Estimating the proportion of SARS-CoV-2 infections ascertained through diagnostic testing

Ewan Colman[*1], **Jessica Enright**[2], **Gavrila A. Puspitarani**[1], **and Rowland R. Kao**[†1]

[1]Royal (Dick) School of Veterinary Studies and the Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, UK
[2]School of Computing Science, University of Glasgow, Glasgow, UK

Background: The number of diagnosed SARS-CoV-2 infections in the population is one of the most commonly reported measures of epidemic severity. One issue with this metric is that it represents only a fraction of the total number of infections; many infected individuals do not receive a diagnostic test, and a proportion of those who do get tested will receive a false negative result. The proportion of cases ascertained is generally unknown and expected to vary depending on natural factors and changes in test-seeking behaviour.

Methods: We model the relationship between daily reported cases and population prevalence of SARS-CoV-2. The model is applied to surveillance and diagnostic test data from the United Kingdom to estimate the proportion of infections that have been ascertained with a positive diagnostic test. We estimate the number of infections and compare this to mortality data to obtain the infection fatality rate.

Results: Before November 2020, we estimate that approximately 20% of SARS-CoV-2 infections in the UK were ascertained with a positive test with results varying by time and region. Higher case ascertainment was observed during the period dominated by the the Alpha variant compared to the wild-type variants in all age categories and in 9 out of 12 regions of the UK. Ascertainment for the Delta variant was higher than for Alpha in 6 out of 7 age bands and 8 regions. Case ascertainment was higher in adults than in children. The infection fatality rate increased during the increase in the Alpha variant and declined in line with the distribution of vaccines.

Conclusions: Between September 2020 and October 2021 the majority of people infected with SARS-CoV-2 in the UK did not receive a positive diagnostic test result. Ascertainment of cases was found to be lowest in children and was found to increase over different phases of the pandemic.

Testing for SARS-CoV-2 in the UK aims to accomplish two things - first, to rapidly confirm suspected cases of COVID-19 disease via symptomatic testing in order to contain outbreak clusters, and second, to establish the overall burden of infection by taking a random sample of the population. Since not all infected individuals receive a test, and some of those who do will receive a false negative result, the number of positive diagnostic tests provides a *lower* estimate of the number of people exposed to the virus [1]. In contrast, random testing can provide an unbiased estimate of prevalence, but is an inefficient way to rapidly identify infection clusters, and may also have biases depending on the extent to which a positive test indicates the true infection status of the individual.

Both types of data are available in the UK: the number of positive tests from people with suspected infection are published daily on the UK government dashboard [2], and the Office for National Statistics COVID-19 Infection Survey (CIS) regularly publishes estimates of the population prevalence based on unbiased sampling [3]. The existence of these sources creates an opportunity to answer an important question: *what proportion of all infections are being reported through diagnostic testing?* Knowing this can help to estimate true incidence rates, a quantity central to understanding how the virus is spreading, and determine the infection fatality rate (IFR) of the disease.

Here we describe how diagnostic case numbers can be used to model the proportion of the population testing positive. By calibrating this model against surveillance data, we estimate the case ascertainment rate, defined as the proportion of infections that were reported through diagnostic testing; the incidence, defined as the number of newly infected individuals each day; and the IFR. Our differs from previous work as we do not rely on prior assumptions about the IFR to estimate incidence [4,5]. Moreover, we estimate the IFR using a data set substantially larger than any that has previously been used [6].

The model incorporates the different types of test and differences caused by variants of SARS-CoV-2, which have been shown to result in higher severity and a different range of clinical symptoms [7,8]. New variants might have different pathological characteristics that could potentially affect the test-seeking behaviour of those infected, which we expect to directly affect case ascertainment. Examination of age related and regional variation in case ascertainment provides a novel way to consider these developments, and to enrich our understanding of the epidemiology of the virus.

1 Methods

1.1 Data We are primarily concerned with daily Pillar 1 and 2 case data [2], hereafter referred to as diagnostic test cases, which represent tests done in health care settings and the community, respectively. We use the following notation:

$$C_q(t) \;=\; \text{Pillar 1 and 2 cases from test type } q \text{ on day } t.$$

Here, q can be PCR or LFD. These counts come from lab-based PCR tests and lateral flow device (LFD) testing, as performed in many community settings [9]. We use data provided for the 12 regions of the UK (9 regions of England and 3 other nations), and the 19 5-year age bands, which we aggregate into 7 distinct age groups to be consistent with the CIS data. Since the age groups for the the Pillar

[*]ecolman@ed.ac.uk
[†]rowland.kao@ed.ac.uk

1 & 2 data are not perfectly aligned with those for the CIS data, we first distribute them into 1-year age brackets, assuming an equal distribution of cases within each bracket, before re-aggregating.

At the time of writing, the number of cases detected by test type were available for England, but not provided at the regional level or for different age bands. We therefore approximate the proportion of cases that come from each test type by partitioning the total case numbers according to the the proportion calculated at the national scale.

From the CIS [3], we use the reported percentage of people testing positive for coronavirus, and apply it to the population to give

$$M(t) \;=\; \text{number of test-positive people} \tag{1}$$

where t is the midpoint of the time interval that the data represent. These estimates are provided for the 9 regions of England and 7 age groupings and represent data collected over 14-day intervals. We take the 7th day as the representative time point of this estimate. Data for the other nations is provided weekly and so we take it to represent the 4th day of the 7-day period. Population counts for the 9 regions and the 7 age categories were compiled from CIS data [10]. CIS provides the rate and the upper and lower bounds of the 95% credible interval.

The CIS provides an estimate of the proportion of tests that achieve different testing targets using the TaqPath test [11], which we use to estimate the proportion of infections from wild-type, Alpha, and Delta variants of the virus. We consider tests that are negative for the S target gene and positive for the two other targets, known as S-gene target failure (SGTF), to be a proxy for the Alpha variant. Tests that are positive for only one of the other targets may indicate the Alpha variant or any other lineage [12], and so we define

$$V(t) \;=\; \text{proportion of infections that have SGTF} \tag{2}$$

as the estimated proportion of tests that are negative on the S target (after discarding those that are negative on the S target and one other target). We assume the proportion to be 0 before November 1st 2020.

On March 1st 2021 the Alpha variant accounted for almost all infections in England, with $V(t) \approx 1$. After this time S gene positivity began to increase, mainly due to the Delta variant, which was dominant by July 2021 so for convenience we assume all S gene positive results after March 1st are caused by the Delta variant. We therefore consider three variant classes of interest: Alpha, that makes up a proportion p_A of all cases and $p_A(t) = V(t)$; Delta, that makes up a proportion $p_\Delta(t)$ of all cases with $p_\Delta(t) = 0$ if t is earlier than March 1st 2021, and $1-V(t)$ otherwise; and the wild-type, that makes up a proportion $p_w(t) = 1-V(t)$ if t is earlier than March 1st 2021.

Mortality data is given for England and Wales by each year of age, we group the ages as before and consider only deaths with an underlying cause of death as COVID-19 [13].

1.2 Modelling the time from exposure to the time of positive test We define two random variables, x and t^+, representing the time an individual was exposed and the time they received a positive test, respectively. Assuming daily time steps, we express the probability that an individual who received a positive test from a sample taken at time T was first exposed to the virus at time X,

$$P(\, x = X \;\&\; t^+ = T \,) = \frac{P(\, x = X \;\&\; t^+ = T \,)}{\sum_{i \geq 0} P(\, x = T - i \;\&\; t^+ = T \,)}. \tag{3}$$

The joint probability distribution $P(\, x = X \;\&\; t^+ = T \,)$ can be pieced together from various sources by considering the sequence of events that result in an individual testing positive.

First, we consider the time the individual was exposed to the virus and acquired the infection. We denote the prior probability that the infection was acquired at time X by $B(X)$. Next, we consider the time between exposure and the time that they received a test. For symptomatic cases we assume that the test occurs shortly after symptom onset, i.e. the time since exposure $\tau = T - X$ is equal to the sum of the incubation period and a delay parameter δ_k that we assume to be a fixed quantity; $k \in \{\text{PCR}, \text{LFD}\}$ represents the type of test being performed. The probability of a test on day T is thus $R(T - \delta_k - X)$ where $R(i)$ is the probability that the duration of the incubation period is i, which we assume to be Log-normal with a mean of 5.5 days and dispersion parameter 1.52 [14], and is broadly consistent with similar studies [15]. To get a probability distribution expressing the length of the incubation period in discrete days, we integrate over consecutive intervals of length 1.

Once the individual has acquired the infection and has had a test, the test must be positive to become a recorded infection. The probability of testing positive varies as a function, $S_k(\tau)$, of the time since exposure $\tau = T - X$. We use the functions provided by Hellewell et al. [16] and shown in Fig. 1. The PCR curve is similar in shape to the shedding profile found in other studies [17–20] with viral load typically peaking at day 3-5 and persisting for a mean duration of 17 days [21]. Variation is associated with severity of illness but not age or sex [22–24]. Studies that look for a difference between asymptomatic and symptomatic infections do not report consistent results [18, 25]. While one study with a small sample found that the Alpha variant had a longer viral course than the wild type [26], studies generally show that shedding profiles do not differ significantly between variants [27, 28]. In contrast, vaccination has been shown to reduce incidence of high shedding rates and duration of shedding [28–31], and so we focus our analysis on the period prior to the vaccination campaign.

We express the probability that an infected individual was exposed on day X and tested positive on day T by multiplying the probabilities mentioned above,

$$P(\, x = X \;\&\; t^+ = T \,) = B(X) R(T - X - \delta_k) S_k(T - X). \tag{4}$$

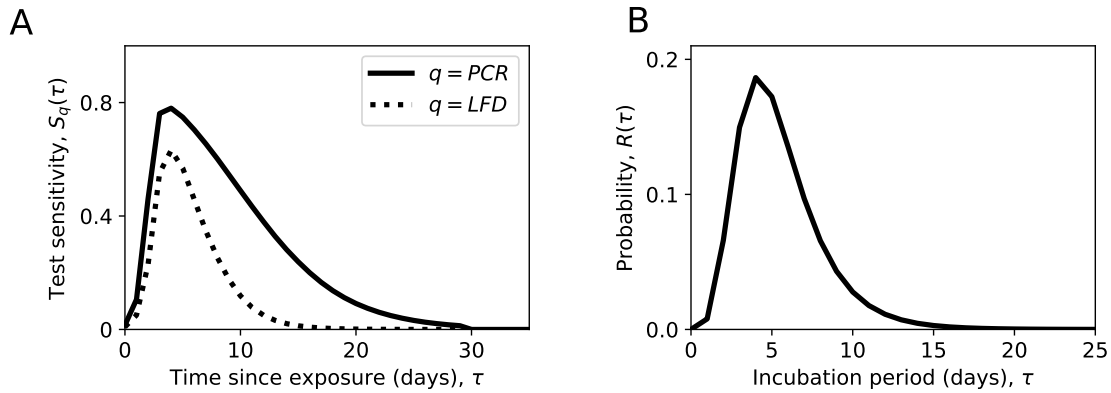

Figure 1: (A) The test sensitivity as a function of time from Hellewel et al. [16]. The function, $S(\tau)$ gives the probability that a PCR test will be positive if performed on an infected person τ days after exposure. (B) The incubation period probability distribution from Lauer et al. [14], shown here are the probability of symptom onset on each day since exposure.

If we assume an uninformative prior probability, $B(X)$, of exposure on day X, then substituting Eq. (4) into Eq. (3) gives

$$P_q(\ x = X \mid t^+ = T\ ) = \frac{R(T - X - \delta_q)S_q(T - X)}{\sum_{i\geq 0} R(i - \delta_q)S_q(i)}. \quad (5)$$

Substituting $\tau = T - X$, and using the notation $P_q(\tau) = P_q(\ x = T - \tau \mid t^+ = T\ )$, we have

$$P_q(\tau) = \frac{R(\tau - \delta_q)S_q(\tau)}{\sum_{i\geq 0} R(i - \delta_q)S_q(i)}, \quad (6)$$

giving the probability distribution of time between exposure and test for the set of ascertained cases corresponding to the test types $q \in \{\mathrm{PCR}, \mathrm{LFD}\}$.

1.3 Estimating the ascertainment rate We define the *ascertainment rate* as the proportion of SARS-CoV-2 infections that result in a positive PCR or LFD test and are recorded in the Pillar 1 & 2 case data. We introduce the time-dependent ascertainment rate $\boldsymbol{\theta}$, a vector whose xth element, θ_x, is the proportion of infections that occurred at time x that get reported through diagnostic testing at any subsequent time. We also consider the incidence, $I(x)$, defined as the number of newly acquired infections on day x.

The number of ascertained cases that were exposed at time x can be expressed in two ways: first, by multiplying the incidence by the ascertainment rate, and second by expressing the number of reported cases that were exposed on day x as a function of the daily case counts. Equating the two gives

$$\theta_x I(x) = \sum_q \sum_{\tau\geq 0} C_q(x + \tau)P_q(\tau). \quad (7)$$

We estimate the number of individuals in the population who would test positive (by PCR) on day t, if tested, by summing over all infections times and weighting by the the probability that each one is test-positive on day t

$$\hat{M}(t) = \sum_{k\geq 0} I(t - k)S_{PCR}(k). \quad (8)$$

Combining with Eq. (7) we can express this as a function of time and the unknown vector of parameters $\boldsymbol{\theta}$

$$\hat{M}(t; \boldsymbol{\theta}) = \sum_{k\geq 0}(S_{PCR}(k)/\theta_{t-k}) \sum_q \sum_{\tau\geq 0} C_q(t - k + \tau)P_c(\tau). \quad (9)$$

We estimate ascertainment by finding the vector $\boldsymbol{\theta}$ that minimizes the difference between the estimated and observed values,

$$\boldsymbol{\theta} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{t\in\mathcal{T}} \left| M(t) - \tilde{M}(t; \mathbf{x}) \right| \quad (10)$$

where $\mathcal{T}$ is the set of time points for which we have empirical estimates of prevalence. Eq. 10 combines the daily diagnostic case counts, the population positivity from surveillance, the incubation period distribution, and the time-dependent test sensitivity of PCR and LFD tests, to provide an estimate of the proportion of infections being reported at time t.

In practice, we estimate θ_x at weekly time points and use linear interpolation to create a daily time series. The solution to Eq. (10) is found numerically. The optimisation is made more efficient by inputting an initial guess based on an approximation to θ given by

$$\theta_t = M(t + 9)^{-1} \sum_q \sum_{k\geq 0} S_{PCR}(k) \sum_{j\geq 0} C_q(t - k + j)P_q(j). \quad (11)$$

Note that this equation uses the reported prevalence shifted forward by 9 days which is approximately the time since exposure of someone who received a positive test result through random surveillance. We estimate the credible interval for θ by substituting the upper and lower bounds of the credible interval for $M(t)$ into Eq (10).

1.4 Time-independent test-seeking rate For a measure of the ascertainment rate that does not change over time we take $\boldsymbol{\theta}$ to be a constant vector. Motivated by the possibility that the Alpha and Delta variants may have different pathological characteristics to the original lineage, we estimate

three ascertainment rates for each age band or region: the original rate, and the rate for cases caused by the Alpha variant and those caused by the Delta variant; θ_w, θ_A and θ_Δ, respectively. Recalling that $p_v(t)$ is the proportion of infections caused by variants of class v, we use a revised estimate of $M(t)$ that weights the contribution of each variant class by its proportion

$$\tilde{M}(t;\theta_w,\theta_A,\theta_\Delta) = \sum_{v\in\{w,A,\Delta\}} p_v(t)\hat{M}(t;\theta_v) \tag{12}$$

We can then estimate by taking the value that minimizes the absolute error between $M(t)$ and $\tilde{M}(t)$ taking only the time points $\mathcal{T}'$ up to March 1st 2021 when SGTF positive tests were associated with the Delta variant. Specifically, the time-independent ascertainment rates are estimated by numerically solving

$$\{\theta_w,\theta_A,\theta_\Delta\} = \underset{x,y,z}{\operatorname{argmin}} \sum_{t\in\mathcal{T}'} \left| M(t) - \tilde{M}(t;x,y,z) \right| \tag{13}$$

1.5 Infection fatality rate The infection fatality rate (IFR) is defined as the proportion of individuals infected who then die as a direct result of the infection. For a given monthly mortality figure, we count the corresponding number of infections from $I(x)$ summed over all corresponding exposure dates. We include a 21 day time from exposure to death (5 days to symptom onset and 16 days between onset and death) to be consistent with previous studies and the time between peaks in case and death data in England [32]. For example, the IFR for September is the number of recorded deaths in that month divided by the number of infections that occurred between August 10th and September 10th.

2 Results

The percentage of cases ascertained estimated with Eq. 10 varies by time, region, and age band (Fig. 2). These results are sensitive to variation in surveillance data, particularly when infection levels are low and there is less data to inform the estimate. For example, from the time that the infection survey began until September 2020, which is not shown in the figure, the results are highly variable and occasionally produce estimates of ascertainment that are above 100%. In general, when case rates are low we see an increase in variability due to the smaller sample size, whereas when case rates are relatively high the estimated ascertainment rate becomes more reliable.

Case ascertainment is related to the proportion of infections that lead to symptomatic infection. This is apparent from the low ascertainment rates observed in the lowest age categories, which are known to be less likely to develop symptoms [33]. There were notable increases after March 2021 in school age children, possibly indicating that the mass testing in that age category that coincided with school reopening caused a higher detection rate of asymptomatic infections. Similarly, since vaccination is effective at preventing infections from becoming symptomatic, the decreasing ascertainment rate seen in the two highest age bands from January to April 2021 may have resulted from vaccination in those groups.

Increases from April to June 2021 occur in every group, and appear to coincide with the rise in cases of the Delta variant. While this could imply that the Delta variant is more likely to cause symptomatic infection, it could also be the result of behavioural factors as restrictions to physical contact were being removed and lateral flow tests were being more widely used. There is similarity between the time series of age groups that are close to each other, whereas changes in the ascertainment rate in any given region appears to be independent of neighbouring regions (Fig. S3).

To compare regions, ages, and different phases of the pandemic, we consider three different variant classes: the wild-type that existed before the emergence of the Alpha variant, the Alpha variant, and the Delta variant, where we have used S-gene target failure to approximate the proportion of cases belonging to each class. Modelling a ascertainment rate that varies according to the variant, but is constant over time, provides a reasonable agreement between the modelled population prevalence and the value reported by the surveillance study, shown in Figure S1. The best-fit ascertainment rates are shown in Fig. 3. Differences between these variants may reflect varying symptomatic responses, or they may reflect other behavioural factors that have changed over time.

The ascertainment rate for the wild-type is consistently lower than the rate for the Alpha and Delta variants across all age bands and 9 out of 12 regions of the UK. The percentage of infections ascertained is generally lower than the percentage of infections that are symptomatic, estimated to be around 70% [34, 35], implying that a considerable number of symptomatic infections do not get ascertained. Ascertainment rates are lowest in the youngest age band and increases with age up to the 25 to 34 band. The ascertainment rate for the Delta variant was higher than Alpha in all but the oldest age band, however this trend is not visible when the same data are aggregated by region.

The trajectory of the epidemic seen in diagnostic case data is largely in agreement with that obtained from surveillance data from random samples. Divergence between these trajectories occurred at the time that the Alpha variant became dominant, and again at the the time that the Delta variant began to dominate. However, by assuming the fitted variant-dependent ascertainment rates from Fig. 3, we see agreement throughout the time period, shown in Fig. S1. Notable exceptions occur in January 2021 coinciding with the introduction of restrictions that caused infections to peak. In London the model overestimates the peak positivity rate and underestimates the subsequent data points as infections decrease.

We calculate the IFR for each month for the 4 oldest age groups (Figure 4). We have chosen not to show lower ages as the low numbers of deaths in these groups make the results highly variable do not provide a reliable estimate of the true IFR. Within the 4 age bands for which data are sufficent, the IFR increases with age. We also see considerable variation over time; the increasing trend Novemeber

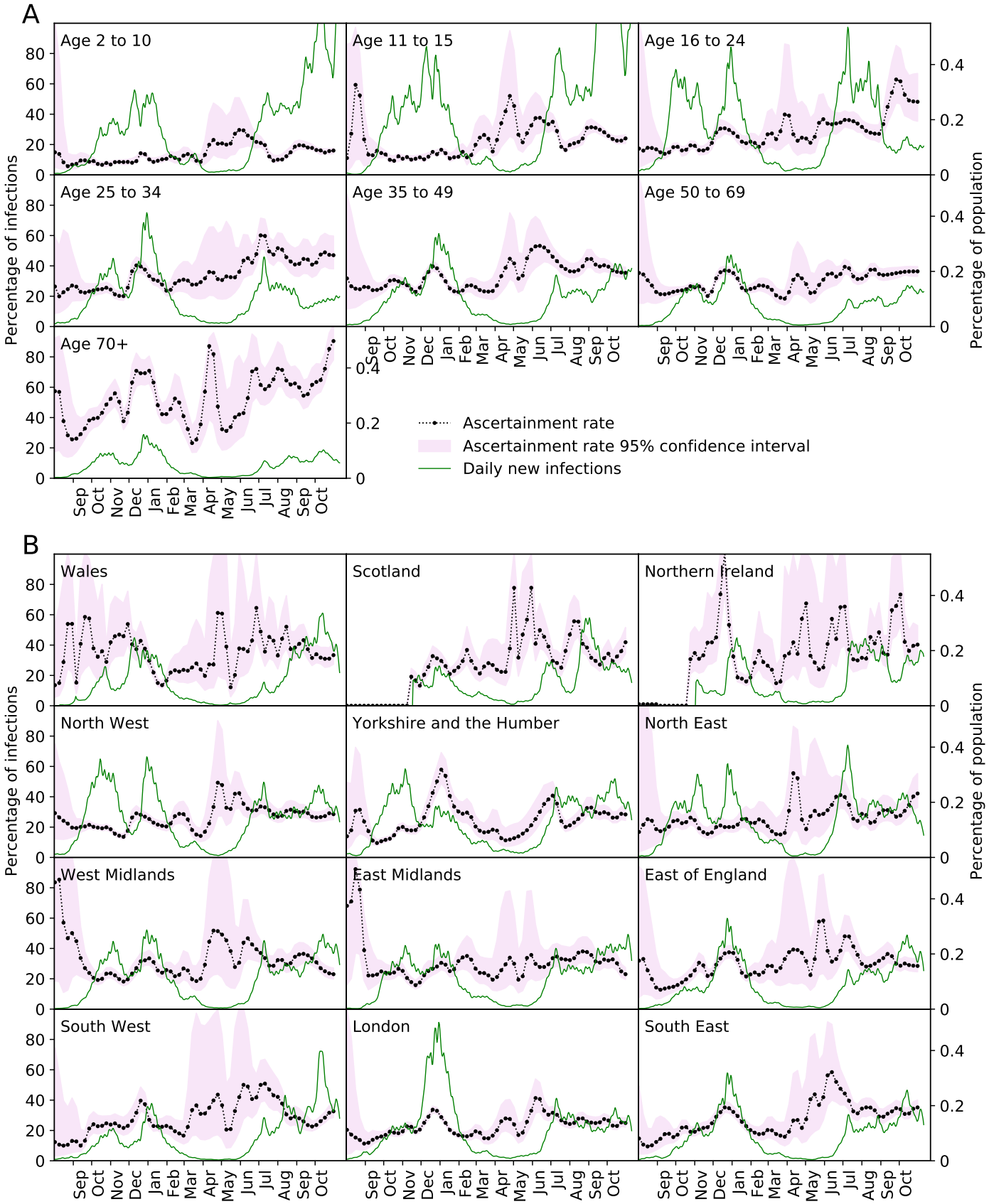

Figure 2: Ascertainment rate in the 7 age bands and 12 regions of the UK expressed as a percentage ($\theta(\mathbf{t})$ given by Eq. (10)). Confidence intervals, shown by the shaded area, are created by recomputing θ using the upper and lower bounds of the 95% confidence interval provided by the surveillance data as the rate $M(t)$. The incidence I is shown as a percentage of the population.

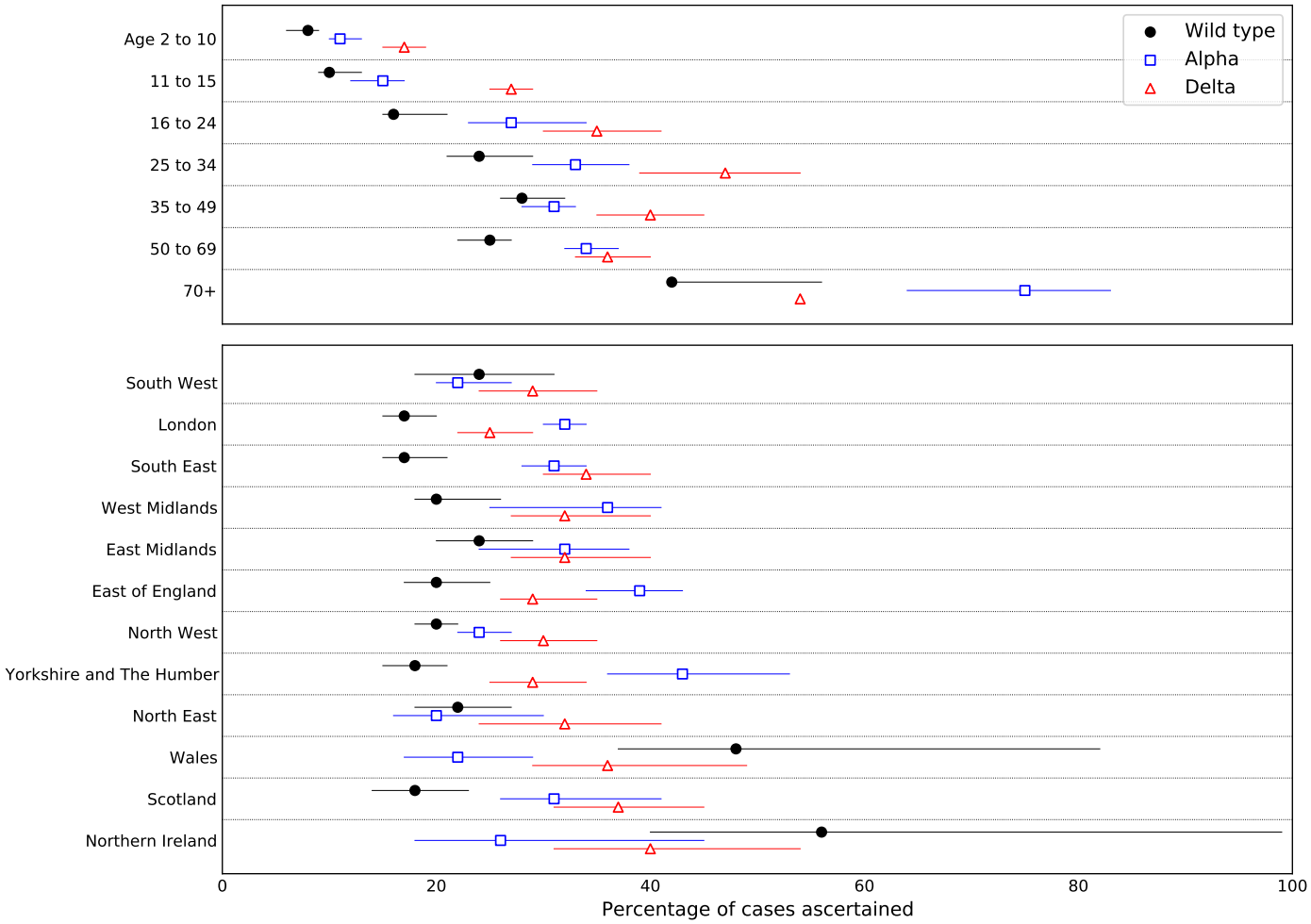

Figure 3: Estimates of time-independent ascertainment rates for the wild-type variants, θ_o, the Alpha variant, θ_A and Delta, θ_Δ. Confidence intervals, shown by the horizontal lines, are created by recomputing θ using the upper and lower bounds of the 95% confidence interval provided by the surveillance data as the rate $M(t)$.

2020 may be a combination of higher severity of the Alpha variant [36], the increased pressure on the healthcare system as hospitalisations increased, or a seasonal affect on immunity. The decrease in IFR coincides with the uptake of vaccines.

We tested the robustness of these results against reasonable changes in the assumptions of our model. Firstly, viral clearance may occur more rapidly in individuals who have been vaccinated [28]. While we cannot model this effect precisely, a model that makes liberal assumptions about vaccine effectiveness and its effect on the test-sensitivity profile gives results that are within a few percent (Supplementary analysis S1.1 and Fig. S2 at end of this document). A more substantial effect, however, is observed in older age bands. Furthermore, the IFR presented in Fig. 4 may be an overestimate during times when vaccine coverage is high. Fig. S4 shows the range of values that are plausible given the duration of viral shedding in vaccinated individuals.

The results presented are derived from a model that assumes the delay between symptom onset and receiving a PCR test is $\delta_{\text{PCR}} = 1$ day. While we believe this to be a reasonable assumption we do not have observational evidence to support it. Repeating the analysis with $\delta_{\text{PCR}} = 2$ yields a mean decrease of 0.18 percentage points with a standard deviation of 1.2 to the results reported in Fig.3, suggesting relatively minor sensitivity to this modelling decision. Some estimates of test sensitivity are higher than the maximum of S_{PCR} [37]. Repeating our analysis using an adjusted versions S_{PCR} and S_{LFD} that are linearly scaled up so that S_{PCR} peaks at 1, we find ascertainment rates increase by a mean of 8.0 percentage points with standard deviation of 3.6, suggesting that the assumed test sensitivity could substantially affect the outcome.

3 Discussion

The extensive efforts in the United Kingdom to monitor the COVID-19 epidemic have provided the opportunity to quantify a critically important parameter - the likelihood that an infected individual will get tested and receive a positive diagnosis. Here we compared the daily reported number of cases to an unbiased estimate of population prevalence to estimate the proportion of cases that are ascertained through community testing. Case ascertainment was revealed to be related to age, with the youngest age

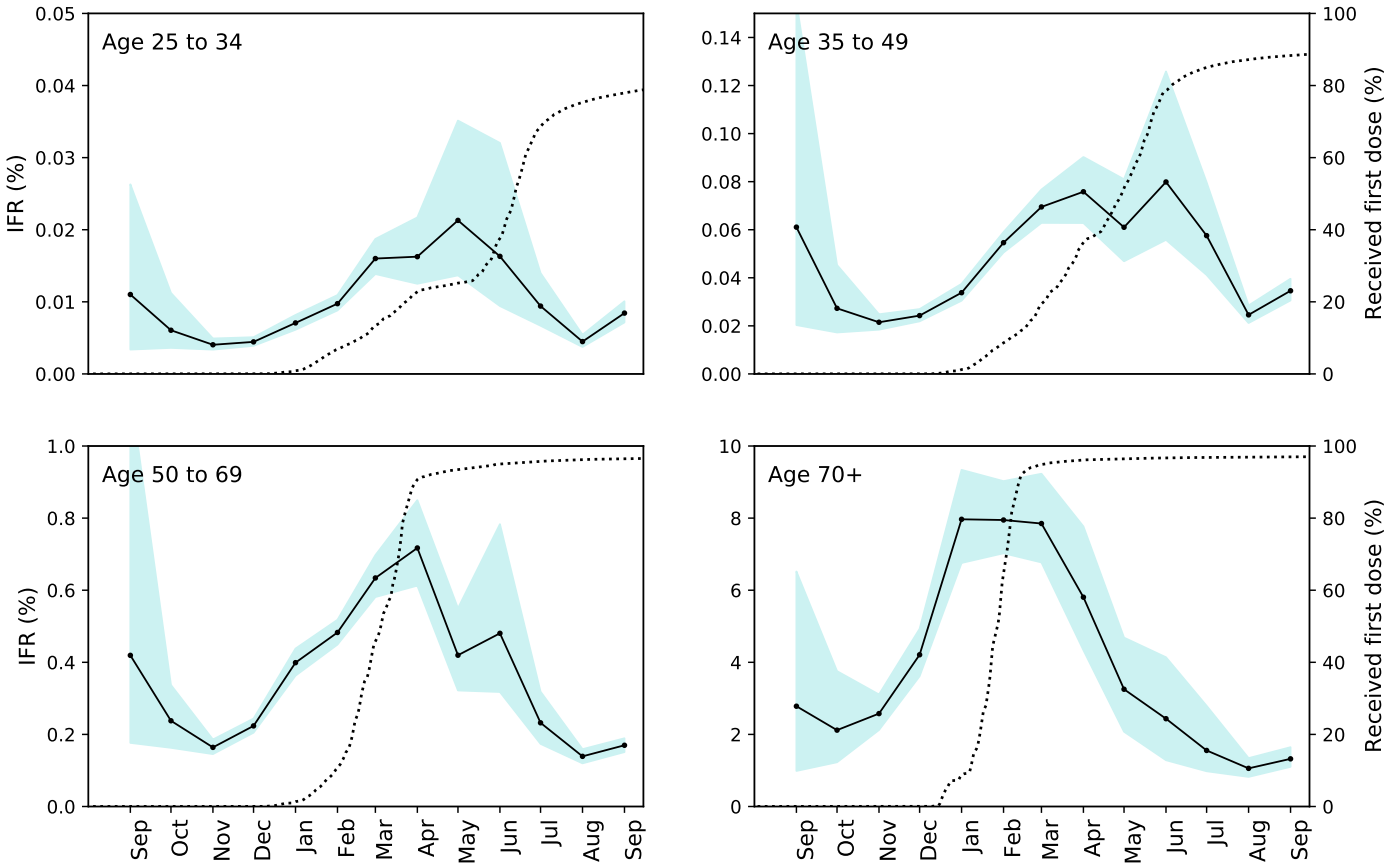

Figure 4: Infection fatality rate (IFR). The estimated percentage of infections that cause mortality. The shaded region shows the 95% confidence interval computed by using $I(x)$ computed from the upper and lower estimates of prevalence given by the surveillance data. Dashed lines show the population in the respective age group that had received at least one dose of a COVID-19 vaccine.

bands having lowest case ascertainment, and infections related to the Alpha and Delta variants more likely to be ascertained compared to the strains that were circulating in earlier months. Following this result, we obtained a time-dependent IFR that showed substantial variation, increasing into winter and declining with the distribution of vaccines.

The challenge when comparing the trend seen in random survey data to that seen in reported community cases is that the former is a measure of prevalence and the latter a measure of incidence. Our analysis resolves this by modelling the relationship between the two. Moreover, changes the case ascertainment rate at over time, whether sudden or gradual, indicate changes in the relationships between the two datastreams, and may be useful information for policy evaluation.

Our results rest on a number of simplifying assumptions. We have applied a model that assumes all individuals experience similar viral dynamics once infected, and the time for between exposure and receiving a test follows the same distribution regardless of age, location, or vaccination status (this is considered in the supplement). Moreover, due of limitations of the data, we have assumed that the proportion of positives that come from LFD tests (and not PCR) is equal across all subgroups. In reality the proportion is likely to be higher in younger age groups as these tests have been widely used in schools. Since LFD tests are expected to be used for asymptomatic screening the time between exposure and receiving an LFD test may be shorter than we have assumed.

Variation in case ascertainment may result from differences in clinical presentation, public perception, availability of testing, or other reasons. It is not possible to determine the extent to which the higher rates of ascertainment for the Alpha and Delta variants estimated here are caused by changes in symptomatic response or by external factors that may alter the propensity of the individual to seek a test. However, rises in case ascertainment across different regions of the UK in the last 2 months of 2020 align with increases in the Alpha variant, despite occurring at different times in different regions, suggesting that it is at least in part due to inherent immune-response related effects.

After accounting for the effects of the Alpha and Delta variants on the ascertainment rate, we see that the two data sources are largely in agreement. That fitting time-independent ascertainment rates to the wild-type, Alpha, and Delta variants leads to a consistently good fit across the entire duration studied is a strong indicator that variation in ascertainment on a short time-scale are more likely to be due to management and behavioural factors, than they are to any inherent differences in test performance.

These methods provide a combined estimate of infec-

tion incidence, offering an alternative to methods that use serological data [38]. This allows for more accurate representation of key metrics related to epidemic control such as the reproduction rate, generation time, case doubling rates, and infection fatality rates. Our analysis revealed considerable variability in the IFR that goes beyond that expected from age and vaccination status alone. The fourfold increase in IFR beginning in Novemeber 2020 suggests that multiple factors contribute to the risk of death from infection and therefore there may be multiple ways to minimize mortality in future winter seasons.

Finally, we highlight that the methods here may be translated to a variety of current and future epidemiological studies. As the COVID-19 pandemic has expanded the scale and scope of health surveillance data to an unprecedented level, the methods required to parse such data, and create interpretations useful to inform decision makers and increase public awareness, need also to adapt. The methods presented here are novel, although built from established mathematical concepts, and this reflects constant requirement to re-evaluate and refresh the set of mathematical and statistical tools available to analysts as the landscape of public health continues to evolve.

Acknowledgments To do.

Code availability

`https://github.com/EwanColman/Estimating_SARS-CoV-2_case_ascertainment`

References

[1] Timothy W Russell, Nick Golding, Joel Hellewell, Sam Abbott, Lawrence Wright, Carl AB Pearson, Kevin van Zandvoort, Christopher I Jarvis, Hamish Gibbs, Yang Liu, et al. Reconstructing the early global dynamics of under-ascertained covid-19 cases and infections. *BMC medicine*, 18(1):1–9, 2020.

[2] Uk coronavirus dashboard: cases. `https://coronavirus.data.gov.uk/details/cases`.

[3] Office for national statistics infection survey. `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/latest`.

[4] Jungsik Noh and Gaudenz Danuser. Estimation of the fraction of covid-19 infected people in u.s. states and countries worldwide. *PLOS ONE*, 16(2):1–10, 02 2021.

[5] Heather Reese, A Danielle Iuliano, Neha N Patel, Shikha Garg, Lindsay Kim, Benjamin J Silk, Aron J Hall, Alicia Fry, and Carrie Reed. Estimated Incidence of Coronavirus Disease 2019 (COVID-19) Illness and Hospitalization—United States, February–September 2020. *Clinical Infectious Diseases*, 72(12):e1010–e1017, 11 2020.

[6] Gideon Meyerowitz-Katz and Lea Merone. A systematic review and meta-analysis of published research data on covid-19 infection-fatality rates. *International Journal of Infectious Diseases*, 2020.

[7] Pengfei Wang, Lihong Liu, Sho Iketani, Yang Luo, Yicheng Guo, Maple Wang, Jian Yu, Baoshan Zhang, Peter D. Kwong, Barney S. Graham, John R. Mascola, Jennifer Y. Chang, Michael T. Yin, Magdalena Sobieszczyk, Christos A. Kyratsous, Lawrence Shapiro, Zizhang Sheng, Manoj S. Nair, Yaoxing Huang, and David D. Ho. Increased resistance of sars-cov-2 variants b.1.351 and b.1.1.7 to antibody neutralization. *bioRxiv*, 2021.

[8] Sean Wei Xiang Ong, Calvin J Chiew, Li Wei Ang, Tze-Minn Mak, Lin Cui, Matthias Paul HS Toh, Yi Ding Lim, Pei Hua Lee, Tau Hong Lee, Po Ying Chia, et al. Clinical and virological features of sars-cov-2 variants of concern: A retrospective cohort study comparing b. 1.1. 7 (alpha), b. 1.315 (beta), and b. 1.617. 2 (delta). 2021.

[9] Uk coronavirus dashboard: tests. `https://coronavirus.data.gov.uk/details/tests`.

[10] Office for national statistics population estimates. `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates`.

[11] Investigation of novel sars-cov-2 variant. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/950823/Variant_of_Concern_VOC_202012_01_Technical_Briefing_3_-_England.pdf`.

[12] New-variant compatibility in the ons infection survey. `https://theo.io/post/2021-01-22-ons-data/`. Accessed on 01/02/2021.

[13] Office for national statistics: Single year of age and average age of death of people whose death was due to or involved covid-19. `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/`.

[14] 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.

[15] Hualei Xin, Jessica Y Wong, Caitriona Murphy, Amy Yeung, Sheikh Taslim Ali, Peng Wu, and Benjamin J Cowling. The Incubation Period Distribution of Coronavirus Disease 2019: A Systematic Review and Meta-analysis. *Clinical Infectious Diseases*, 06 2021. ciab501.

[16] Joel Hellewell, Timothy W. Russell, Rupert Beale, Gavin Kelly, Catherine Houlihan, Eleni Nastouli, and Adam J. Kucharski. Estimating the effectiveness of routine asymptomatic pcr testing at different frequencies for the detection of sars-cov-2 infections. *medRxiv*, 2020.

[17] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature medicine*, 26(5):672–675, 2020.

[18] Quan-Xin Long, Xiao-Jun Tang, Qiu-Lin Shi, Qin Li, Hai-Jun Deng, Jun Yuan, Jie-Li Hu, Wei Xu, Yong Zhang, Fa-Jin Lv, et al. Clinical and immunological assessment of asymptomatic sars-cov-2 infections. *Nature medicine*, 26(8):1200–1204, 2020.

[19] Roman Wölfel, Victor M Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A Müller, Daniela Niemeyer, Terry C Jones, Patrick Vollmar, Camilla Rothe, et al. Virological assessment of hospitalized patients with covid-2019. *Nature*, 581(7809):465–469, 2020.

[20] Rebecca L Smith, Laura L Gibson, Pamela P Martinez, Ruian Ke, Agha Mirza, Madison Conte, Nicholas Gallagher, Abigail Conte, Leyi Wang, Richard Fredrickson, Darci C Edmonson, Melinda E Baughman, Karen K Chiu, Hannah Choi, Tor W Jensen, Kevin R Scardina, Shannon Bradley, Stacy L Gloss, Crystal Reinhart, Jagadeesh Yedetore, Alyssa N Owens, John Broach, Bruce Barton, Peter Lazar, Darcy Henness, Todd Young, Alastair Dunnett, Matthew L Robinson, Heba H Mostafa, Andrew Pekosz, Yukari C Manabe, William J Heetderks, David D McManus, and Christopher B Brooke. Longitudinal Assessment of Diagnostic Test Performance Over the Course of Acute SARS-CoV-2 Infection. *The Journal of Infectious Diseases*, 224(6):976–982, 06 2021.

[21] Muge Cevik, Matthew Tate, Ollie Lloyd, Alberto Enrico Maraolo, Jenna Schafers, and Antonia Ho. Sars-cov-2, sars-cov-1 and mers-cov viral load dynamics, duration of viral shedding and infectiousness: a living systematic review and meta-analysis. *SARS-CoV-1 and MERS-CoV Viral Load Dynamics, Duration of Viral Shedding and Infectiousness: A Living Systematic Review and Meta-Analysis*, 2020.

[22] Paul Z Chen, Niklas Bobrovitz, Zahra Premji, Marion Koopmans, David N Fisman, and Frank X Gu. Sars-cov-2 shedding dynamics across the respiratory tract, sex, and disease severity for adult and pediatric covid-19. *eLife*, 10:e70458, aug 2021.

[23] Lael M Yonker, Julie Boucau, James Regan, Manish C Choudhary, Madeleine D Burns, Nicola Young, Eva J Farkas, Jameson P Davis, Peter P Moschovis, T Bernard Kinane, Alessio Fasano, Anne M Neilan, Jonathan Z Li, and Amy K Barczak. Virologic features of SARS-CoV-2 infection in children. *The Journal of Infectious Diseases*, 10 2021. jiab509.

[24] Terry C. Jones, Barbara Mühlemann, Talitha Veith, Guido Biele, Marta Zuchowski, Jörg Hofmann, Angela Stein, Anke Edelmann, Victor Max Corman, and Christian Drosten. An analysis of sars-cov-2 viral load by patient age. *medRxiv*, 2020.

[25] Stephen M. Kissler, Joseph R. Fauver, Christina Mack, Scott W. Olesen, Caroline Tai, Kristin Y. Shiue, Chaney C. Kalinich, Sarah Jednak, Isabel M. Ott, Chantal B.F. Vogels, Jay Wohlgemuth, James Weisberger, John DiFiori, Deverick J. Anderson, Jimmie Mancell, David D. Ho, Nathan D. Grubaugh, and Yonatan H. Grad. Sars-cov-2 viral dynamics in acute infections. *medRxiv*, 2020.

[26] Stephen M Kissler, Joseph R Fauver, Christina Mack, Caroline Tai, Mallery Breban, Anne E Watkins, Radhika Samant, Deverick Anderson, David Ho, Nathan D Grubaugh, et al. Densely sampled viral trajectories suggest longer duration of acute infection with b. 1.1. 7 variant relative to non-b. 1.1. 7 sars-cov-2. *medRxiv*, 2021.

[27] Ruian Ke, Pamela P Martinez, Rebecca L Smith, Laura L Gibson, Agha Mirza, Madison Conte, Nicholas Gallagher, Chun Huai Luo, Junko Jarrett, Abigail Conte, et al. Daily sampling of early sars-cov-2 infection reveals substantial heterogeneity in infectiousness. *medRxiv*, 2021.

[28] Stephen M Kissler, Joseph R Fauver, Christina Mack, Caroline G Tai, Mallery I Breban, Anne E Watkins, Radhika M Samant, Deverick J Anderson, Jessica Metti, Gaurav Khullar, et al. Viral dynamics of sars-cov-2 variants in vaccinated and unvaccinated individuals. *medRxiv*, 2021.

[29] Emma Pritchard, Philippa C Matthews, Nicole Stoesser, David W Eyre, Owen Gethings, Karina-Doris Vihta, Joel Jones, Thomas House, Harper VanSteenHouse, Iain Bell, et al. Impact of vaccination on new sars-cov-2 infections in the united kingdom. *Nature Medicine*, pages 1–9, 2021.

[30] Ruian Ke, Pamela Martinez, Rebecca Lee Smith, Laura Gibson, Chad Achenbach, Sally McFall, Chao Qi, Joshua Jacob, Etienne Dembele, Camille Bundy, et al. Longitudinal analysis of sars-cov-2 vaccine breakthrough infections reveal limited infectious virus shedding and restricted tissue distribution. *medRxiv*, 2021.

[31] Michela Antonelli, Rose S Penfold, Jordi Merino, Carole H Sudre, Erika Molteni, Sarah Berry, Liane S Canas, Mark S Graham, Kerstin Klaser, Marc Modat, Benjamin Murray, Eric Kerfoot, Liyuan Chen, Jie Deng, Marc F Österdahl, Nathan J Cheetham, David A Drew, Long H Nguyen, Joan Capdevila Pujol, Christina Hu, Somesh Selvachandran, Lorenzo Polidori, Anna May, Jonathan Wolf, Andrew T Chan, Alexander Hammers, Emma L Duncan, Tim D Spector, Sebastien Ourselin, and Claire J Steves. Risk factors and disease profile of post-vaccination sars-cov-2 infection in uk users of the covid symptom study app: a prospective, community-based, nested, case-control study. *The Lancet Infectious Diseases*, 2021.

[32] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, 19(3):141–154, 2021.

[33] Piero Poletti, Marcello Tirani, Danilo Cereda, Filippo Trentini, Giorgio Guzzetta, Giuliana Sabatino, Valentina Marziano, Ambra Castrofino, Francesca Grosso, Gabriele Del Castillo, et al. Association of age with likelihood of developing symptoms and critical disease among close contacts exposed to patients with confirmed sars-cov-2 infection in italy. *JAMA network open*, 4(3):e211085–e211085, 2021.

[34] Diana Buitrago-Garcia, Dianne Egli-Gany, Michel J Counotte, Stefanie Hossmann, Hira Imeri, Aziz Mert Ipekci, Georgia Salanti, and Nicola Low. Occurrence and transmission potential of asymptomatic and presymptomatic sars-cov-2 infections: A living systematic review and meta-analysis. *PLoS medicine*, 17(9):e1003346, 2020.

[35] Pratha Sah, Meagan C. Fitzpatrick, Charlotte F. Zimmer, Elaheh Abdollahi, Lyndon Juden-Kelly, Seyed M. Moghadas, Burton H. Singer, and Alison P. Galvani. Asymptomatic sars-cov-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, 118(34), 2021.

[36] Nicholas G Davies, Christopher I Jarvis, W John Edmunds, Nicholas P Jewell, Karla Diaz-Ordaz, and Ruth H Keogh. Increased mortality in community-tested cases of sars-cov-2 lineage b. 1.1. 7. *Nature*, 593(7858):270–274, 2021.

[37] Ingrid Arevalo-Rodriguez, Diana Buitrago-Garcia, Daniel Simancas-Racines, Paula Zambrano-Achig, Rosa Del Campo, Agustin Ciapponi, Omar Sued, Laura Martinez-García, Anne W. Rutjes, Nicola Low, Patrick M. Bossuyt, Jose A. Perez-Molina, and Javier Zamora. False-negative results of initial rt-pcr assays for covid-19: A systematic review. *PLOS ONE*, 15(12):1–19, 12 2020.

[38] Kayoko Shioda, Max SY Lau, Alicia NM Kraay, Kristin N Nelson, Aaron J Siegler, Patrick S Sullivan, Matthew H Collins, Joshua S Weitz, and Benjamin A Lopman. Estimating the cumulative incidence of sars-cov-2 infection and the infection fatality ratio in light of waning antibodies. *Epidemiology (Cambridge, Mass.)*, 32(4):518, 2021.

S1 Supplementary analysis

S1.1 Effect of vaccination on ascertainment rate We apply different test sensitivity functions to the proportion of infections that are in individuals who have received a vaccine and those who have not. Here we first describe how the proportion of infections that are in vaccinated, and unvaccinated, people is estimated. We then describe how this was accommodated into our analysis.

Vaccine effectiveness is defined as

$$e = 1 - \frac{P(I|vaccinated)}{P(I|unvaccinaed)}. \tag{14}$$

It varies depending on the time since vaccination, the number of doses, the specific vaccine given, and the outcome measured e.g. infection, symptoms, hospitalisation. Effectiveness is lowest when the measured outcome is infection of any kind, regardless of symptoms. This is estimated to be $e = 0.56$ (56%) [29]. We choose to use this low value to avoid underestimating the effect of vaccines; lower estimates of effectiveness results in higher proportions of infections that are subject to the effects of the vaccine.

We have that $P(I|vaccinated) = p_x I(t)/V(t)$ where p_x is the proportion of infections that are in people who have received the vaccine by day t, $I(t)$ is the number of new infections on day t, V is the number of people vaccinated by day t, and N is the total population. Similarly $P(I|unvaccinated) = (1 - p_x)I(t)/(N - V(t))$ where N is the population. Substituting into Eq.(14) gives

$$p_x = \frac{(1 - V(t))e}{1 - eV(t)} \tag{15}$$

Vaccination has been reported to reduce the time until viral clearance of those infected [28]. It was reported that the time from viral peak to viral clearance was 2 days shorter for vaccinated individuals. To model this we assume that there is no viral shedding detectable by either PCR or LFD test 10 or more days after exposure, shortening the time to viral clearance considerably more than the reported effect to ensure we do not underestimate the effect of vaccines in this sensitivity analysis. For infections in vaccinated people we denote these modified functions using $S'q(\tau)$, and $P'(\tau)$ for the equivalent of Eq.(6) with S' substituted for S. The modified version of Eq.(7) is

$$\theta_x I(x) = (1-p_x)\sum_q \sum_{\tau \geq 0} C_q(x+\tau)P_q(\tau) + p_x \sum_q \sum_{\tau \geq 0} C_q(x+\tau)P_q^{'}(\tau) \tag{16}$$

and finally the modified Eq.(8) is

$$\hat{M}(t) = (1-p_x)\sum_{k \geq 0} I(t-k)S_{PCR}(k) + p_x \sum_{k \geq 0} I(t-k)S_{PCR}^{'}(k). \tag{17}$$

The effect of these modifications is shown in Fig. S2.

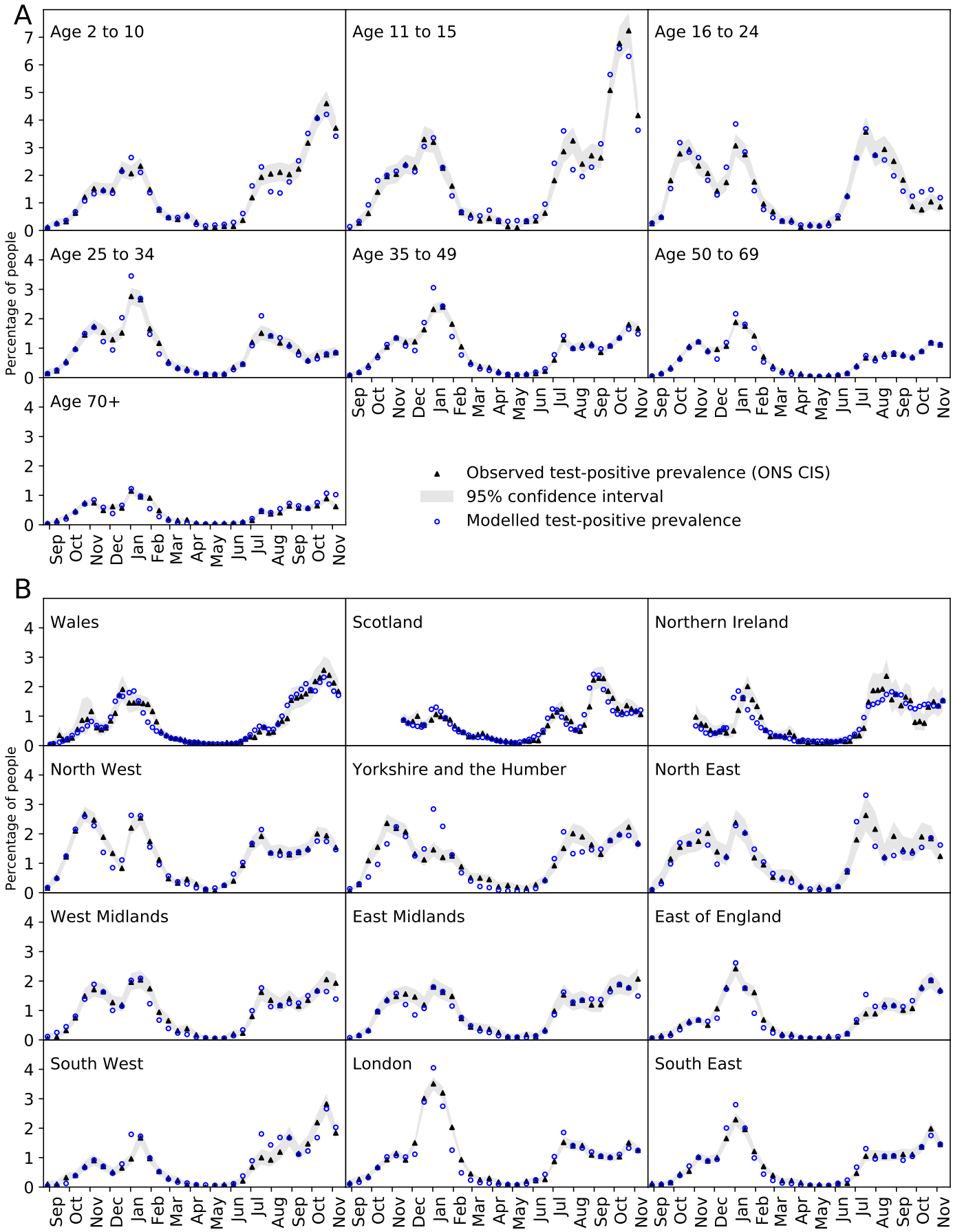

Figure S1: The percentage of people who would test positive if tested. Comparison of the ONS CIS to the modelled value based on reported cases and the estimated ascertainment rates given in Figure 3.

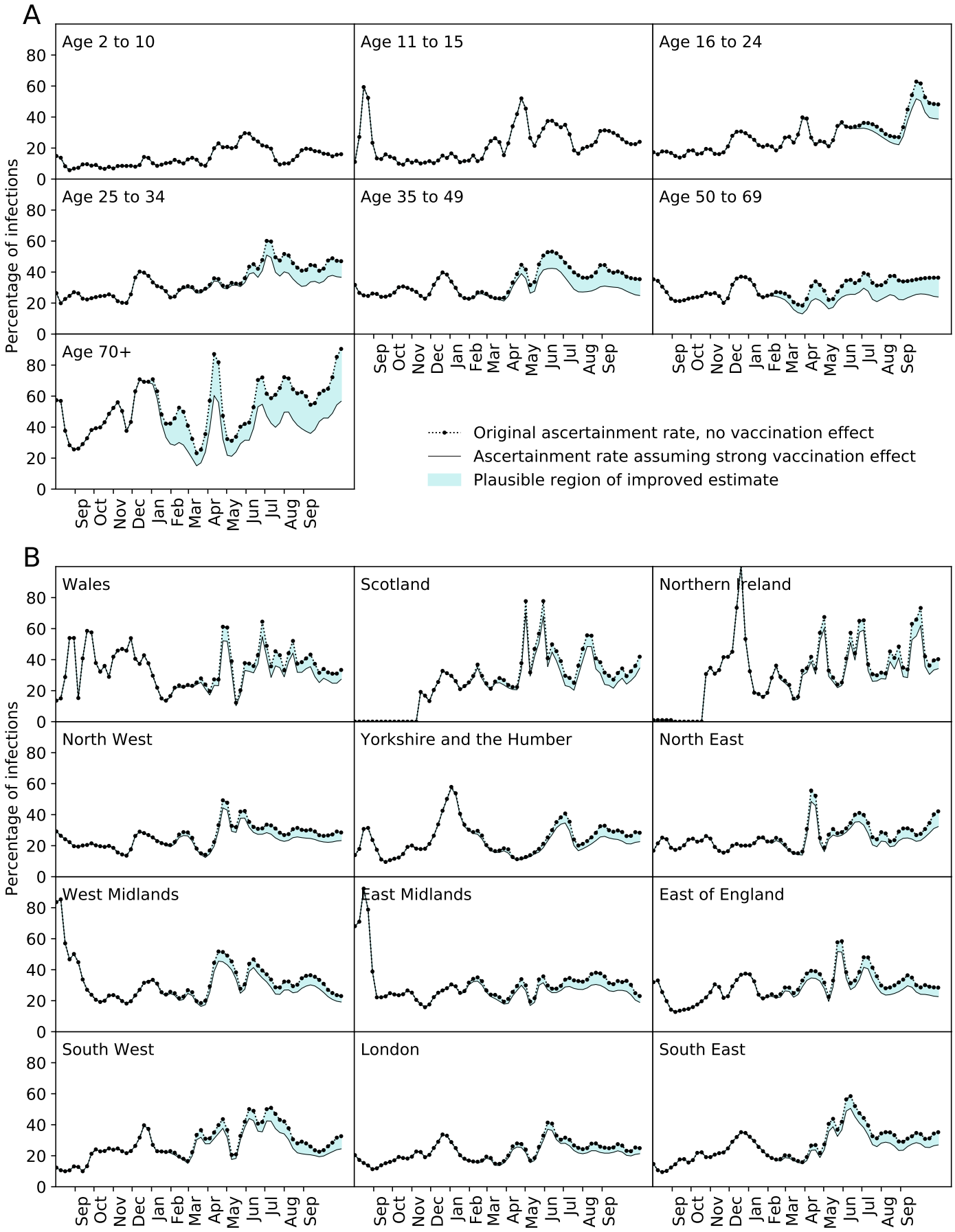

Figure S2: Ascertainment rates shown in Fig. 10 compared to the equivalent value that incorporates the effect of vaccination. Modelling assumptions have been made to provide the largest reasonable deviation from the original ascertainment estimate with the data available. Therefore it is likely that a precise treatment of vaccination in the model would yield a result within the shaded area between the curves.

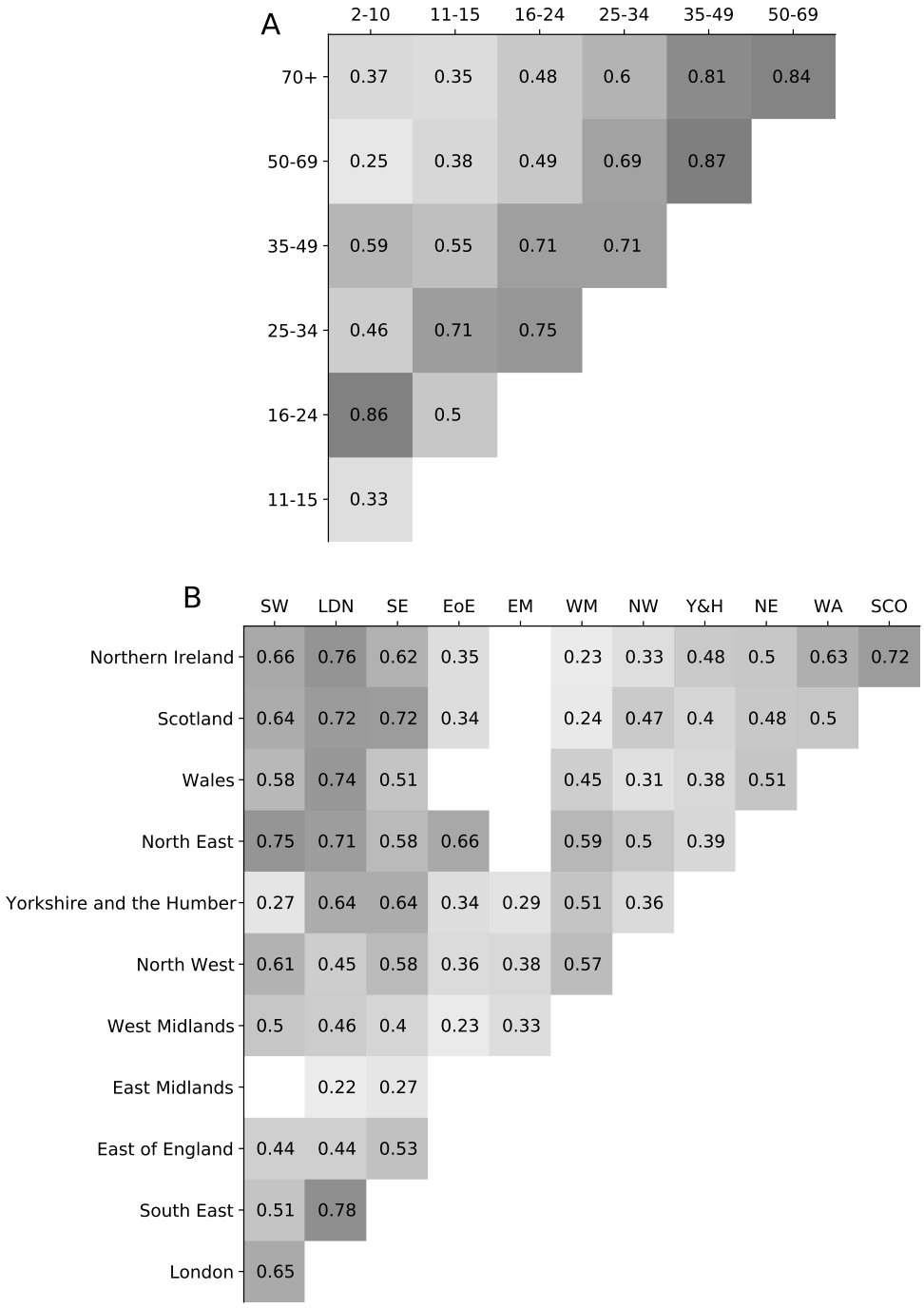

Figure S3: Infection fatality rates shown in Fig. 4 compared to the equivalent value that incorporates the effect of vaccination. Modelling assumptions have been made to provide the largest reasonable deviation from the original ascertainment estimate with the data available. Therefore it is likely that a precise treatment of vaccination in the model would yield a result within the shaded area between the curves.

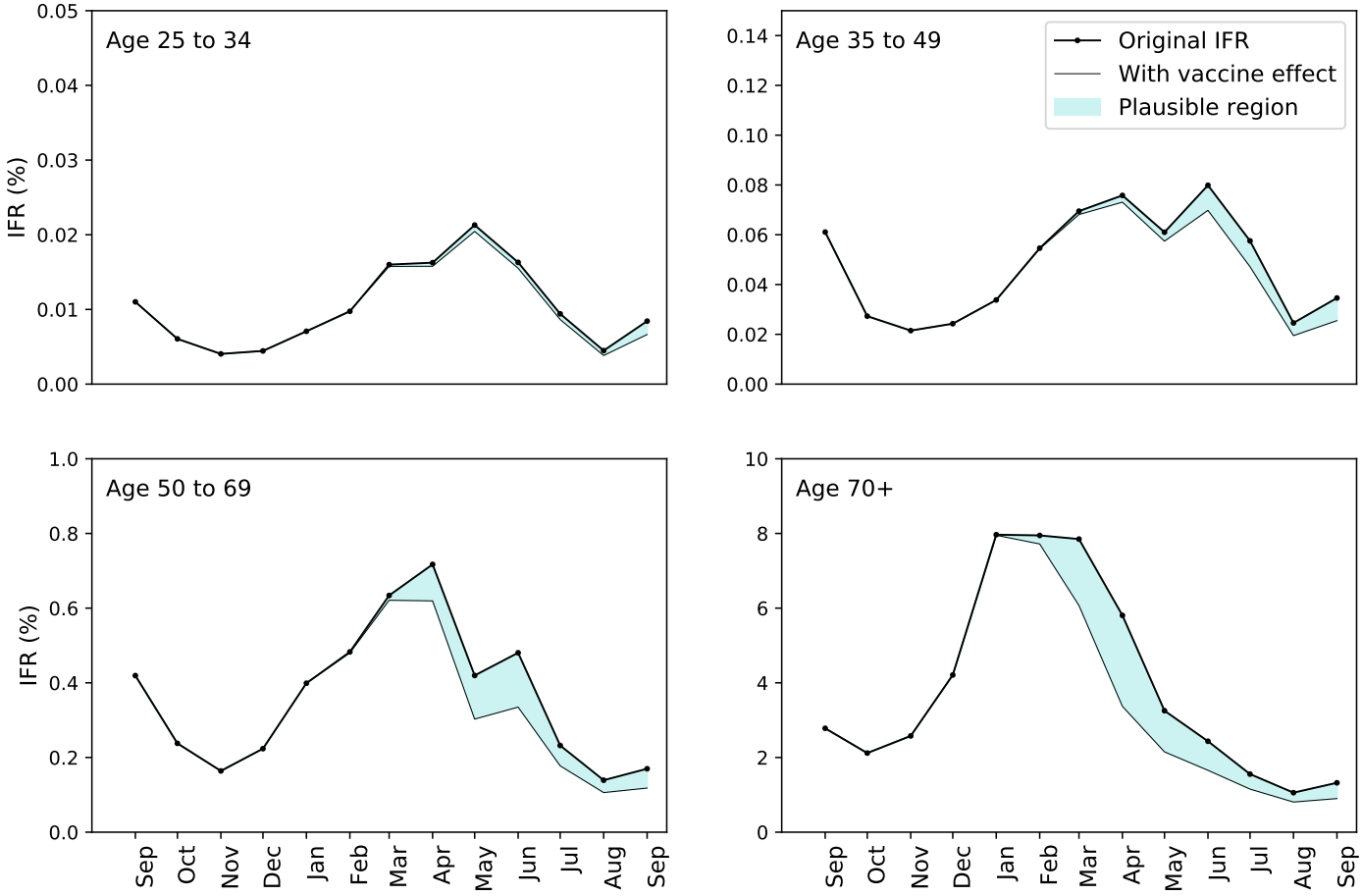

Figure S4: Correlations between the ascertainment rate time series of (A) all age groups and (B) all regions. Values show the Pearson correlation coefficient, r. Correlations where $r < 0$ or $p > 0.05$ are not displayed.