporting ratio was calculated by

$$\bar{I}_{England}(t) = \frac{1}{7} \sum_{i=t-3}^{t+3} I_{England}(i), \qquad t = 4, 5, \dots, n-4$$

*Equation (3)*

$$r(t) = \frac{\bar{I}_{England}(t)}{C(t)}$$

*Equation (4)*

Here, $C$ is the 7-day average reported cases in England.

While testing the relationship between model performance and data abundance, we firstly obtained all the data and codes from paper [16] and rerun the model by adding the seroprevalence measurements one by one into the model.

**Acknowledgments:** The authors received no financial support for the research.

**Author contributions:**

L.J.W. and S.C conceived and designed the study. S.C. cleaned the data, S.C. and L.J.W. developed the methodology and conducted the formal analysis. S.C. and L.J.W. wrote the original manuscript. All authors reviewed and provided analytical input and approved the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

**Data and materials availability:** All codes and materials used in the analyses can be accessed at: https://github.com/SiyuChenOxf/Exposure_ONS-modelling. All parameter estimates and figures presented can be reproduced using the code provided. 

**References**

1. Aburto, J.M., et al., Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries. International journal of epidemiology, 2022. **51**(1): p. 63-74.
2. Ozili, P.K. and T. Arun, Spillover of COVID-19: impact on the Global Economy, in Managing Inflation and Supply Chain Disruptions in the Global Economy. 2023, IGI Global. p. 41-61.
3. Metcalf, C.J.E., D.H. Morris, and S.W. Park, Mathematical models to guide pandemic response. Science, 2020. **369**(6502): p. 368-369.
4. Aguas, R., et al., Modelling the COVID-19 pandemic in context: an international participatory approach. BMJ global health, 2020. **5**(12): p. e003126.
5. Pagel, C. and C.A. Yates, Role of mathematical modelling in future pandemic response policy. bmj, 2022. **378**.

6. Bollyky, T.J., et al., Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021. The Lancet, 2022. **399**(10334): p. 1489-1512.
7. Kennedy, B., et al., App-based COVID-19 syndromic surveillance and prediction of hospital admissions in COVID Symptom Study Sweden. Nature Communications, 2022. **13**(1): p. 2110.
8. Desjardins, M.R., Syndromic surveillance of COVID-19 using crowdsourced data. The Lancet Regional Health–Western Pacific, 2020. **4**.
9. Pullano, G., et al., Underdetection of cases of COVID-19 in France threatens epidemic control. Nature, 2021. **590**(7844): p. 134-139.
10. Rippinger, C., et al., Evaluation of undetected cases during the COVID-19 epidemic in Austria. BMC Infectious Diseases, 2021. **21**(1): p. 1-11.
11. Caldwell, J.M., et al., Understanding COVID-19 dynamics and the effects of interventions in the Philippines: A mathematical modelling study. The Lancet Regional Health-Western Pacific, 2021. **14**: p. 100211.
12. Trauer, J.M., et al., Understanding how Victoria, Australia gained control of its second COVID-19 wave. Nature Communications, 2021. **12**(1): p. 6266.
13. Phipps, S.J., R.Q. Grafton, and T. Kompas, Robust estimates of the true (population) infection rate for COVID-19: a backcasting approach. Royal Society Open Science, 2020. **7**(11): p. 200909.
14. GOV.UK, National COVID-19 surveillance reports https://www.gov.uk/government/publications/national-covid-19-surveillance-reports. 2020.
15. Arora, R.K., et al., SeroTracker: a global SARS-CoV-2 seroprevalence dashboard. The Lancet Infectious Diseases, 2021. **21**(4): p. e75-e76.
16. Chen, S., et al., Levels of SARS-CoV-2 population exposure are considerably higher than suggested by seroprevalence surveys. PLOS Computational Biology, 2021. **17**(9): p. e1009436.
17. Pouwels, K.B., et al., Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. The Lancet Public Health, 2021. **6**(1): p. e30-e38.
18. Office of National Statisics Coronavirus (COVID-19) Infection Survey UK,2023 https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/latest#strengths-and-limitations. 2023.
19. GOV.UK. Coronavirus (COVID-19) in the UK 2023; Available from: https://coronavirus.data.gov.uk/details/deaths?areaType=nation&areaName=England.
20. Knock, E.S., et al., Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. Science Translational Medicine, 2021. **13**(602): p. eabg4262.
21. Russell, T.W., et al., Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. BMC medicine, 2020. **18**(1): p. 1-9.