Journal of Theoretical Biology 611 (2025) 112177

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Estimating the contribution of age-structure to the COVID-19 epidemic in England

Robert Hinch [a,*], Jasmina Panovska-Griffiths [a,b], Christophe Fraser [a]

[a] Pandemic Sciences Institute, University of Oxford, Oxford, UK
[b] The Queen's College, University of Oxford, Oxford, UK

ARTICLE INFO

Keywords:
Data-driven models
COVID-19
Reproduction number R
Age-dependence.

ABSTRACT

The spread of epidemics in populations is often inhomogeneous, consequently infection incidence varies between sub-populations. Age-structure is often particularly important in the dynamics of epidemics, due to the contact patterns between individuals of different ages. Public health interventions are often targeted at specific age-groups, therefore analysing the age-structure of transmission patterns is essential to evaluate the efficacy of these interventions. We develop a Bayesian model to estimate the contribution of different age-groups to the reproduction number (R) and to new infections for COVID-19 in England throughout 2021, using the ONS Infection Survey. We model a dynamic next-generation matrix in a novel way by splitting it into a static survey-derived social-contact matrix, multiplied by a low-rank dynamic matrix. We show that whilst R was typically highest for school-age children (5–11y and 12–17y) and lowest for the elderly (60y+), the former typically rose during term-time and fell during the school-holidays. The dynamics for young adults (18–29y) were particularly interesting, which increased relative to older adults in late-spring 2021 following the re-opening of entertainment venues. The R peaked for young adults in July 2021 coinciding with the period of the Euros football tournament, before rapidly dropping as the national vaccination program reached this group in August 2021. Our model is an important tool that can estimate R and attribute new infections by the infector's age, thus identifying core groups which sustain the epidemic and informing the design of targeted interventions.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spread in England from 2020 and has been characterised by the emergence of novel variants which quickly became dominant. In late 2020, the wild type variants were replaced by the Alpha variant, followed by the Delta variant in spring 2021 and the Omicron variants in 2022. By 30 December 2022, over 20.4 million confirmed cases and over 183,000 deaths related to COVID-19 had been reported in England (Coronavirus , COVID-19).

Throughout the pandemic there were notable differences in the distribution of burden by age, observed in confirmed cases, population cohort infection and serology surveys, hospitalisations, and mortality rates (Sorensen et al, 2022; O'Driscoll et al, 2021). Whilst the predominant age-specific feature of COVID-19 was the concentration of disease burden in the elderly, infection rates were often highest in other age demographics, especially children and young adults. Understanding how infection rates vary by age is important to quantify the contribution of different age-groups to transmission, and thus their role in sustaining the epidemic. Specifically, by quantifying the role of each age-group in transmission can help inform how to target non-pharmaceutical interventions and vaccination programs. For example, should social contact mixing be reduced in a blanket way (*i.e.* via a national lockdown), or should they be targetted at specific aspects of society?

During the epidemic, vast amounts of data were collected recording confirmed cases, hospitalisations and mortality, much of it stratified by age. Whilst hospitalisation and mortality statistics gave an accurate measurement of the disease burden, confirmed case counts were not a true reflection of actual infections due to different ascertainment rates. Case ascertainment rates not only varied with time throughout the epidemic, but also differed dynamically between age-group. In England, the testing strategies changed throughout the epidemic leading to age-specific under-reporting of cases due to the policy at the time. For example, testing in schools and workplaces was compulsory with automatic reporting of positive test over late 2020 but this shifted to self-testing and self-reporting of positive tests in 2021. In addition to case statistics, the Office for National Statistics carried out a longitudinal population cohort study, the ONS Infection Survey, Office for National Statistics (2022a), to estimate infection prevalence by age-group and geographic region. For this survey, representative households were recruited and

* Corresponding author.
E-mail address: robert.hinch@ndm.ox.ac.uk (R. Hinch).

https://doi.org/10.1016/j.jtbi.2025.112177
Received 31 October 2024; Received in revised form 1 June 2025; Accepted 5 June 2025
Available online 7 June 2025

each household periodically submitted samples for PCR testing. Test positivity rates were reported bi-weekly, with approximately 150,000 samples being tested per reporting period in 2021.

Whilst age-specific infection rates tell us the distribution of the infection burden, they do not tell us directly about age-specific transmission rates. This is due to social-mixing between age-groups, whereby an infected person of one age might transmit the virus to a recipient of a different age. For example, a high incidence of infections amongst school-age children could lead to a high incidence of infections amongst their parents, even if the reproduction number of the parents was less than 1. The link between infection incidence and transmission rates is thus complex, with mathematical modelling being a useful tool to estimate one from the other.

Modelling epidemics in structured populations has been an area of extensive research in epidemiology for many years (Anderson, 1991; Diekmann et al., 1998, 2013). Examples of characteristics that cause structured epidemics are age, sex, geography, genetic composition, and individual risk tolerance (Diekmann et al., 2013; Pickles et al., 2021). A standard modelling approach is to divide the population into compartments based on relevant characteristics, with transmission of the pathogen between compartments described by the next generation matrix (NGM) (Diekmann et al., 2010). During the pandemic, modelling was extensively used to characterise temporal changes in the epidemiology of SARS-CoV-2 (Rozhnova et al, 2021; Hinch et al., 2020; Hinch et al., 2021; Hozé et al, 2021; Milne et al, 2021; World Health Organization, 2022b; Viana et al, 2021). Most of these studies focused on answering specific questions and were not age-specific (Pei et al., 2021; Saad-Roy et al, 2020; Kissler et al., 2020), explored specific time-periods: the first epidemic wave (Dekker et al., 2023; Gatto et al, 2020; Bertuzzo et al, 2020; Giordano et al, 2020; Panovska-Griffiths et al., 2020) or periods of Alpha (Hozé et al, 2021; Viana et al, 2021; Panovska-Griffiths et al., 2022a; Hinch et al., 2022) and Delta variants (Sonabend et al, 2021; Bosetti et al, 2022; Panovska-Griffiths et al., 2022b). Some modelling studies have explored the importance of different age-groups in the epidemic (Hozé et al, 2021; Davies et al., 2020; Monod et al, 2021; Kiem et al., 2021; Boldea et al., 2024). Davies et al. (2020) fitted an age-structured model to data from China, Italy, Japan, Singapore, Canada and South Korea, to show that susceptibility to infection in < 20y old was approximately half that those >20y old, while that clinical symptoms manifested in 21 % (95 % CI: 12–31 %) of infections in 10–19y old, rising to 69 % (57–82 %) of infections in people aged over 70y. Monod et al (2021) used a longitudinal, and age-specific population mobility and COVID-19 mortality data from USA to show that as of October 2020, individuals 20–49y old were the group sustaining resurgent SARS-CoV-2 transmission and responsible for 65 % of COVID-19 infections. Kiem et al. (2021) fitted a model to age-stratified hospital admissions and positivity rates among symptomatic people to show that 20–29y old had the largest contribution to the French COVID-19 epidemic over the summer 2020. Hozé et al (2021) extended this work to include all of 2020, showing that the proportion infected was twice as high in 20–49y old than in those >50y old. A recent study by Boldea et al. (2024) used inference modelling to reconstruct the burden of true infections and hospital admissions in children, adolescents, and adults over the seven waves of four variants (wild-type, Alpha, Delta, and Omicron BA.1) during the first 2 years of the pandemic in Netherlands. They found that case reporting differed by age and the number of hospital admissions in adolescents and children remained much lower than those of adults. They also saw a shift in infections and notably higher hospitalisation towards children and younger age-groups in later (Omicron/Delta) compared to earlier (Alpha/wild-type) waves.

These existing studies point to the inhomonegeinty in the incidences of infections, hospitalisations and deaths from COVID-19 across different age-cohorts and settings. Our work here complements this existing work, and explores such inhomogeneity over the Delta epidemic wave in England in 2021. Specifically, we develop a Bayesian model for estimating the contribution of different age-groups to new infections throughout the English epidemic over the Delta epidemic wave in 2021. Infections in each age-group are modelled using a deterministic exposure-infection model with a time-dependent next-generation matrix. With N age-groups and T time points, the time-dependent next-generation matrix has N^2T parameters to estimate, which is not possible directly given the available data. Therefore, and to the best of our knowledge for the first time, we introduce a model for the next-generation matrix which requires only N parameters to be estimated at each step, with each parameter following a log-normal process through time. We then use the model to calculate the positivity rate for each age-group in every round of the infection survey (Office for National Statistics, 2022a), which allows the likelihood of the observed survey data to be calculated (the observation model). The posterior distribution of the model parameters is calculated using MCMC, which allows the calculation of age-specific reproduction rates and attribution of new infections. The results are then compared with the timeline of public health interventions.

2. Methods

2.1. Data

During the COVID-19 epidemic in the UK, extensive PCR testing was carried out to detect infections, with 100,000 s of tests administered each day. The largest proportion of these test were case based, administered after self-reported symptoms. Whilst this is the most extensive daily data set, it suffers from multiple ascertainment biases, both cross-sectional between age-group and temporally. In parallel to the case-based testing, the UK Office for National Statistics (ONS) carried out a bi-weekly household infection survey throughout the pandemic. The survey reported both the number of tests administered and the number of positive tests segregated by age-group, as well as estimates of the overall population positivity (Fig. 1). As a household survey, it suffers from less ascertainment biases than case-based testing, especially temporally. However, since it is a measure of positivity, there may be ascertainment biases between age-groups due to people of different ages testing positive for different lengths of time.

At the end of January 2021, in the UK almost all SARS-CoV-2 infections were of the Alpha variant. In late March 2021, the first cases of the Delta variants were recorded in the UK, with almost all cases being of the Delta variant by the end of June 2021. Throughout the epidemic a large number of virus samples were sequenced by the COVID-19 Genomics UK Consortium (COG-UK), which gave a comprehensive view of the circulating variants in the UK (COVID-19, 2024).

In this paper we will develop and use a model to estimate the contribution of each age-group to sustaining the epidemic through 2021. The two main datasets we use will be the ONS Infection survey (Office for National Statistics, 2022a) and the COG-UK genomics time-series (COVID-19, 2024). Additionally, we use a single contact matrix, which we use as a base and estimate adjustments through time. The base contact matrix is from the Comix study in September 2020 (Gimma et al., 2022).

2.2. Model description

The aim of the model will be to estimate the contribution of each age-group to the reproduction number of the epidemic throughout time. There are many factors which contribute to this, such as social mixing patterns, population infection history, mix of circulating variants, vaccination programs, test-trace and isolate programs, the presence of lockdowns and other non-pharmaceutical interventions. Whilst it is possible to directly model all of these effects in mechanistic models, such as agent-based models, it is very difficult to estimate all the parameters in these models. The ability of these models to estimate the effect of different interventions relies heavily on strong priors on these parameters or estimating them independently from targeted studies. We will

not attempt to attribute the rate of transmissions to these different effects, therefore we will use a simple exposure-infection model where we directly estimate change to the force of infection.

2.2.1. Generative model

The underlying infection dynamics are modelled as a determinisitc discrete-time exposure-infectious (EI) model with age stratification and stochastic transmission parameters (e.g. the reproduction number). Consider an age-structured population, where each individual is in one of N_{age} groups. For each age-group i, we model the number of people exposed (but not infectious) E_i and the number of infectious people I_i. Due to mixing between age-groups, new infections can be transmitted between people in different age-groups. This rate of new infections in the i^{th} age-group from individuals in the n^{th} at time t is given by the component of the next-generation matrix $G_{i,n}(t)$. Schematically our model is where k is the rate at which people leave the exposed and infectious states (assuming it is the same for both states, then it is equivalent to the generation time being gamma distributed with shape parameter 2) and we assume that the parameter k is the same for all age-groups.

In addition to the age-stratification, we consider multiple circulating variants, with the introduction of the Delta variant in spring 2021 of primary interest. The next-generation matrix for each variant is assumed to be the same up to a variant specific transmission multiplier. For convenience we express the age profile of exposed and infectious individuals as vectors (*i.e.* $\mathbf{E}(t) = \{E_1(t), .., E_{N_{\text{age}}}(t)\}$), and the number of exposed individuals with variant υ as $\mathbf{E}_\upsilon(t)$, then the infection model is

$$\begin{aligned}
\mathbf{E}_\upsilon(t+1) &= (1-k)\mathbf{E}_\upsilon(t) + \mathbf{N}_\upsilon(t+1),\\
\mathbf{N}_\upsilon(t+1) &= \mu_\upsilon G(t)\mathbf{I}_\upsilon(t) + \mathbf{S}_\upsilon(t),\\
\mathbf{I}_\upsilon(t+1) &= (1-k)\mathbf{I}_\upsilon(t) + k\mathbf{E}_\upsilon(t),
\end{aligned} \tag{1}$$

where $\mathbf{N}_\upsilon(t+1)$ is the number of new infections, $\mathbf{S}_\upsilon(t+1)$ are seed infections (see Appendix A.1.5), and μ_υ is the variant transmissibility multiplier for the υ^{th} variant.

The infection dynamics model (1) is relatively standard, the novel part of the model is for the next-generation matrix $G(t)$, which has N_{age}^2 components per time-step. Clearly the ONS survey data is insufficient to determine all these parameters independently, therefore we must impose some structure on how the components vary through time. However, there are some basic patterns of social mixing which must be respected, therefore as a base we use the Comix contact matrix from September 2020. The base matrix is adjusted on each time-step by multiplying it element-wise (*i.e.* Hadamard product, represented by $\odot$) with a rank-1 matrix (*i.e.* a matrix formed by taking the outer product of 2 vectors). Each time step we estimate a vector $\mathbf{A}(t)$, then form the multiplicative rank-1 matrix by taking the outer product of $\mathbf{A}^{1-\rho}$ with $\mathbf{A}^\rho$, where ρ is an estimated mixing parameter. The operation of element-wise multiplication by a rank-1 matrix is equivalent to pre- and post-multiplying the contact matrix by diagonal matrices, therefore it is equivalent to adjusting the susceptibility and infectiousness of each age-group, with ρ determining the split between the two. An example of this adjustment with 3 age-groups is

$$\begin{aligned}
C \odot (\mathbf{A}^{1-\rho}(\mathbf{A}^\rho)^T) &= \text{diag}(\mathbf{A}^{\mathbf{1-\rho}}(t))\, C\, \text{diag}(\mathbf{A}^\rho(t))\\
&= \begin{bmatrix} C_{11}A_1 & C_{12}A_1^{1-\rho}A_2^\rho & C_{13}A_1^{1-\rho}A_3^\rho \\ C_{21}A_2^{1-\rho}A_1^\rho & C_{22}A_2 & C_{23}A_2^{1-\rho}A_3^\rho \\ C_{31}A_3^{1-\rho}A_1^\rho & C_{32}A_3^{1-\rho}A_2^\rho & C_{33}A_3 \end{bmatrix},
\end{aligned} \tag{2}$$

note that this introduces N_{age} parameters to be estimated at each step instead of N_{age}^2 if the full contact matrix was re-estimated. Finally, the generation matrix is multiplied by a scalar $R(t)$, which models global changes in the transmissibility which effect all age-groups (e.g. lockdowns). The next-generation matrix is thus

$$G(t) = k\,R(t)\,\text{diag}(\mathbf{A}^{\mathbf{1-\rho}}(t))\, C\, \text{diag}(\mathbf{A}^\rho(t)), \tag{3}$$

where C is the Comix contact matrix normalised by its largest eigenvalue (see Appendix A.1.3). The reason for this normalisation and the factor of k, is that if $\mathbf{A} = \mathbf{1}$, then $R(t)$ is the reproduction number. The values of $R(t)$ and $\mathbf{A}(t)$ are expected to be highly auto-correlated in time, therefore we model them as zero-drift log-normal processes.

$$\begin{aligned}
\log(R(t+\tau)) &= \log(R(t)) + \epsilon(t),\\
\epsilon(t) &\sim N\left(-\sigma_R^2/2, \sigma_R^2\right),\\
\log(\mathbf{A}(t+\tau)) &= \log(\mathbf{A}(t)) + \boldsymbol{\mu}(t)\\
\boldsymbol{\mu}(t), &\sim N\left(-\mathbf{1}\sigma_A^2/2, \sigma_A^2 I\right),
\end{aligned} \tag{4}$$

where σ_R^2 and σ_A^2 are the variances of the movements between time-steps and are estimated parameters; and τ is the time between survey rounds (2 weeks). For the duration of each survey round we use a constant value for R and $\mathbf{A}$, as it is not possible to estimate the parameters at a finer scale than the data to which we fit the model. Note that although the next-generation matrix is constant within each round of the survey, the generative model calculates new infections daily.

2.2.2. Observation model

The generative model is linked to the survey data using an observation model. The positivity rate, $p_i(t)$, for an age-group during each survey period is estimated by summing the mean values of the exposed and infectious individuals across variants. In the model we use a mean generation time of 5.5 days, therefore we are assuming that individuals test positive for an average of 11 days. Individuals in the survey are assumed to be drawn randomly from the population, so the number of positive test samples in each round is distributed binomially.

$$\begin{aligned}
p_i(t) &= \frac{1}{\tau Pop_i}\sum_{s=0}^{\tau-1}\sum_{\upsilon=1}^{N_{\text{var}}}\left([\mathbf{E}_\upsilon(t-s)]_i + [\mathbf{I}_\upsilon(t-s)]_i\right)\\
n_{\text{pos},i}(t) &\sim \text{Bin}(N_{\text{surv},i}(t), p_i(t)),
\end{aligned} \tag{5}$$

where Pop_i is the total population in the i^{th} age-group; $n_{\text{pos},i}(t)$ is the number of positive tests and $N_{\text{surv},i}(t)$ is the number of samples in the i^{th} age-group of the survey at time t. For the genomic data we assume that all samples are drawn equally from positive cases over the reporting period (i.e. there is no bias towards a particular variant or age-group). Let $x_\upsilon(t)$ be the probability that a sample in the reporting period ending at t is of variant υ, then

$$\begin{aligned}
x_\upsilon(t) &= \frac{1}{\sum_j x_j(t)}\sum_{s=0}^{\tau-1}\left(\mathbf{1}.(\mathbf{E}_\upsilon(t-s)] + \mathbf{I}_\upsilon(t-s))\right)\\
\mathbf{m}(t) &\sim \text{MultiNomial}\left(M(t), \mathbf{x}(t)\right),
\end{aligned} \tag{6}$$

where $\mathbf{x}(t) = \{x_1(t), .., x_{N_{\text{var}}}(t)\}$, $m_\upsilon(t)$ are the number of samples of variant υ at time t, $\mathbf{m}(t) = \{m_1(t), .., m_{N_{\text{var}}}(t)\}$. and $M(t)$ is the total number of samples at time t.

2.3. Parameter estimation

The model parameters were estimated in a Bayesian framework. Range priors are put on parameters (see Appendix A.2.1) with the exception of the base contact matrix and the generation time which are determined from the literature (see Appendix A.2.2). The posterior distribution of the parameters is estimated using MCMC, with the model implemented in RStan (Carpenter et al., 2017). We ran 12 chains of length 1000 with a burn-in period of 500. The largest Rhat across all parameters was 1.02 (mean < 1.01) suggesting the chains were well mixed. Full details are in the Appendix A.1.7 and the code is available on Github as the R package *EpiAgeVar* (Hinch, 2024).

3. Results

3.1. Validity of fit

The first test of the model is that it is able to fit the survey data. Fig. 1 shows the model fit and ONS Infection Survey data of the positivity by

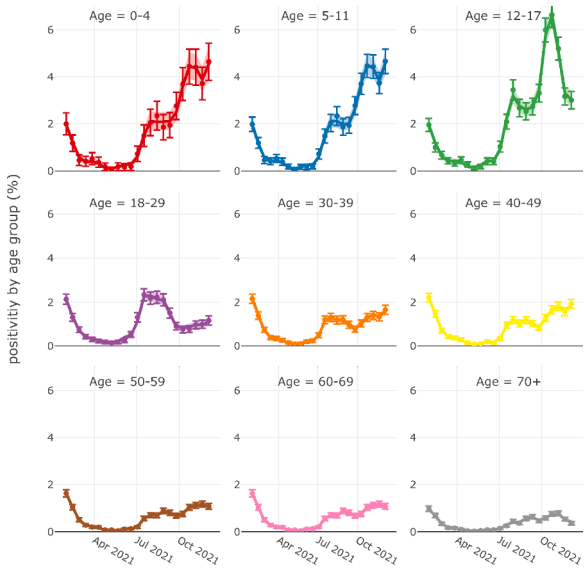


Fig. 1. Positivity by age-group throughout 2021. The points are from the ONS Survey data, with the error bars being the 95 % credible interval (see Appendix A.1.6). The solid line is the median of the posterior distribution of the fitted model and the shaded area is the 95 % credible interval. The charts show that the model successfully fitted the survey data.

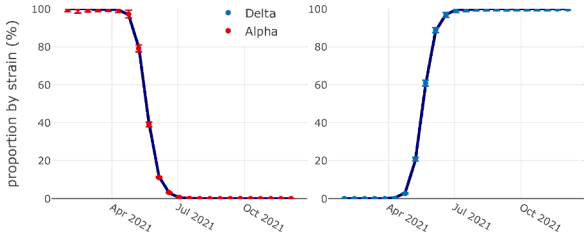


Fig. 2. Proportion of infection samples for the Alpha and Delta variants. The points are from the COG-UK analysis of cases samples. The solid line is the median of the posterior distribution of the fitted model and the shaded area is the 95 % credible interval. The charts show that the model successfully fitted the variant data.

age-group and demonstrates that at all points the posterior median lies within the 95 % credible interval of the survey data and *vice versa*. Fig. 2 shows the model fit to the COG-UK variant data and demonstrates that at all points the posterior median lies within the 95 % credible interval of the survey data and *vice versa*.

The second test of the model is whether its parameters are identifiable. Fig. 3 compares the posterior distributions (blue) to the prior distributions (grey) for the key model parameters, and the data were able to identify all of them. The model estimates the relative transmissibility of the Delta variant to the Alpha variant of 2.21 (CrI: 2.19-2.25), which is consistent with previous estimates (Earnest et al., 2022). The estimate of the age dynamics split in R between transmissibility and susceptibility (ρ) is 0.34 (CrI: 0.15-0.47), suggesting that age dynamics are driven more by changes in susceptibility. The estimates of the standard deviation of the log-normal process for overall R (σ_R) was 0.23 (CrI: 0.16-0.34) and for the age-specific R (σ_A) was 0.23 (CrI: 0.19-0.28), measured in units of fraction change in R per survey interval (2 weeks). This suggests that approximately half of the change in R over time could be attributed to population-wide changes and about half to change in specific age-groups.

3.2. Age-attributed $R(t)$ and new infections

We now look at the model estimates of the reproduction number R by age-group. We define R for an age-group as the expected number

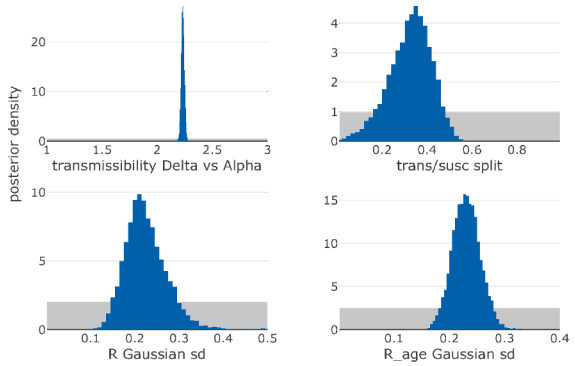


Fig. 3. Posterior distribution of key model parameters. The blue bars are the posterior distribution and the light grey is the prior distribution. The data was able to identify all these parameters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

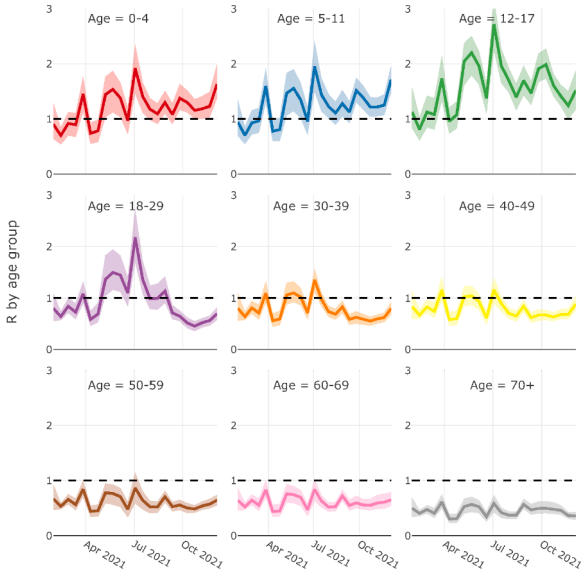


Fig. 4. R by age-group through time. The solid line is the median of the posterior distribution and the shaded areas is the 95 % credible interval of the posterior. The dashed black line is for $R = 1$. The charts show that R was consistently highest amongst school age children and lowest in older adults.

of new infections (of any age) generated by an individual of that age. Mathematically within the model this is

$$\mathbf{R}_v(t) = \mu_v R(t)(\mathbf{A}^{1-\rho(t)})^T C \operatorname{diag}(\mathbf{A}^{\rho(t)}), \tag{7}$$

where $\mathbf{R}_v(t) = \{R_{1,v}(t), .., R_{N_{\text{Age}},v}(t)\}$ and $R_{i,v}$ is the R of the i^{th} age-group infected with the v^{th} variant at time t. Given that the difference in R between variants is simply the transmissibility multiplier, we report the mean value across all variants weighted by the positivity of the variants at that time. Fig. 4 shows the posterior distributions of R by age through time. R was consistently the highest amongst school-age children and the lowest amongst older adults. The dashed line is at $R = 1$, with $R < 1$ consistently for adults over the age of 40 years.

An alternative measure of the importance of each age-group is to attribute new infections by the age of the infector. Mathematically within the model this is

$$\boldsymbol{\phi}(t) = \frac{\sum_{v=1}^{N_{\text{var}}} \mathbf{R}_v(t) \odot \mathbf{I}_v(t)}{\sum_{v=1}^{N_{\text{var}}} \mathbf{R}_v(t).\mathbf{I}_v(t)} \tag{8}$$

where $\boldsymbol{\phi}(t)$ as the proportion of new infections by the age of the infector and $\odot$ is element-wise multiplication. Fig. 5 shows the posterior

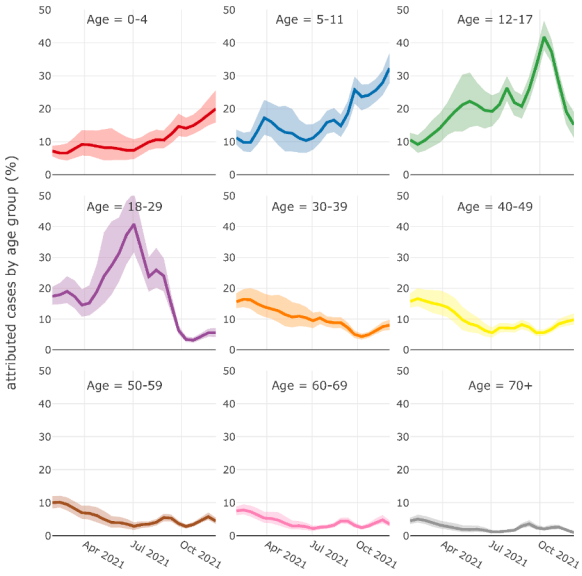
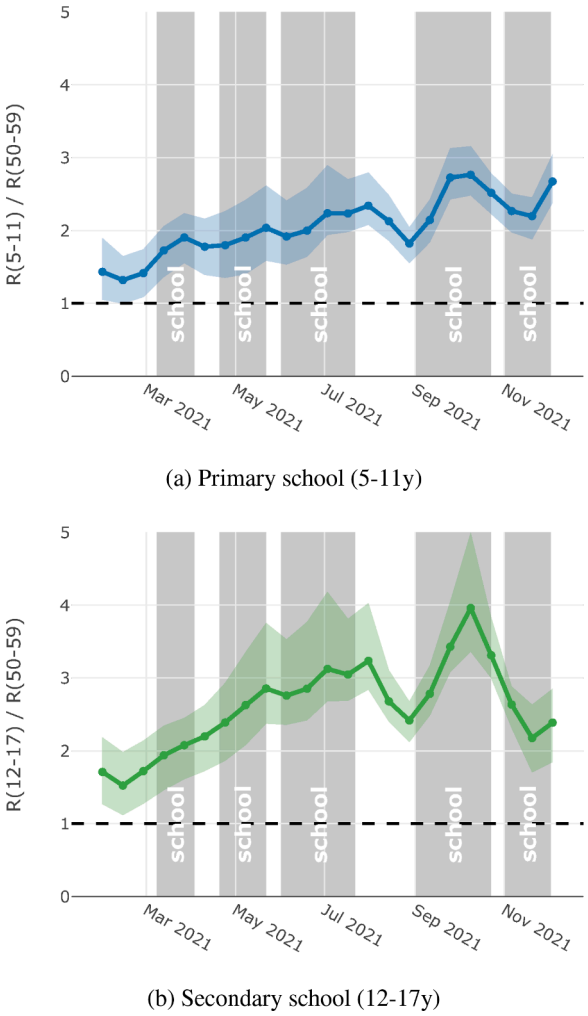


Fig. 5. Attribution of new infections by age of infector. The solid line is the median of the posterior distribution and the shaded areas is the 95 % credible interval of the posterior. In early 2021, infectors were split approximately evenly by age, before being concentrated sequentially in young adults followed by school age children.



(a) Primary school (5-11y)

(b) Secondary school (12-17y)

Fig. 6. Relative R in different age-groups compared to the 50y-59y your group. The solid line is the posterior median and the shaded area is the 95 % credible interval. School openings, are overlayed to give context as possible explanations for the changes in R.

distributions of new infections by the age of the infector. The attribution of new infections is very similar to that of R, but the temporal patterns are amplified because the high values of R in an age-group are correlated with a higher incidence of infections.

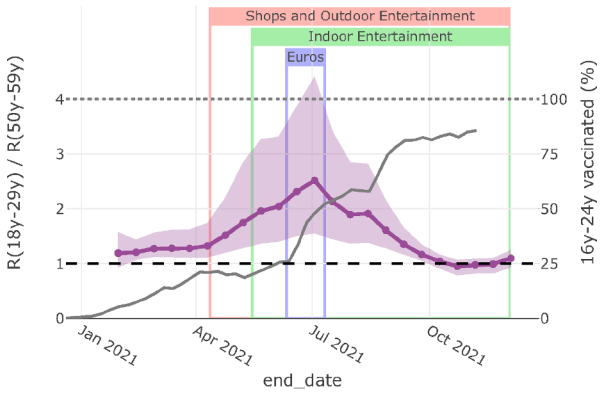
We now compare the estimates of R by age to the timing of events and interventions introduced to control the epidemic. Given that about half of the changes in R are correlated across age-groups (since $\sigma_R \approx \sigma_A$), clear temporal patterns are observed when we consider the relative value of R between age-groups. As a denominator we use the R of the age 50–59 year group, which was the most stable through time and is a group where the majority are employed but do not have school-age children. Fig. 6(a) and (b) shows the R for primary aged (5–11y) and secondary aged (12–17y) school children, along with the times at which the schools were open. In January and February 2021 the schools were closed in the UK during a national lockdown, and in this time R of children was only 20 % to 50 % higher than that of the 50–59 year group. Schools re-opened on March 8th 2021, at which point the relative value of R began to rise amongst school age children, only falling during the school holidays. Additionally it should be noted that the 50–59 year group received their first vaccine from late-February, reaching 97 % coverage by mid-April.

Next we consider the relative R of young adults (18–29y), which was stable and just above 1 at the start of the year (Fig. 7(a)). From mid-April and in May the relative R of young adults began to increase, which coincided with the Step 2 (April 12th: shops and outdoor venues) and Step 3 (May 17th: indoor restaurants and pubs) of the UK Reopening Roadmap for easing of restrictions. The relative value of R for young adults peaked at the start of July, which coincided with the Euro 2021 football tournament when large groups gathered in pubs to watch the matches. A similar trend was observed in the UK COVID-19 App, where both contacts and transmissions peaked on the day of matches (Kendall et al., 2024). From mid-July the relative R in young adults began to fall, reaching parity with the 50–59y age-group by October. It is not clear the exact reason for this decline, although it coincided with the time at which this age-group started to receive their COVID-19 vaccinations.

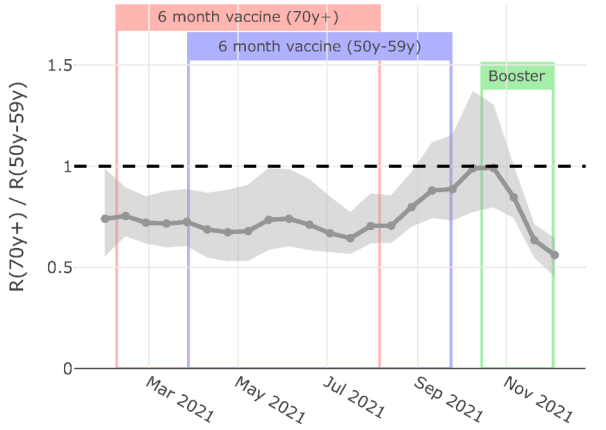
The final age-group we consider is the over 70y age-group, which up until August 2021 showed a consistent relative R of about 0.8 to the 50–59y group (Fig. 7(b)). In August and September 2021 the gap closed to almost parity, before the relative R dropped dramatically in October. A possible explanation for this is the waning of vaccine-induced immunity in the older age-group. The UK followed a schedule of vaccination by age starting with the oldest age-groups, therefore people in the over 70y group were typically vaccinated against 6 to 8 weeks prior to the 50–59y group. In late September 2021, the UK commenced a booster vaccination programme, starting with the elderly and people with pre-existing medical conditions, which provides a possible explanation for the relative reduction of R in the over 70y group in October 2021.

In addition to the age-attributed R, the model can also be used to generate an overall R estimate (Fig. 8), which is the mean age-attributed R weighted by the number of positive individuals. The overall R shows 3 periods where it rose rapidly. Firstly, in March 2021, R increased rapidly following the initial school openings after the 3rd national lockdown, before falling again over the Easter holidays. Secondly, in May 2021, R increased rapidly, driven by arrival of the more transmissible Delta variant, the school term, and the re-opening of shops, restaurants and entertainment venues. Finally, there was a brief but sharp peak in late-June and early-July, probably caused by increased social gatherings linked to the Euro 2021 football tournament.

An alternative way to age attribute $R(t)$, is to allocate to each age-group proportional to the number of new infections generated by each group (Fig. 9). Since the total number of new infections generated by



(a) Young Adult (18-29y)



(b) Elderly (>70y)

Fig. 7. Relative R in different age-groups compared to the 50–59y your group. The solid line is the posterior median and the shaded area is the 95 % credible interval. Key events are overlayed to give context as possible explanations for the changes in R.

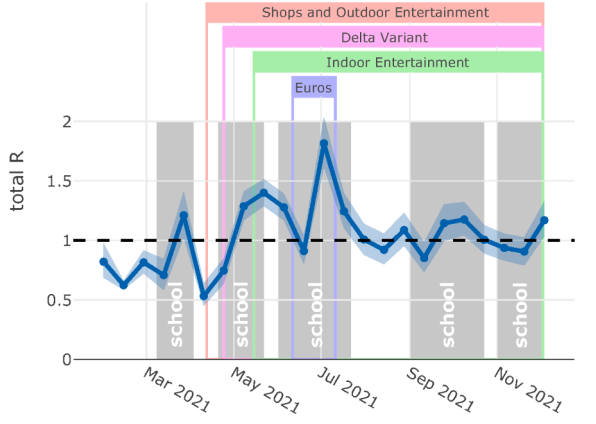


Fig. 8. Overall estimate of $R(t)$, annotated with key events which would likely effect it. The line is the posterior median estimate and the shaded error is the 95 % credible interval.

each group is the R for the age-group multiplied by the number of infectious people in that age-group, the difference between age-groups is larger because ages with higher R typically also have a higher incidence of infections. The results show that in early 2021 that new infections were spread across all age-groups, however, by late-spring new infections were predominately transmitted by school-age children and young

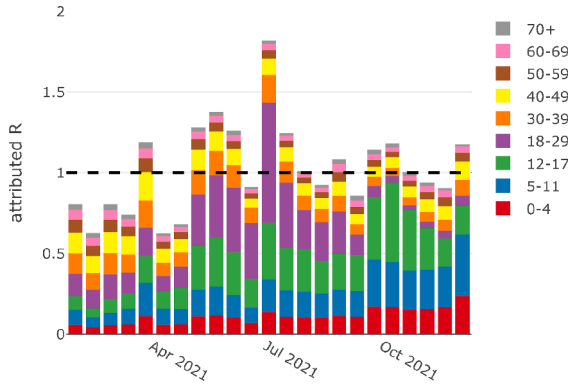


Fig. 9. Overall estimate of $R(t)$ attributed to each group. The total height of the bars is posterior mean of $R(t)$ (*i.e.* Fig. 8), and the divisions by age are proportional to the total number of new infections generated by individuals in each age-group.

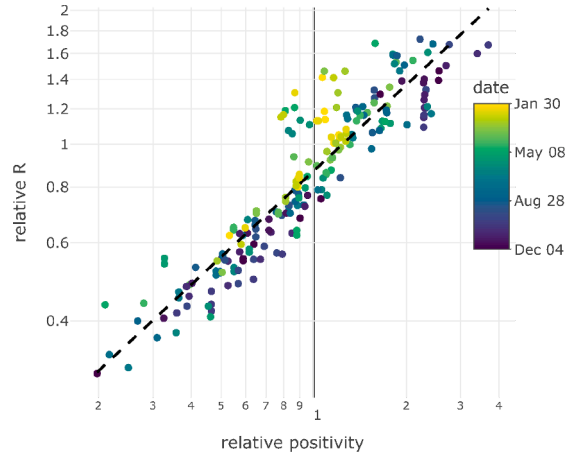


Fig. 10. Relation between age-attributed R relative to the population R and positivity by age-group relative to the overall positivity.

adults. In autumn 2021, following the start of the school term, the vast majority of new infections were transmitted by school-age children.

3.3. Positivity and r(t)

The model parameters are estimated using age-dependent survey data, therefore the differences in age-attributed R will be driven by positivity differences between age-groups. To investigate this further, we plotted the relative R against the relative positivity for each age-group at each time point (Fig. 10), note the axis are logarithmic. Relative R and relative positivity are highly correlated (Pearson correlation 0.869; CrI 0.828–0.897), however their logarithms have a substantially higher correlation (Pearson correlation 0.921; CrI 0.897–0.939), suggesting that the data is better fit by a power-law than a linear function. Regressing the logarithmic values yields $R_{\text{rel}} \propto \text{Pos}_{\text{rel}}^{0.64}$ (exponent CrI: 0.60–0.67), meaning that the variation in R is less than that in the positivity.

4. Discussion

In this paper we estimated the contribution of different age-groups to the reproduction number (R) of COVID-19 in England throughout 2021, using the ONS Infection Survey. We developed a Bayesian model which estimated a dynamic next-generation matrix comprised of a static survey-derived social-contact matrix, multiplied by a low-rank dynamic matrix. This split of the next-generation matrix is novel and extends existing work in the field, such as the EpiEstim toolbox (Cori et al., 2013), allowing us to quantify the contribution of

different age-groups. Our findings suggest that whilst R was typically highest for school-age children (5–11y and 12–17y) and lowest for the elderly (60–69y and 70y+), different dynamic patterns were observed in each age-group. For example, R for school-aged children (5–11y and 12–17y) typically rose during term-time and then fell during the school-holidays. The dynamics for young adults (18–29y) were particularly interesting, which increased relative to older adults in late-spring 2021 following the re-opening of entertainment venues. The R peaked for young adults (18–29y) in July 2021 coinciding with the period of the Euros football tournament, before rapidly dropping as the national vaccination program reached this group in August 2021.

The key data which generated the age-specific R estimates was the difference in test positivity between age-groups in the ONS Infection Survey, along with a base estimate of the social-mixing matrix. It should be noted that the positivity time-series could be fit by a model with no mixing between age-groups, which would have led to very similar estimates of $R(t)$ for each age-group, due to high correlation between the positivity time-series. However, if we tried to estimate an unconstrained social-mixing matrix to only the positivity data, we would run into the problem of parameter non-identifiability due to the high-dimensionality of the time-dependent next-generation matrix. The novel approach of our work was to constrain the form of the social-mixing matrix, using a base matrix from a social-mixing survey and adjusting it using an (estimated) rank-1 matrix *i.e.* only estimating N_{age} parameters per time-step instead of N_{age}^2. The time-dependent parameters were further constrained by modelling them as log-normal processes, thus constraining the social-mixing matrix to vary gradually with time. With these 2 constraints on the social-mixing matrix, the model parameters were identifiable in the sense that their posterior distributions were distinct from the prior distributions, and that the inter-age difference in dynamic patterns were greater than the posterior credible intervals. One of the strengths of our model is that it only requires a base estimate of the social-mixing matrix and age-stratified estimates of positivity. Therefore, the approach is translatable to other settings and diseases where pathogens are transmitted over a social-network, providing there are age-stratified estimates of infection incidence.

Our modelling approach has some limitations. Firstly, the constraints on the dynamic factor of the next-generation matrix might be too simplistic, for example, dynamic changes in relative transmissibility and susceptibility of an age-group are fully correlated. However, when we consider an intervention such as vaccination, it is likely to have a larger effect in reducing susceptibility than transmissibility. Secondly, we assume that individuals test positive for the same length of time regardless of age. Estimates of viral load kinetics suggest that older adults (> 50 years) will take 2–3 days longer (about 25 % of total time) to clear detectable virus than younger adults (<50 years), see Fig. 2 of Hay et al. (2022). Thus our model will over estimate the number of infections in older adults relative to younger adults, although this effect would be reasonably small because the relative difference in the length of time of detectable virus is small. Additionally, the latency and infectious period are assumed to be the same for all age-groups, which is a simplification given the difference in viral load kinetics with age. Finally, within the model we do not attempt to attribute difference in R to known interventions, thus we do not directly estimate the effect of interventions. For example, if we included data on the proportion of age-group vaccinated by time, we could then try to directly estimate the vaccine efficacy.

In summary, we developed a Bayesian model that could estimate the age-specific contribution to sustaining the SARS-CoV-2 epidemic in England throughout 2021. Our findings suggest that the largest contribution to the epidemic over this period in England was from school-age children (5–11y or 12–17y) or from young adults (18–29y) while the least contribution came from the elderly age-cohorts (60y+). These findings are in line with existing studies that have identified the 20–29y old as a crucial group to sustaining the epidemic across different settings over 2020 and 2021, e.g Kiem et al. (2021); Boldea et al. (2024). Our model and the results it generates are an important tool that can inform R across ages, and hence the impact of different public health policies during an epidemic. By identifying the core groups that sustain the epidemic, furthermore our work can contribute to the design and evaluation of targeted non-pharmaceutical and pharmaceutical interventions.

CRediT authorship contribution statement

Robert Hinch: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization; **Jasmina Panovska-Griffiths:** Writing – review & editing, Writing – original draft; **Christophe Fraser:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Conceptualization..

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jasmina Panovska-Griffiths is a co-author of this paper and is also an editor of this special issue,. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

RH's and JPG's work was funded by research funding from the UK Health Security Agency (UKHSA) and UK Department of Health and Social Care (DHSC) to CF. This funder had no role in the study design, data analysis, data in the interpretation, or writing of the report. The views expressed in this article are those of the authors and not necessarily those of the UKHSA or the UK DHSC.

Appendix A

A.1. Further methods

A.1.1. Alignment of age-groups

The split of age-groups in the Comix survey and the ONS Infection survey are not exactly the same. Therefore, we used the Comix survey age-groups, and mapped the ONS Infection survey to these age-groups by assuming a uniform distribution of participants by age within each age-group.

A.1.2. Alignment of reporting periods

The ONS Infection survey reporting period is bi-weekly, whilst the COG-UK reporting period is weekly (although both align to the same day of the week). When the model is fit to the data, we combine the COG-UK data in to bi-weekly reporting periods which match the ONS reporting periods.

A.1.3. Base contact matrix details

The base contact matrix (C) is taken from the Comix survey, in particular we use the matrix labelled "4. School Reopening" from September 2020. We chose this matrix since it came from a survey when there was not a national lockdown, schools were open, but some social-mixing restrictions did apply, so thus best reflected 2021. The contact matrix was normalised by dividing each element of the matrix by the principle eigenvalue. This is equivalent of shifting the prior on $\log R(0)$ and is done so that a value of $R(0) = 1$ is when the reproduction number is 1.

A.1.4. Initial conditions

The initial conditions for $\log R(0)$ and $\log \mathbf{A}(0)$ are drawn from range priors (see Appendix A.2.1). For the Alpha variant, the initial number of people in the exposed and infected compartments were fixed to match the positivity rates from the first round (split equally between fixed and exposed). For the Delta variant, there were initial no infections.

A.1.5. Seeding the Delta variant

The Delta variant was not present in England at the beginning of the study period, with the first cases being observed at the end of March 2021. Therefore, the Delta variant was seeded in to the model by externally introducing new infections (spread equally by age) on a daily basis at the end of March 2021 and in early April 2021 (see Appendix A.2.2). The daily seeding rate is an estimated model parameter with a range prior (see Appendix A.2.1).

A.1.6. Estimation of confidence intervals

In Figs. 1 and 2 we display confidence intervals for the ONS Infection Survey estimates of positivity and COG-UK estimates of proportion of each variant. Due to the large sample sizes, we can use the standard normal approximation to estimate the Wald intervals, and show the 95 % symmetric confidence interval.

A.1.7. MCMC parameter estimation

The model was implemented in RStan and samples were drawn from the posterior distribution using the default NUTS sampler. The *adapt_delta* parameter was increased to 0.90 to prevent *divergent transitions* which were observed with the default sampler settings, and the *maximum_tree_depth* parameter was increased to boost efficiency. The parameters were estimated by running 12 chains with 1,000 samples per chain in batches of 3 on a 2020 MacBook Pro, each batch took approximately 3.5 hours to run.

A.2. Table of model parameters

A.2.1. Table of priors on the model parameters

Parameter and Description	Min	Max
Log of initial reproduction number, $R(0)$	log(0.8)	log(1.2)
Log of initial age factors, $\mathbf{A}(0)$	log(0.4)	log(2.5)
S.d. of (bi-weekly) change in R, σ_R	0.0001	0.5
S.d. of (bi-weekly) change in $\mathbf{A}$, σ_A	0.0001	0.4
Maximum bi-weekly change in R	0.5	2
Maximum bi-weekly change in A	0.5	2
Susceptible *vs* transmissible factor, ρ	0.01	0.99
Seeding rate of Delta variant	0	50
Transmissibility of Delta *vs* Alpha	1.00	3.00

A.2.2. Table of the fixed model parameters

Parameter and Description	Value
Data frequency	14 days
Generation time	5.5 days
Seeding period for Delta variant	20th March–20th April

A.2.3. Table of the Stan sampling parameters

Parameter	Value
Number of MCMC chains	12
Samples per Chain	1,000
Burn-in period	500
Maximum Tree Depth	12
Adapt delta	0.90

References

Anderson, R.M., 1991. Infectious Disease of Humans: Dynamics and Control. Oxford University Press.

Bertuzzo, E., Mari, L., Pasetto, D., Miccoli, S., Casagrandi, R., Gatto, M., Rinaldo, A., 2020. The geography of COVID-19 spread in Italy and implications for the relaxation of confinement measures. Nat. Commun. 11 (1), 4264. https://doi.org/10.1038/s41467-020-18050-2

Boldea, O., Alipoor, A., Pei, S., Shaman, J., Rozhnova, G., 2024. Age-specific transmission dynamics of SARS-CoV-2 during the first 2 years of the pandemic. PNAS Nexus 3 (2), 24.

Bosetti, P., et al., 2022. Epidemiology and control of SARS-CoV-2 epidemics in partially vaccinated populations: a modeling study applied to France. BMC Med. 20, 33.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: a probabilistic programming language. J. Stat. Softw. 76 (1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S., 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178 (9), 1505–1512.

Coronavirus (COVID-19), (2024). Coronavirus (COVID-19) in the UK: UK summary. Accessed July 12, https://coronavirus.data.gov.uk

COVID-19 (2024). Genomic surveillence. Accessed July 20, https://covid19.sanger.ac.uk/lineages/raw

Davies, N.G., Klepac, P., Liu, Y., Prem, K., Jit, M., 2020. Cmmid covid-19 working group; eggo rm. age-dependent effects in the transmission and control of covid-19 epidemics. Nat. Med. 26 (8), 1205–1211.

Dekker, M.M., Coffeng, L.E., Pijpers, F.P., Panja, D., Vlas, S. J.D., 2023. Reducing societal impacts of SARS-CoV-2 interventions through subnational implementation. eLife. 12, e80819.

Diekmann, O., Gyllenberg, M., Metz, J., Thieme, H.R., 1998. On the formulation and analysis of general structured population models I: Linear theory. J. Math. Biol. 36, 349–388.

Diekmann, O., Heesterbeek, H., Britton, T., 2013. Mathematical Tools for Understanding Infectious Disease Dynamics. Princeton University Press.

Diekmann, O., Heesterbeek, J., Roberts, M.G., 2010. The construction of next-generation matrices for compartmental epidemic systems. J. R. Soc. Interface 7 (47), 873–885.

Earnest, R., et al., 2022. Comparative transmissibility of SARS-CoV-2 variants Delta and Alpha. Cell Rep. Med. 3, 100583.

Gatto, M., et al., 2020. Spread and dynamics of the Covid-19 epidemic in Italy: effects of emergency containment measures. Proc. Natl. Acad. Sci. 117 (19), 10484–10491.

Gimma, A., Munday, J.D., Wong, K., Coletti, P., Zandvoort, K.V., Prem, K., et al., 2022. Changes in social contacts in England during the Covid-19 pandemic between March 2020 and March 2021 as measured by the Comix survey: a repeated cross-sectional study. PLoS Med. 19 (3), 1003907.

Giordano, G., et al., 2020. Modelling the Covid-19 epidemic and implementation of population-wide interventions in Italy. Nat. Med. 26, 855–860.

Hay, J. A., et al., 2022. Quantifying the impact of immune history and variant on SARS-CoV-2 viral kinetics and infection rebound: A retrospective cohort study. eLife, 11, e81849.

Hinch, R., Probert, W.J.M., et al., (2020). Effective configurations of a digital contact tracing app: A report to NHSX. https://github.com/BDI-pathogens/covid-19_instant_tracing/blob/master/Report%20-%20Effective%20Configurations%20of%20a%20Digital%20Contact%20Tracing%20App.pdf

Hinch, R., Probert, W.J.M., Nurtay, A., Kendall, M., Wymant, C., Hall, M., et al., 2021. OpenABM-Covid19 – an agent-based model for non-pharmaceutical interventions against covid-19 including contact tracing. PLoS Comput. Biol. 17 (7), 1009146. https://doi.org/10.1371/journal.pcbi.1009146

Hinch, R., Panovska-Griffiths, J., Probert, W., Ferretti, L., Wymant, C., Lauro, F.D., Bya, N., Ghafari, M., Abeler-Dörner, L., Covid-19 genomics uk (cog-uk) consortium, Fraser, C., 2022. Estimating SARS-CoV-2 variant fitness and the impact of interventions in England using statistical and geo-spatial agent-based models. Phil. Trans. R. Soc. A 380, 20210304.

Hinch, R., 2024. R-package EpiAgeVar. https://github.com/BDI-pathogens/EpiAgeVar.

Hozé, N., et al., 2021. Monitoring the proportion of the population infected by SARS-CoV-2 using age-stratified hospitalisation and serological data: a modelling study. Lancet Glob. Health 6 (6), 408–e415.

Kendall, M., Ferretti, L., Wymant, C., Tsallis, D., Petrie, J., Francia, A. D., Lauro, F. D., Abeler-Dörner, L., Manley, H., Panovska-Griffiths, J., Ledda, A., Didelot, X., Fraser, C., 2024. Drivers of epidemic dynamics in real time from daily digital covid-19 measurements. Science 385 (6710), 8103.

Kiem, C. T., Bosetti, P., Paireau, J., Crépey, P., Salje, H., Lefrancq, N., Fontanet, A., Benamouzig, D., Boëlle, P.Y., Desenclos, J.C., Opatowski, L., Cauchemez, S., 2021. SARS-CoV-2 transmission across age-groups in France and implications for control. Nat. Commun. 12 (1), 6895.

Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H., Lipsitch, M., 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science 368 (6493), 860–868.

Milne, G., et al., 2021. Does infection with or vaccination against SARS-CoV-2 lead to lasting immunity?, Lancet Respir Med. 9 (12), 1450–1466.

Monod, M., et al., 2021. Age-groups that sustain resurging covid-19 epidemics in the United States. Science 371, 8372.

O'driscoll, M., et al., 2021. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature 590 (7844), 140–145.

Office for National Statistics (2022a). Coronavirus (covid-19) infection survey, UK. https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/covid19infectionsurvey

Panovska-Griffiths, J., Kerr, C.C., Stuart, R.M., Mistry, D., Klein, D.J., Viner, R.M., Bonell, C., 2020. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second covid-19 epidemic wave in the UK: a modelling study. Lancet Child Adolesc. Health 4 (11), 817–827.

Panovska-Griffiths, J., Stuart, R.M., Kerr, C.C., Rosenfield, K., Mistry, D., Waites, W., Klein, D.J., Bonell, C., Viner, R.M., 2022a. Modelling the impact of reopening schools in the UK in early 2021 in the presence of the Alpha variant and with roll-out of vaccination against SARS-CoV-2. J. Math. Anal. Appl. 514 (2), 126050.

Panovska-Griffiths, J., Swallow, B., Hinch, R., Cohen, J., Rosenfeld, K., Stuart, R.M., Ferretti, L., Lauro, F.D., Wymant, C., Izzo, A., Waites, W., Viner, R., Bonell, C., Fraser, C., Klein, D., Kerr, C.C., 2022b. Covid-19 genomics UK (COG-UK) consortium. statistical and agent-based modelling of the transmissibility of different SARS-CoV-2 variants in England and impact of different interventions. Philos. Trans. A Math. Phys. Eng. Sci. 380, 20210315.

Pei, S., Yamana, T.K., Kandula, S., Galanti, M., Shaman, J., 2021. Burden and characteristics of covid-19 in the United States during 2020. Nature 598 (7880), 338–341.

Pickles, M., Cori, A., Probert, W.J.M., 2021. Popart-IBM, a highly efficient stochastic individual-based simulation model of generalised HIV epidemics developed in the context of the hptn 071. PLoS Comput. Biol. 17, 1009301.

Rozhnova, G., et al., 2021. Model-based evaluation of school and non-school-related measures to control the Covid-19 pandemic. Nat. Commun. 12 (1), 1614.

Saad-Roy, C.M., et al., 2020. Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years. Science 370 (6518), 811–818.

Sonabend, R., et al., 2021. Non-pharmaceutical interventions, vaccination, and the SARS-CoV-2 Delta variant in England: a mathematical modelling study. Lancet 398, 1825–1835.

Sorensen, R.J., et al., 2022. Variation in the covid-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. Lancet 399, 1469–1488.

Viana, J., et al., 2021. Controlling the pandemic during the SARS-CoV-2 vaccination roll-out. Nat. Commun. 12 (1), 3674.

World Health Organization (2022b). Tracking SARS-CoV-2 Variants. World Health Organization. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants