THE LANCET
Public Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Varsavsky T, Graham MS, Canas LS, et al. Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study. *Lancet Public Health* 2020; published online Dec 3. https://doi.org/10.1016/S2468-2667(20)30269-3.

Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study

Supplementary Appendix

Supplementary Methods

Incidence

We use Bayes rule to estimate the number of symptomatic COVID-19 cases in the general population,

$$P(C) = P(C|S)\, P(S) / P(S|C)$$

P(C) is the probability that a randomly selected person is infected with COVID-19 on a given day. P(S) is the probability of being newly sick according to the data entered in the app, defined as somebody who logs as healthy for at least nine days before reporting any of the symptoms asked about in the app (appendix pp 3-6). These newly sick users are invited to take a swab test. P(C|S) is the probability of a user testing COVID-19 positive given they are newly sick on the app. This is estimated as the percentage testing positive amongst the newly sick users that accept the test invite, with the assumption that the number of positive cases in this population is representative of the full population of newly sick users. This test positivity rate is considered as a binomial proportion, for which the 95% confidence intervals are derived using the Wilson score.The limits of this confidence interval are substituted in the same conditional probability equation to get the confidence interval for p(C) and in turn for incidence. P(S|C) is the probability of developing symptoms given that one has COVID-19. We set P(S|C) = 1 as we focus on the prediction of symptomatic cases. Note the calculation of p(C) reflects a simplifying assumption that a person becomes COVID-19 positive on the day they first report symptoms, in reality the start of infection will be before this.

Symptom-based classifier

The symptom-based classifier used in this work was developed and described in Menni et. al., Nature Medicine (2020). In brief, data from the COVID Symptom Study was used to generate a symptom-based classifier among a subset of 6,452 COVID-19-positive cases and 9,186 COVID-19-negative controls to generate a linear model for likelihood of COVID-19 infection. The model was divided into a train and testing set on a 80:20 ratio. Predictive features were chosen on the basis of a stepwise linear regression, to produce the following model:

$$\text{Predicted COVID-19 infection} = -(0{\cdot}01 \times \text{age}) + (0{\cdot}44 \times \text{male sex}) + (1{\cdot}75 \times \text{loss of smell/taste}) + (0{\cdot}31 \times \text{significant/severe persistent cough}) + (0{\cdot}49 \times \text{severe fatigue}) + (0{\cdot}39 \times \text{skipped meals}) - 1.32$$

where all symptoms are coded as 1 if the person self-reports the symptom and 0 if not. The sex feature is also binary, with 1 indicative of male participants and 0 representing females. The obtained value is then transformed into predicted probability using $\exp(x)/(1 + \exp(x))$

transformation followed by assigning cases of predicted COVID-19 for probabilities >0.5 and controls for probabilities <0.5.

In the UK test set, the prediction model had a sensitivity of 0.65 (95% CI 0.62–0.67), a specificity of 0.78 (95% CI0.76–0.80), an area under the curve (AUC) of the receiver operating characteristic curve (ROC) (that is, ROC-AUC) of 0.76 (95% CI 0.74–0.78), a positive predictive value of 0.69 (95% CI 0.66–0.71) and a negative predictive value of 0.75 (95% CI 0.73–0.77). A cross-validation ROC-AUC was 0.75 (95% CI 0.74–0.76) in the 15,638 UK users who were tested for SARS-CoV-2.

Model of recovery

The model of recovery, M_t, describes the proportion of users who become infected on day 0 who will recover on day t. We estimate the model from our data. We looked at users who reported a positive RT-PCR test and were also predicted positive from our symptom-based model. We defined onset as the first appearance of any symptom that occurred less than seven days before a positive test or being predicted positive. We defined recovery as either seven days of uninterrupted healthy reporting in the app or the date of a negative RT-PCR test, selecting the smallest value if both occurred. For our prevalence estimate, we truncate the model of recovery at 30 days, meaning we consider anyone who has suffered from COVID-19 for greater than 30 days as a long-term COVID-19 patient who is no longer infectious. This agrees with (Wajnberg et. al., 2020) where the authors suggest that although patients can test positive on RT-PCR beyond 28 days, they are unlikely to be infectious. This gives a probability distribution for the number of days it takes to recover from COVID-19.

Using the incidence estimates per NHS region, I_S, we can produce prevalence estimates per NHS region using the following dot-product:

$$P = I_{t:t+30} \bullet (1 - CDF(M))$$

Where $I_{t:t+30}$ is a vector of incidence estimates in the 30 days up to time t and CDF(M) is the cumulative distribution function of the model of recovery, M.

Supplementary Tables

Table S1. Questions asked in the COVID Symptom Study App.

The application is available to users in both private and non-private accommodation (e.g. care homes, student accommodation). Users are asked to enter their location on sign-up, and there is an option to update the location if a user moves.

Baseline questions (asked on sign-up)

About your work	
Are you a health care worker (including hospital, elderly care or in the community)?	- Yes, currently treat patients - Yes, do not currently treat patients - Yes, I currently interact with patients - Yes, but I do not currently interact with patients - No
Do you care for multiple people in the community, with direct contact with your patients?	- Yes/No
About you	
What year were you born?	
What sex were you assigned at birth?	- Female - Male - Prefer not to say - Intersex
What gender do you most identify with?	- Male - Female - Transgender - Do not identify as female, male or transgender - Prefer not to say
Which of the following best describes your ethnicity?	- Asian/Asian British - Indian, Pakistani, Bangladeshi, other - Black/Black British - Caribbean, African, other - Mixed race - White and Black/Black British - Mixed race - other - White - British, Irish, other - Chinese/Chinese British - Middle Eastern/Middle Eastern British - Arab, Turkish, other - Other ethnic group - Prefer not to say
What is your height?	
What is your weight?	
What is your postcode?	
In general, do you have any health problems that require you to stay at home?	- Yes/No

About your health	
Do you have heart disease?	- Yes/No
Do you have heart disease?	- Yes/No
Do you have lung disease or asthma?	- Yes/No
Do you smoke?	-Yes - Not currently - Never
Do you have kidney disease?	Yes/No
Are you living with cancer?	Yes/No

Follow up questions (asked daily)

Tests	
Have you had a test for COVID-19?	Yes/No
Did you test positive for COVID-19?	- Positve - Negative - Test failed - Waiting for result
How was this test performed?	- Nose swab - Throat swab - Spit tube - Blood sample
When was your test?	
Symptoms	
How do you feel right now?	- I feel as healthy as normal - I'm not feeling quite right
If not feeling quite right:	
Do you have a fever?	Yes/No
Do you feel chills or shivers (feel too cold)?	- Yes/No
If you are able to measure it, what is your temperature?	
Do you have a persistent cough (coughing a lot for more than an hour, or 3 or more	- Yes/No

coughing episodes in 24 hours)?	
Are you experiencing unusual fatigue?	- No - Mild fatigue - Severe fatigue - I struggle to get out of bed
Are you experiencing unusual shortness of breath?	- No - Yes. Mild symptoms - slight shortness of breath during ordinary activity - Yes. Significant symptoms - breathing is comfortable only at rest - Yes. Severe symptoms - breathing is difficult even at rest"
Do you have a loss of smell/taste?	Yes/No
Do you have an unusually hoarse voice?	Yes/No
Are you feeling an unusual chest pain or tightness in your chest?	Yes/No
Do you have an unusual abdominal pain?	Yes/No
Are you experiencing diarrhoea?	Yes/No
Do you have a headache?	- Yes/No
Do you have any of the following symptoms: confusion, disorientation or drowsiness?	Yes/No
Do your eyes have any unusual eye-soreness or discomfort (e.g. light sensitivity, excessive tears, or pink/red eye)?	Yes/No
Have you been skipping meals?	Yes/No
Are you experiencing dizziness or light-headedness?	Yes/No
Do you have a sore throat?	Yes/No
Do you have unusual strong muscle pains?	Yes/No
Have you had raised, red, itchy welts on the skin or sudden swelling of the face or lips?	Yes/No
Have you had any red/purple sores or blisters on your feet, including your toes?	Yes/No

Are there other important symptoms you want to share with us?	Free text entry

Supplementary Figures

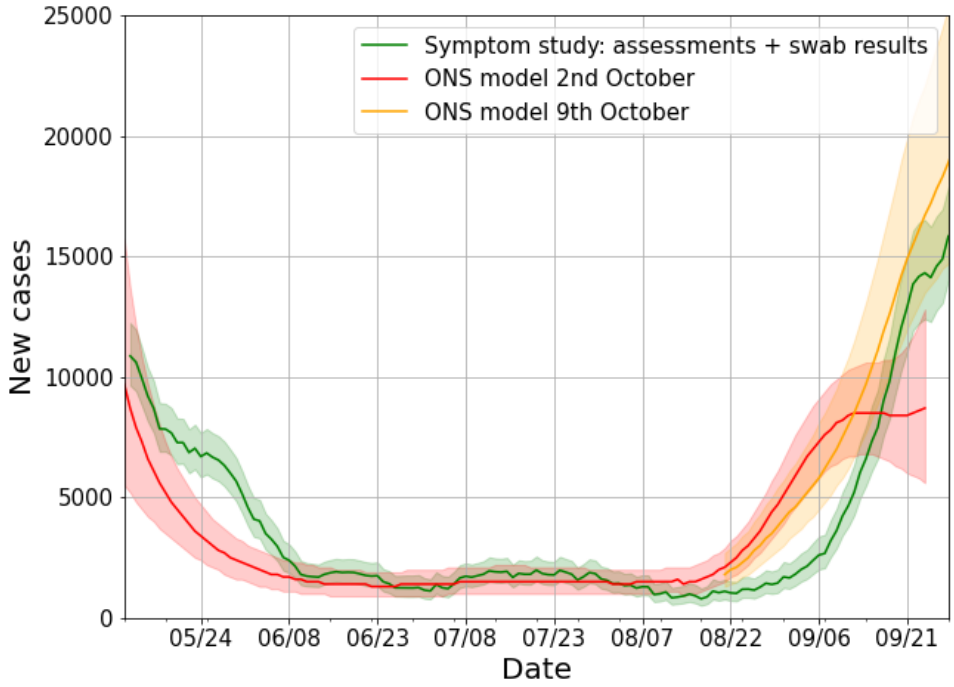

Figure S1. Comparison of incidence data with two ONS models.

Comparison of our incidence data with two ONS models, released on 2 October and 9 October. The model released on 2 October showed incidence levelling off in September, in disagreement with our released incidence estimates. By contrast the 9 October model showed a rapid increase in daily cases throughout September.

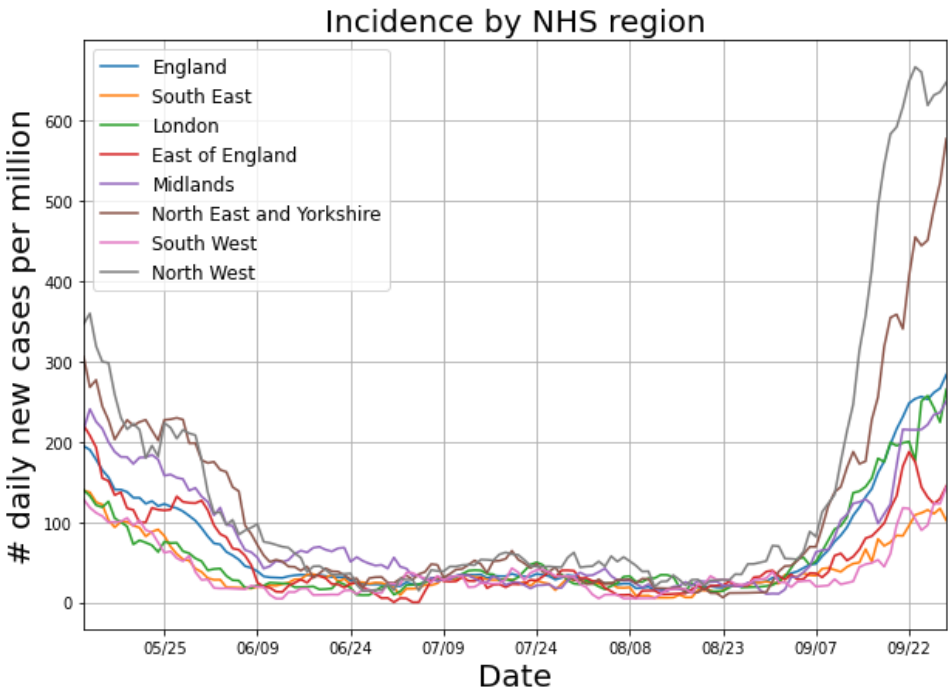

Figure S2. Estimated incidence for each NHS region in England.

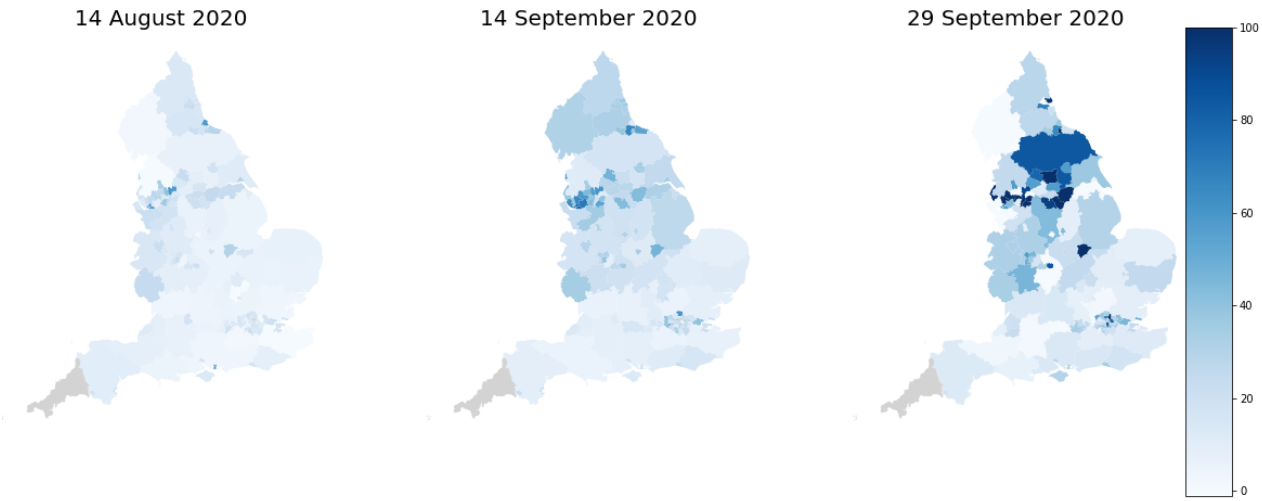

Figure S3. Map of daily cases per 100,000 people per UTLA.

Maps show daily cases per 100,000 people in each UTLA at three time-points, selected to show the lowest number of cases in mid-August, and the rapid increase through September.

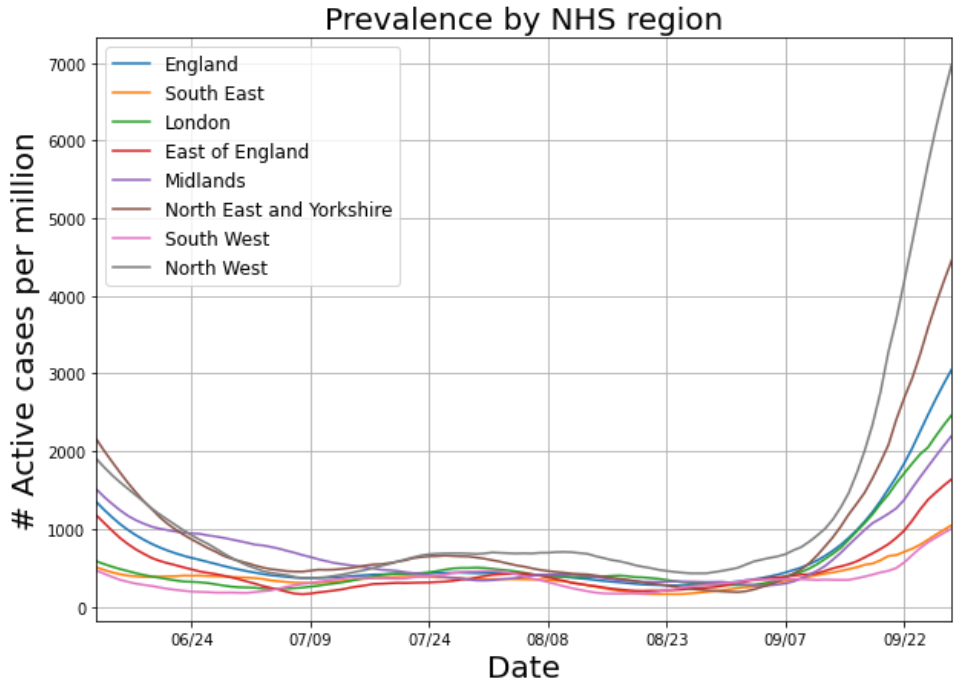

Figure S4. Estimated prevalence for each NHS region in England.

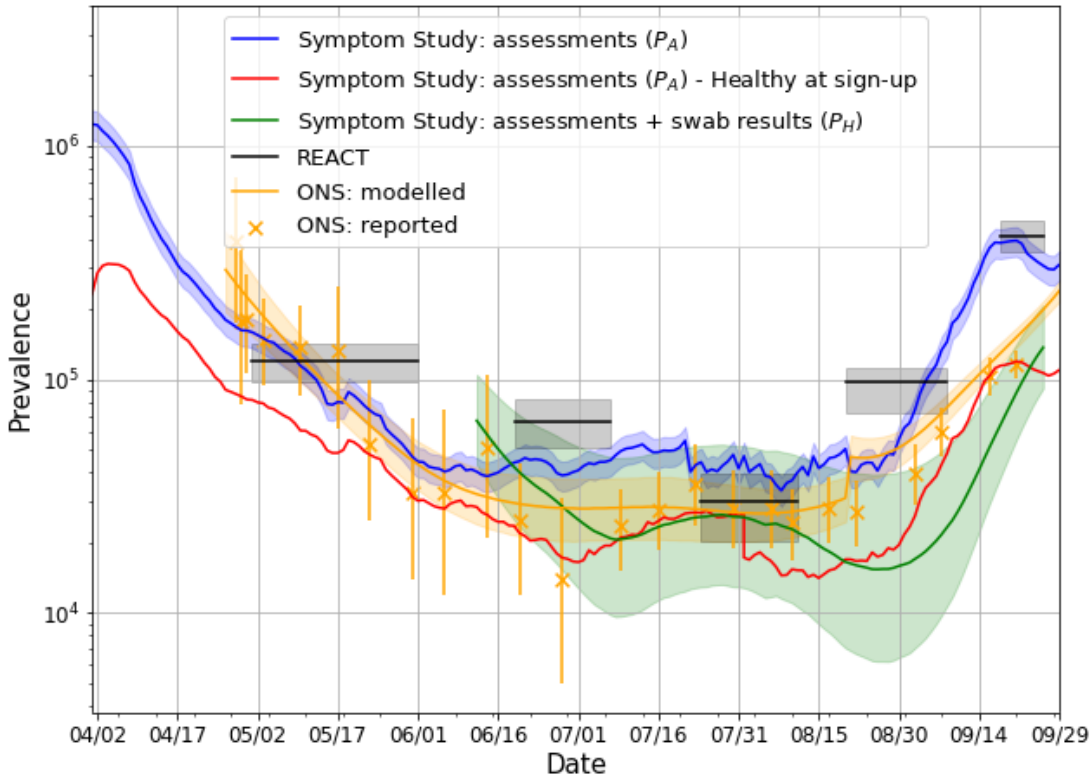

Figure S5. Estimates of prevalence, including an estimate obtained when all users who are sick upon sign-up are dropped from the analysis.

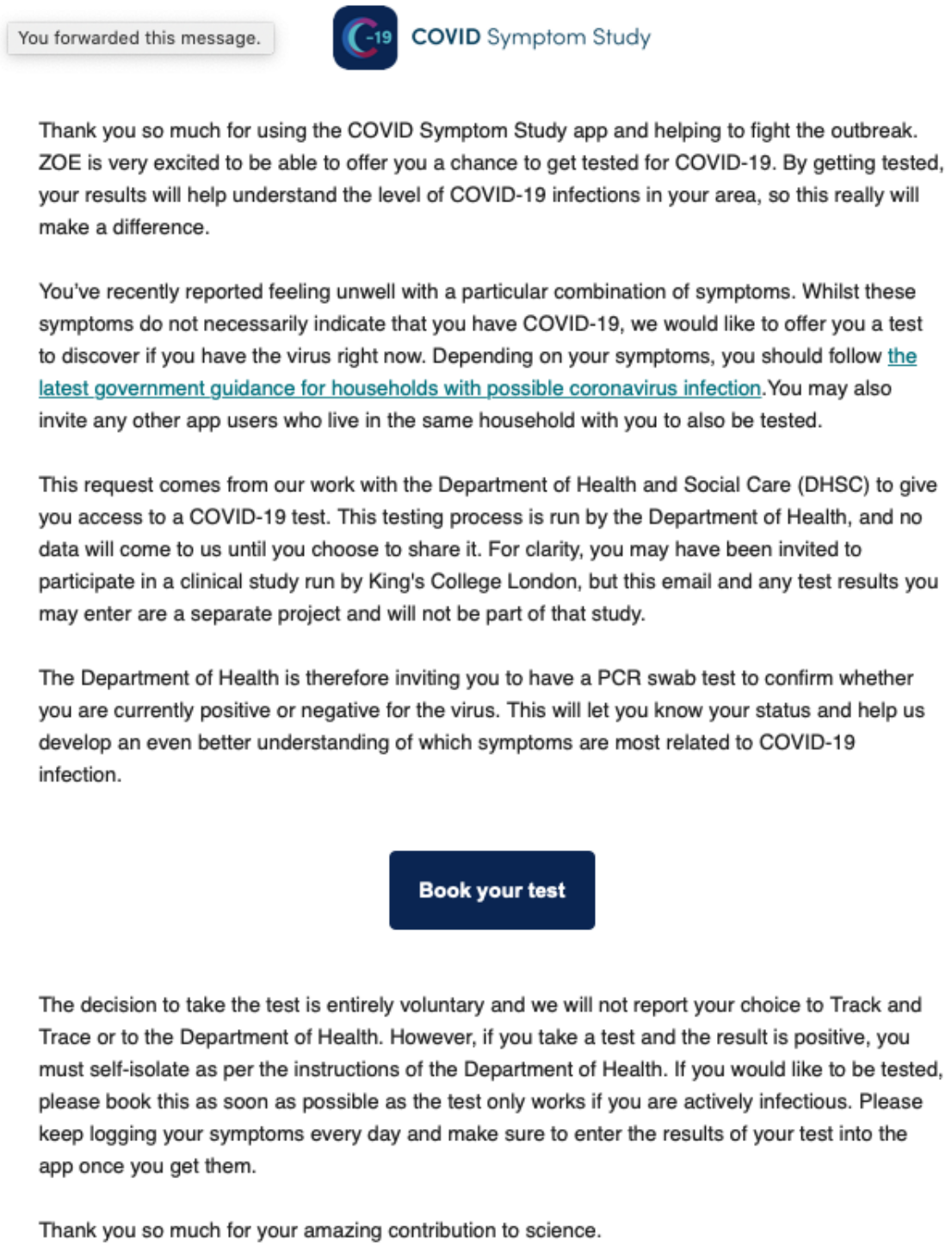

Figure S6. The testing invite sent to app users.

Figure S7. Estimates of R(t) for each NHS region in England.

Estimated R(t), for NHS regions in England between 24 June and 28 September, with 95% credible intervals shown. UK government estimates published every 7-12 days from 12 June.