Community prevalence of SARS-CoV-2 in England during April to November 2020: Results from the ONS Coronavirus Infection Survey

Koen B Pouwels, PhD[1,2*†], Thomas House, PhD[3,4*], Emma Pritchard, MSc [2,5] Julie V Robotham, PhD[6], Paul J Birrell, PhD[6,7], Andrew Gelman, PhD[8], Karina-Doris Vihta,[2,5] Nikola Bowers, [9] Ian Boreham,[9] Heledd Thomas, MSc[9], James Lewis, BA[9], Iain Bell, BSc[9], John I Bell, MD[10], John N Newton, FRCP[11], Jeremy Farrar, PhD,[12] Ian Diamond, PhD[9], Pete Benton, MSc[9], Ann Sarah Walker, PhD[2,5,13,14] and the COVID-19 Infection Survey team

[*] contribution considered equal

[†] Correspondence to koen.pouwels@ndph.ox.ac.uk

[1] Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford, UK

[2] The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford, Oxford, UK.

[3] Department of Mathematics, University of Manchester, Manchester, UK

[4] IBM Research, Hartree Centre, Sci-Tech Daresbury, UK

[5] Nuffield Department of Medicine, University of Oxford, Oxford, UK

[6] National Infection Service, Public Health England, London, UK

[7] MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Cambridge, UK

[8] Department of Statistics, Columbia University, New York, NY, United States

[9] Office for National Statistics, Newport, UK

[10] Office of the Regius Professor of Medicine, University of Oxford, Oxford, UK

[11] Health Improvement Directorate, Public Health England, London, UK

[12] Wellcome Trust, London, UK

[13] The National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK

[14] MRC Clinical Trials Unit at UCL, UCL, London, UK

See Acknowledgements for the Coronavirus Infection Survey team

Abstract

Background: Decisions regarding the continued need for control measures to contain the spread of SARS-CoV-2 rely on accurate and up-to-date information about the number of people and risk factors for testing positive. Existing surveillance systems are not based on population samples and are generally not longitudinal in design.

Methods: Samples were collected from individuals aged 2 years and over from a representative sample of private households from England using repeated cross-sectional household surveys with additional serial sampling and longitudinal follow-up. Participants completed a questionnaire and nose and throat swabs were taken. The percentage of individuals testing positive for SARS-CoV-2 RNA was estimated over time using dynamic multilevel regression and post-stratification, to account for potential residual non-representativeness. Potential changes in risk factors for testing positive over time were also evaluated.

Findings Between 26 April and 1 November 2020, in total, results were available from 1,191,170 samples from 280,327 individuals, of which 5,231 were positive overall from 3,923 individuals. The percentage of people testing positive for SARS-CoV-2 changed substantially over time, with an initial decrease between end of April and June, followed by low levels during the summer, before marked increases occurred starting end of August 2020. Having a patient-facing role and working outside your home were important risk factors for testing positive in the first period but not (yet) in the second period of increased positivity rates, whereas age (young adults) was an important initial driver of the second period of increased positivity rates. A substantial proportion of infections were in individuals not reporting symptoms (45%-68%, dependent on calendar time).

Interpretation Important risk factors for testing positive varied substantially between the initial and second periods of higher positivity rates, and a substantial proportion of infections were in individuals not reporting symptoms, indicating that continued monitoring for SARS-CoV-2 in the community will be important for managing the epidemic moving forwards.

Funding

This study is funded by the Department of Health and Social Care. KBP, ASW, EP and JVR are supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) (NIHR200915). AG is supported by U.S. National Institute of Health and Office of Naval Research. ASW is also supported by the NIHR Oxford Biomedical Research Centre and by core support from the Medical Research Council UK to the MRC Clinical Trials Unit [MC_UU_12023/22] and is an NIHR Senior Investigator. We gratefully acknowledge the support of the Huo Family Foundation. The views expressed are those of the authors and not necessarily those of the National Health Service, NIHR, Department of Health, or PHE.

Research in context

Evidence before this study

Unprecedented control measures, such as national lockdowns, have been widely implemented to contain the spread of SARS-CoV-2. Decisions regarding the continued need for social distancing measures in the overall population, specific subgroups and geographic areas heavily rely on accurate and up-to-date information about the number of people and risk factors for testing positive. We searched PubMed and medRxiv and bioRxiv preprint servers up to 15 November 2020 for epidemiological studies using the terms "SARS-CoV-2" and "prevalence" or "incidence" without data or language restrictions. Most studies were small or had only information about current presence of the virus for a small subset of patients, or used data not representative of the community, such as hospital admissions, deaths or self-reported symptoms. Large population-based studies, such as the current study, are required to understand risk factors and the dynamics of the epidemic.

Added value of this study

This is the first longitudinal community survey of SARS-CoV-2 infection at national and regional levels in the UK. With more than 1,000,000 swabs from almost 300,000 individuals this ongoing study provides robust evidence that the percentage of individuals from the general community in England testing positive for SARS-CoV-2 clearly declined between end of April and June 2020, followed by consistently low levels during the summer, before marked increases started end of August 2020. Risk factors for testing positive varied substantially between the initial and second periods of higher positivity rates, with having a patient-facing role and working outside your home being important risk factors in the first period but not (yet) in the second period of high positivity, and age (young adults) being an important driver of the second period of increased positivity rates. Positive tests commonly occurred without symptoms being reported.

Implications of all the available evidence

This survey demonstrates that community supervised self-swabbing RT-PCR-based surveillance is achievable and practical. This survey may serve as a model for other countries and potential future pandemics. The observed decline in the percentage of individuals testing positive adds to the increasing body of empirical evidence and theoretical models that suggest that the lockdown imposed on 23 March 2020 in England was associated, at least temporarily, with a decrease in infections. Important risk factors for testing positive varied substantially between the initial and second periods of higher positivity rates, and a substantial proportion of infections were in individuals not reporting symptoms, indicating that continued monitoring for SARS-CoV-2 in the community will be important for managing the epidemic moving forwards. Using multilevel regression and poststratification to account for potential residual non-representativeness of the sample, the survey provided early warnings that certain regions, such as the North West of England, were likely going to experience increases in hospital admissions and deaths.

Introduction

Since severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) started causing severe respiratory illness in Wuhan, China, in late 2019,[1] as of 1 November, there have been nearly 46 million confirmed cases and 1.2 million deaths reported to the WHO.[2] Control measures, such as national lockdowns, have been widely implemented to contain the spread of the virus in a, at least temporarily successful,[3-5] attempt to prevent the collapse of healthcare systems and even larger numbers of deaths. Although such measures are important for control of the pandemic, they also affect the economy, unemployment rates, and global supply-chains.[6,7] Politicians continuously make difficult decisions between continuing strict control measures or relaxing them in ways that would be safe enough from a public health perspective yet beneficial more broadly across society.

Importantly, early detection of population subgroups driving new increases in infections is crucial to potentially tailor interventions or messaging without having to implement drastic measures affecting the whole society.

There are several reasons why risk factors may vary over time. First, behaviour and contact patterns of subgroups change over time without intervention, e.g. students starting university. Adherence to non-mandatory infection prevention measures may reduce more over time among subgroups with a low risk of COVID-19-related hospital admission and death than those that are more vulnerable. Moreover, subgroups that have been disproportionally affected in a first wave may have acquired sufficient immunity and may have better access to effective measures that reduce the risk of infection making them less likely to acquire a new infection during a second wave.

Here, we use data from the Office of National Statistics (ONS) Coronavirus Infection Survey (CIS). This ongoing large national survey with more than 1 million swab results to 1 November is designed to be representative of the target population, offering a unique opportunity to identify risk factors that are driving recent new increases in the positivity rate, as well as investigating the proportion of individuals testing positive that do not report symptoms, potential false-positivity rate, and other factors that can directly inform policy around COVID-19-related control measures. We used Bayesian dynamic multilevel regression and poststratification (MRP) to account for any residual unrepresentativeness, a potential problem often ignored with surveillance data.

Methods

Data were collected between 26 April and 1 November 2020 from individuals from randomly selected private households from address lists and previous ONS surveys to provide a representative sample of the population of England (details on sampling design in Supplementary file). Individuals aged 2 years and older living in private households were eligible. If one or more individuals from a household agreed to participate, a study worker visited the household and directly collected information from individuals about any symptoms (current until 23 July, then in the last 7 days before the visit) and contacts, together with demographic information. The study worker provided instructions on how to self-swab the nose and throat and monitored the self-swabbing, which has been shown to be comparable or even more sensitive than swabs performed by healthcare workers.[9] Parents/carers took swabs from children under 12 years old. The nose and throat self-swabs were couriered directly to the UK's national Lighthouse laboratories at Milton Keynes (National Biocentre) (from 26 April) and Glasgow (from 16 August), where the samples were tested using identical methodology for the presence of SARS-CoV-2 (3 gene targets, N protein, S protein and ORF1ab) using reverse transcriptase polymerase chain reaction (RT-PCR) as part of the national testing programme,[10] using the Thermo Fisher TaqPath RT-PCR COVID-19 Kit, analysed using UgenTec Fast Finder 3.300.5 (TagMan 2019-nCoV Assay Kit V2 UK NHS ABI 7500 v2.1). The assay Plugin contains an

Assay specific algorithm and decision mechanism that allows conversion of the qualitative amplification Assay PCR raw data from the ABI 7500 Fast into test results with minimal manual intervention. Samples are called positive in the presence of at least single N gene and/or ORF1ab but may be accompanied with S gene (1, 2, or 3 gene positives). S gene is not considered a reliable single gene positive (as of mid-May 2020).

After the first visit, participants were asked whether they were willing to participate in further follow-up visits: either every week for the first 5 weeks of the study, or this and then monthly thereafter. The study protocol and questionnaires are available at https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey (accessed 12/11/2020).

The survey was designed to test 150,000 individuals every fortnight across England in October, to provide 15,000-20,000 individuals in each of the nine governmental office regions to provide approximately a 0.1%, 0.2% and 0.5% margin of error on a 0.1%, 0.5% and 2% prevalence, respectively.

The survey has been reviewed and given ethical approval by South Central - Berkshire B Research Ethics Committee (20/SC/0195).

Trend in proportion of positive tests over time

We analysed the proportion of the private-residential population testing positive for SARS-CoV-2 from nose and throat swabs over time using Bayesian dynamic MRP.[11,12] MRP was used to correct for any residual non-representativeness in terms of age, sex and region. In several empirical and simulation studies MRP was superior at both the national and regional levels compared to classical survey weighted and unweighted approaches, including when using small sample sizes.[11-16] Partial pooling through the use of random effects in the multilevel model ensures stable estimates can be obtained for subnational levels from relatively small samples that would be problematic using more traditional survey-weighting approaches.[11-16] MRP consists of two steps. First, a multilevel regression model is used to generate the outcome of interest as a function of (socio)demographic and geographic variables. Next, the resulting outcome estimates for each demographic-geographic respondent type are poststratified by the percentage of each type in the actual overall population.[11]

We used a Bayesian multilevel generalised additive regression model to model the swab test result (positive/negative) as a function of age, sex, time and region. We did not post-stratify for other factors (e.g. ethnicity) because reliable estimates in the target population where not available and a model for the full period did not converge with ethnicity in the model (divergent transitions). A model including ethnicity for the last 7 weeks did converge and showed similar estimates and trend as the main model (Figure S1). Besides the 9 regions in England, we also took into account the fact that certain local authorities within regions (boosted areas) were purposefully oversampled at the end of July (see supplementary file). Because there were very few missing values (≤1%) in these factors, we restricted all analyses to observations with non-missing data. A complementary log-log link was used due to the ability to interpret regression coefficients as arising from an infection process with varying levels of exposure (see Supplementary File).[17] MRP models with random effects for individual participant and/or household nested within region did not converge. Therefore MRP models were run with only a random intercept for region (including separate levels for the boosted areas within region, i.e. Yorkshire & The Humber non-boosted and Yorkshire & The Humber boosted), without a random intercept for participant and/or household. However, a model with only one participant sampled from each household gave similar results with somewhat wider 95% credible intervals mainly due to the smaller sample size (Figure S2). Time, measured in days since the start of the study (26 April 2020), was modelled using thin-plate splines and allowed to vary by

region. We set *k*, the number of basis functions, to 10 to control the smoothness of the fitted function.[18] We used a normal prior with location set to 4 for the standard deviation of the smooth. Very similar results were obtained when using different values for k (Figure S3A) or different priors for the standard deviation of the smooth (Figure S3B).Subsequently, we poststratified the resulting positivity estimates for each demographic-geographic respondent type by the percentage of each type in the overall population and in each region.

Because the effect of potential risk-factors may change over time, and it was not feasible in terms of run-time and available cpu to fit a model with a much more flexible thin-plate spline for the entire period 26 April-1 November, we also ran an MRP model using the most recent 7 weeks. This analysis was performed using the rstanarm package in R version 3.6.1.[19]

Time-varying risk factors

To assess whether particular subgroups were more likely to test positive for SARS-CoV-2 during the first wave in England we performed a multilevel regression analysis (without poststratification) on the data between 26 April and 28 June 2020 including variables on which we did not post-stratify: work location, having a job that directly involved patients/care-home residents, ethnicity, household size, and number of children in the household. Given the short timescale included, and the fact that questions were not always asked at every visits, we carried non-missing data forward and backwards to adjacent visits with missing data. After this, there were very few missing values (≤1%) so we again restricted all analyses to observations with non-missing data only. Results are shown in Table S3.

We evaluated to what extent different factors were potentially driving recent increases in the positivity rate. Given that age appeared to be such a strong factor in driving the increase (see Results), all other factors were subsequently stratified by age (<35 and 35+ years). We evaluated the same factors as for the first wave (through 28 June) in generalised additive models with thin-plate splines that varied by each level of the factor of interest. These models additionally included a random intercept for region to account for any regional differences. As it was not possible to fit all factors with these time interactions in one model, and given the limited evidence of confounding (Table S3), we fitted separate models for each factor of interest.

Presence of symptoms among those testing positive

To evaluate the number of positive tests where participants reported symptoms around time of visit (same visit, visit before or after), or no symptoms around the time of visit, we used the same MRP model as for the overall positivity rate.

To assess the impact of potential false positive tests we classified each positive into 3 categories:[20]

i) 'Higher evidence'; two or three genes detected (irrespective of cycle threshold (Ct) value).
ii) 'Moderate evidence'; single gene detections if a) the Ct value was <97.5th percentile of 'higher evidence' positives (<34) or b) there was a higher pre-test probability of infection, i.e. any symptoms at or around the test (visit before or after) or reporting working in patient-facing healthcare role or resident-facing care home role.

iii) 'Lower evidence'; all other positives, which by definition were all in asymptomatic individuals not having patient- or resident-facing roles with a single gene detected with Ct ≥34.

Results

Between 26 April and 1 November 2020, in total, results were available from 1,191,170 nose and throat swabs from 280,327 individuals, of which 5,231 were positive overall from 3,923 individuals in 3,056 households. The study is still ongoing and many participants were only recruited recently to achieve the target sample size in October; nevertheless the median number of visits per individual was 4 (interquartile range 3-5, max 13). Table S1 shows that, of those enrolled sufficiently early to have multiple study visits before 1 November, the vast majority had at least 5 study visits.

Characteristics of participating individuals are shown in Table S2. Representativeness of the sample was visualised by plotting proportions of the sample within each region and age- and sex-category by comparing with known distributions for individuals living in private households in England (Figure S4). Small under/overrepresentation of certain groups, such as individuals aged 2-11 being slightly underrepresented, was corrected for using dynamic MRP with poststratification performed on a daily basis. Positivity rates dropped to consistently low levels during the summer before increasing markedly again starting end of August (Figure 1). When restricting the analysis to the most recent 7 weeks in order to make the flexible spline more responsive to recent changes, the increase in positivity rates appeared to start levelling off at the end of October, before the lockdown was implemented on 5 November.

Observed patterns in positivity rate were similar between participants reporting symptoms and those not reporting symptoms, although in October the positivity rate among those reporting symptoms started to increase less steeply (Figure 2A). The modelled percentage of positives with reported symptoms around the test was lowest around mid-July (32%) and highest around beginning of October (55%). The increase in positivity starting end of August 2020 was almost entirely due to high evidence positives, although the levels of moderate and to a lesser extent low evidence positives started to increase slightly in September 2020 as well (Figure 2B). People may have become infected with lower viral loads and fewer symptoms during the summer, when there were small increases in low-evidence positives and few people reported symptoms when testing positive, but with higher viral loads in September potentially leading to a higher proportion of cases with symptoms.

Positivity rates showed marked regional differences, with increases in late August-October largely occurring in the North of England and to a lesser extent the Midlands (Figure 1). The most important factor underlying the observed sharp increase in positivity was age, with earlier and greater increases apparent in younger adults (Figure 3) as also evident from results from a model categorising age (Figure S5). These figures also show that near the end of October the prevalence started to decrease in young adults. Importantly, there was clear diffusion of risk from initial increases in younger age groups at lower risk of hospitalisation and death, into older ages at higher risk.

While working outside their home and in patient-facing healthcare roles were clear risk factors during the initial period of high positivity, as was contact with hospitals (26 April to 28 June, Table S3), there was no evidence that those working outside their home, working in patient-facing roles or with hospital contact were driving initial increases after the summer (Figure S6-8). Non-White

ethnicity was also associated with greater positivity rates during the initial period but not the initial increases after the summer (Figure S9 and Figure 4). While the probability of testing positive increased in all age-groups after the summer, the increase was particularly pronounced in individuals aged <25 that shared a household with 17-24 year olds (Figure 5).

Discussion

Here we demonstrate substantial changes over time in the percentage of people in private-residential households in the community in England testing positive for SARS-CoV-2 using RT-PCR testing, with an initial decrease between end of April and June 2020, followed by consistently low levels during the summer, before marked increases between end of August and 1 November 2020. Our estimates have been regularly updated and shared with the UK Government and Scientific Advisory Group for Emergencies (SAGE) sub-group Scientific Pandemic Influenza sub-group on Modelling (SPI-M) to directly inform decisions about potential changes to the current alert level or relaxation of certain restrictions. Notably, we found that a substantial proportion (45%-68%, dependent on calendar time) of individuals that tested positive did not report any symptoms on the day of the visit or at visits before or after the swab was taken.

The Bayesian dynamic multilevel generalised additive models are useful tools for monitoring the effect of different factors on positivity rates over time. In particular, they show that the epidemic restarted in young people, and that factors associated with an increased risk of testing positive during the initial high-positivity period in April-May 2020, such as working outside the home and having a job with direct patient contact were not important drivers of initial increases after the summer.

While false-positives may be a concern when prevalence is low, the low positivity at the end of June (0.05%) is also reassuring, since it indicates that the specificity of the test used in the national UK programme is very high. A test specificity lower than 99.95% would lead to observed positivity rates above 0.05%, even in the purely hypothetical situation that the virus was not circulating in June.

Comparison with other studies

According to a recent systematic review of population-based prevalence surveys from 19 countries, during the COVID-19 pandemic the vast majority of studies (n=25, 68%) reported only antibody testing, with many of those studies having a high risk of bias. The few PCR-based surveys, such as the current study, were generally found to have a low risk of bias, and importantly provide information about people currently being infected and potentially able to transmit the virus.[21]

An important advantage of our population-based study is that it can detect increases in the positivity rate potentially earlier and more systematically than surveillance based on confirmed cases, hospital admissions or deaths (https://coronavirus.data.gov.uk/, accessed 12/11/2020).[22,23] This is likely especially the case when new increases initially occur in a subgroup of the population with low risk of hospitalisation and death, but does contribute to transmission including if asymptomatic,[24] as observed after the summer with the increase in positives among young adults. For example, the sharp rise in cases in young adults that started in August in the North West, that subsequently resulted in increases in other age-groups as well, preceded sharp rises in intensive care (ICU) bed occupancy by COVID-19 patients in larger cities in the North West, such as Manchester where 35% of beds were occupied by COVID-19 patients on 22 October.[25]

Furthermore, interpretation of changes in incidence and positivity rate from tests that are taken for contact tracing or clinical cases is likely confounded by substantial changes in testing practice over time. Our study is based on a representative sample of the population, with further correction for residual non-representativeness using MRP, thereby preventing difficulties with interpretation due to changes in testing practice.

There are a few other studies that aimed to assess the prevalence of SARS-CoV-2 infection in the general population. A repeated cross-sectional population-based study from England also found a similar decline in the prevalence among the general population between 1 May and 1 June.[26] Another cross-section from that study showed also an increase in in the prevalence in September.[27] Among individuals that tested positive in that study, the percentage reporting no symptoms varied between 50-81% in different cross-sections.[27] A study from Vo, an Italian town with a population of 3275 individuals, found that the percentage of those who tested positive that did not report any symptoms was 41.0%-44.8%.[28] As part of a larger study from Iceland, where participants were recruited via an open invitation, which may bias the sample towards people with symptoms, 57% of individuals testing positive reported having symptoms, although 29% of individuals testing negative also reported having symptoms.[29] A recent meta-analysis of studies focusing on close contacts of confirmed COVID-19 cases suggested that only 17% (95% CI 14%-20%) of infected individuals are asymptomatic.[30] However, informing participants that they were recently in close contact with a confirmed COVID-19 case, may result in recall bias and overestimate the true prevalence of symptoms among a representative sample of infected persons. Although we may have underestimated the true prevalence of symptoms among SARS-CoV-2 cases in the community, partly due to asking about current symptoms at visits through 23 July (meaning that very transient symptoms only occurring between visits would have been missed) and symptoms in the last 7 days thereafter, our study adds to the growing evidence that a substantial proportion of SARS-CoV-2 in the community may be asymptomatic.[31,32,33]

This survey demonstrates that community supervised self-swabbing RT-PCR-based surveillance is achievable and practical. It facilitates early detection of changes in the epidemic that are not driven by changes in testing, estimation of prevalence and incidence, evaluation of time-varying risk factors of testing positive, as well as changes in viral burden.[20] This survey may serve as a model for other countries and potential future pandemics.

Limitations of this study

An important limitation of this study is that the number of people in the community that test positive is low, limiting power and leading to relatively large uncertainty around estimates, and meaning that our multilevel regression model was not able to incorporate likely correlation within households. However, sensitivity analyses suggested that within-household clustering did not have a large impact on our results, and, assuming the households we sampled are representative of households in general, our estimates will still reflect positivity rates in the target population as a whole.

Furthermore, while we adjusted for potential non-representativeness in terms of age, sex and region, there may be other factors for which we do not have detailed information about population distributions that also are associated with testing positive. For example, there was modest underrepresentation of non-white ethnicity, potentially leading to a small underestimation of the prevalence. Furthermore (lack of) associations with testing positive may be due to residual confounding. We did forwards and backwards imputation for missing data, reflecting the relatively short timescales of the study.

Another limitation is that, in the absence of a true gold standard, we do not know the test sensitivity and specificity, making it difficult to assess what the true prevalence is. However, as detailed above the true specificity is likely very close to 100%. The data cannot inform about the test sensitivity without providing a very informative prior on the true prevalence.[34] Whilst self-swabbing was monitored by study workers, and is used very widely, this could still lead to underestimates of prevalence. However, this should not affect trends over time.

Conclusions

In a rapidly evolving epidemic where ongoing surveillance is essential to guide public health response Bayesian dynamic multilevel regression and poststratification is a powerful tool to ensure population-representative estimates can be obtained. Specifically it showed that the percentage of individuals from the community in England testing positive for SARS-CoV-2 declined between 26 April and 28 June 2020, remained approximately stable for much of the summer before increasing again from the end of August through October. Important risk factors for testing positive varied substantially between the initial and second periods of higher positivity rates, and a substantial proportion of infections were in individuals not reporting symptoms, indicating that continued monitoring for SARS-CoV-2 in the community will be important for managing the epidemic moving forwards.

Contributors: The study was designed and planned by SW, JF, JB, JN, IB, ID, PB, KBP, and JVR. KBP, TH, EP, KDV, NB, IB, HT, JL and ASW contributed to the statistical analysis. KBP drafted the manuscript and all authors contributed to interpretation of the data and results and revised the manuscript. KBP and ASW are the guarantors of the study. All authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work. The corresponding authors attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Declaration of interest: We declare no competing interests.

Data sharing: Data sharing: De-identified study data are available for access by accredited researchers in the ONS Secure Research Service (SRS) for accredited research purposes under part 5, chapter 5 of the Digital Economy Act 2017. For further information about accreditation, contact Research.Support@ons.gov.uk or visit the SRS website

Acknowledgements

COVID-19 Infection Survey Team:

COVID-19 Infection Survey Team:

Office for National Statistics: Iain Bell, Ian Diamond, Alex Lambert, Pete Benton, Emma Rourke, Stacey Hawkes, Sarah Henry, James Scruton, Peter Stokes, Tina Thomas.

Office for National Statistics, Analysis: John Allen, Russell Black, Heather Bovill, David Braunholtz, Dominic Brown, Sarah Collyer, Megan Crees, Colin Daglish, Byron Davies, Hannah Donnarumma, Julia Douglas-Mann, Antonio Felton, Hannah Finselbach, Eleanor Fordham, Alberta Ipser, Joe Jenkins, Joel Jones, Katherine Kent, Geeta Kerai, Lina Lloyd, Victoria Masding, Ellie Osborn, Alpi Patel, Elizabeth Pereira, Tristan Pett, Melissa Randall, Donna Reeve, Palvi Shah, Ruth Snook, Ruth Studley, Esther Sutherland, Eliza Swinn, Heledd Thomas, Anna Tudor, Joshua Weston.

Office for National Statistics, Secure Research Service: Shayla Leib, James Tierney, Gabor Farkas, Raf Cobb, Folkert van Galen, Lewis Compton, James Irving, John Clarke, Rachel Mullis, Lorraine Ireland, Diana Airimitoaie, Charlotte Nash, Danielle Cox, Sarah Fisher, Zoe Moore, James McLean, Matt Kerby.

University of Oxford, Nuffield Department of Medicine: Ann Sarah Walker, Derrick Crook, Philippa C Matthews, Tim Peto, Emma Pritchard, Nicole Stoesser, Karina-Doris Vihta, Alison Howarth, George Doherty, James Kavanagh, Kevin K Chau, Stephanie B Hatch, Daniel Ebner, Lucas Martins Ferreira, Thomas Christott, Brian D Marsden, Wanwisa Dejnirattisai, Juthathip Mongkolsapaya, Sarah Hoosdally, Richard Cornall, David I Stuart, Gavin Screaton.

University of Oxford, Nuffield Department of Population Health: Koen Pouwels.

University of Oxford, Big Data Institute: David W Eyre.

University of Oxford, Radcliffe Department of Medicine: John Bell.

Oxford University Hospitals NHS Foundation Trust: Stuart Cox, Kevin Paddon, Tim James.

University of Manchester: Thomas House.

Public Health England: John Newton, Julie Robotham, Paul Birrell.

IQVIA: Helena Jordan, Tim Sheppard, Graham Athey, Dan Moody, Leigh Curry, Pamela Brereton

National Biocentre: Ian Jarvis, Kirsty Howell, Bobby Mallick, Phil Eeles.

Glasgow Lighthouse Laboratory: Jodie Hay, Harper Vansteenhouse.

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395:497-506.
2. WHO. COVID-19 Weekly Epidemiological Update. 3 November 2020. https://www.who.int/publications/m/item/weekly-epidemiological-update---3-november-2020. Accessed 12 November 2020.
3. Salje H, Kiem CT, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, et al. Estimating the burden of SARS-CoV-2 in France. *Science* 2020;eabc3517
4. Jarvis CI, Van Zandvoort K, Gimma A, Prem K, Klepac P, Rubin GJ, et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med* 2020;18(1):124.
5. Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *Lancet Public Health* 2020;5:E375-E385.
6. Guan D, Wang D, Hallegatte S, Davis SJ, Huo J, Li S, et al. Global supply-chain effects of COVID-19 control measures. *Nat Hum Behav* 2020;4:577-587.
7. U.S. Department of Labour. Unemployment Insurance Weekly Claims Report. 4 June 2020. https://www.dol.gov/sites/dolgov/files/OPA/newsreleases/ui-claims/20201165.pdf. Accessed 12 November 2020.
8. HM Government. COVID Alert Levels. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/884352/slides_-_11_05_2020.pdf. Accessed 12 November 2020.
9. Kojima N, Turner F, Slepnev A, Demig L, Kodeboyina S, Klausner JD. Self-collected oral fluid and nasal swab specimens demonstrate comparable sensitivity to clinician-collected nasopharyngeal swab specimens for the dection of SARS-CoV-2. Clin Infect Dis 2020;ciaa1589;https://doi.org/10.1093/cid/ciaa1589.
10. Office of National Statistics. COVID-19 Infection Survey (Pilot). https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation. Accessed 12 November 2020.
11. Gelman A, Little TC. "Poststratification into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology* 1997;23(2):127-35.
12. Gelman A, Lax J, Phillips J, Gabry J, Trangucci R. Using multilevel regression and poststratification to estimate dynamic public opinion. 2018. http://www.stat.columbia.edu/~gelman/research/unpublished/MRT(1).pdf. Accessed 12 November 2020.
13. Downes M, Carlin JB. Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: a simulation study. *Biom J* 2020; 62(2):479-91.
14. Warshaw C, Rodden J. How should we measure district-level public opinion on individual issues? *J Politics* 2012; 74(1): 203-19.
15. Kennedy L, Gelman A. Know your population and know your model: using model-based regression and post-stratification to generalize findings beyond the observed sample. *ArXiv* 2020; 1906.11323v2.
16. Si Y, Trangucci R, Gabry JS, Gelman A. Bayesian hierarchical weighting adjustment and survey inference. *ArXiv* 2020; 1707.08220.

17. McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society B* 1980;42(2):109–42.
18. Wood, SN. Thin plate regression splines. *J R Stat Soc Series B Stat Methodol* 2003;65: 95–114.
19. Goodrich B, Gabry J, Ali I, Brilleman S. "rstanarm: Bayesian applied regression modeling via Stan." 2020; R package version 2.19.3, https://mc-stan.org/rstanarm.
20. Walker AS, Pritchard E, House T, Robotham JV, Birrell PJ, Bell I, et al. Viral load in community SARS-CoV-2 cases varies widely and temporally. MedRxiv 2020; https://doi.org/10.1101/2020.10.25.20219048
21. Franceschi VB, Santos AS, Glaeser AB, Paiz JC, Caldana GD, Lessa CLM, et al. Population-based prevalence surveys during the COVID-19 pandemic: a systematic review. MedRxiv 2020; https://doi.org/10.1101/2020.10.20.20216259
22. Office of National Statistics. Deaths registered weekly in England and Wales provisional. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregisteredweeklyinenglandandwalesprovisional/weekending30october2020. Accessed 12 November 2020.
23. Public Health England. Weekly coronavirus disease 2019 (COVID-19) surveillance report summary of COVID-19 surveillance system. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/888254/COVID19_Epidemiological_Summary_w22_Final.pdf. Accessed 12 November 2020.
24. Chau NVV, Thanh Lam V, Thanh Dung N, Yen LM, Minh MNQ, Hung LM, et al. The natural history and transmission potential of asymptomatic SARS-CoV-2 infection. Clin Infect Dis. 2020; doi:10.1093/cid/ciaa711.
25. Manchester Evening News. True intensive care unit figures. https://www.manchestereveningnews.co.uk/news/greater-manchester-news/true-intensive-care-unit-figures-19148243. Accessed 14 November 2020.
26. Riley S, Ainslie KEC, Eales O, Jeffrey B, Walters CE, Atchinson CJ, et al. Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study. MedRxiv 2020; https://doi.org/10.1101/2020.07.10.20150524.
27. Riley S, Ainslie KEC, Eales O, Walters CE, Wang H, Atchison C, et al. High prevalence of SARS-CoV-2 swab positivity in England during September 2020: interim report of round 5 of REACT-1 study. MedRxiv 2020; https://doi.org/10.1101/2020.09.30.20204727.
28. Lavezzo E, Franchin E, Ciavarella C, Cuomo-Dannenburg G, Barzon L, Del Vecchio C, et al. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo. *Nature* 2020;584:425-9.
29. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med 2020;382:2302-15
30. Byambuasuren O, Cardona M, Bell K, Clark J, McLaws M, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *MedRxiv* 2020; doi: https://doi.org/10.1101/2020.05.10.20097543
31. Poletti P, Tirani M, Cereda D, Trentine F, Guzzetta G, Sabatino G, et al (2020). Probability of symptoms and critical diseases after SARS-CoV-2 infection. arRxiv 2020; 2006.08471
32. Pollan M, Perez-Gomez B, Pastor-Barriuso R, Oteo J, Hernan MA, Perez-Olmeda M, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet 2020;396:P535-544.
33. Kasper MR, Geibe JR, Sears CL, Riegodedios AJ, Luse T, Von Thun AM, et al. An outbreak of Covid-19 on an aircraft. N Engl J Med 2020; https://doi.org/10.1056/NEJMoa2019375.

34. Lewis IF, Torgerson PR (2012). A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. Emerg Themes Epidemiol 2012;9(1):9.

Figure 1. Percentage of population living in private households testing positive for SARS-CoV-2 over time in England and the 9 regions of England. Shaded areas are 95% credible intervals. The blue curve is from a model fitted on data from the entire period (26 April – 1 November), while the red curve is from a model fitted on data from the last 7 weeks up to 1 November.

Figure 2. A: Percentage of population living in private households testing positive for SARS-CoV-2 with and without reporting symptoms; **B**:Percentage of population living in private households testing positive for SARS-CoV-2 stratified by high, moderate and low evidence positivity. Shaded areas are 95% credible intervals.

Figure 3. Modelled estimates (posterior medians) of the distribution of positive SARS-CoV-2 tests across age over time. Note that different scales are being used for each region.

Figure 4. Percentage of population living in private households testing positive for SARS-CoV-2 stratified by ethnicity and age (<=34 and >34 years of age). Shaded areas are 95% credible intervals.

Figure 5. Percentage of population living in private households testing positive for SARS-CoV-2 stratified by their household composition. HH=household; Preschool HH = household with a preschool child (< 5 years of age); Primary HH = household with a primary school child (<= school year 6, <=11/12 years of age); Secondary HH = household with a secondary school child (<= school year 11, <=16/17 years of age). The first part of the title of each plot indicates the type of household (e.g. Preschool HH) and the second part indicates the age of the individual (e.g. age <25). Households were classified according to the following hierarchy: anyone aged 17-24 (17-24 HH); anyone in secondary school age (Secondary HH); anyone in primary school age (Primary HH); anyone in preschool age (Preschool HH); anyone aged 50+ (50+HH); all 25-49 (25-49 HH). Shaded areas are 95% credible intervals.

Figure 1

Click here to access/download;Figure;regions_prevalence_plot.png

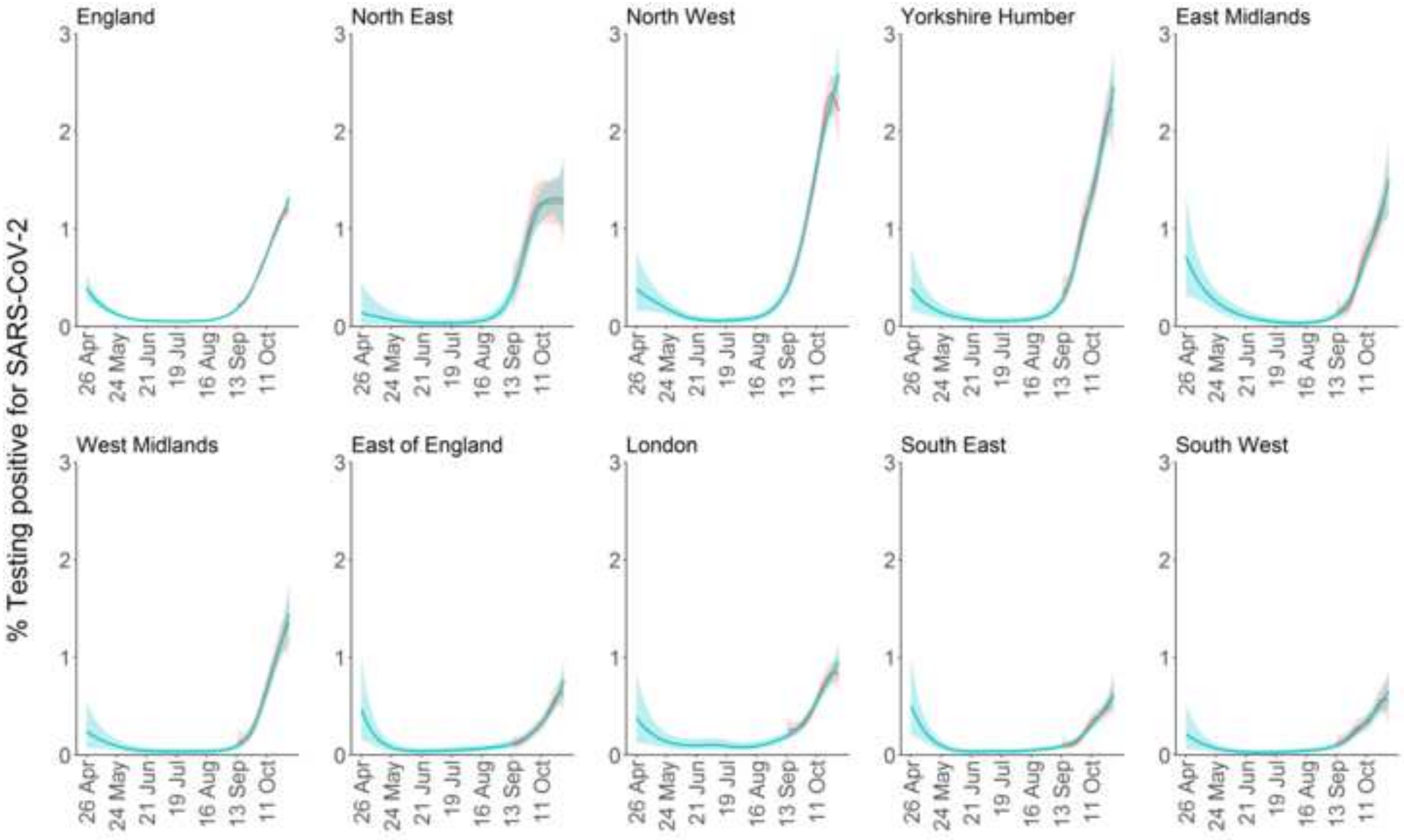


Figure 2

Click here to access/download;Figure;symp_evidence_combined.png

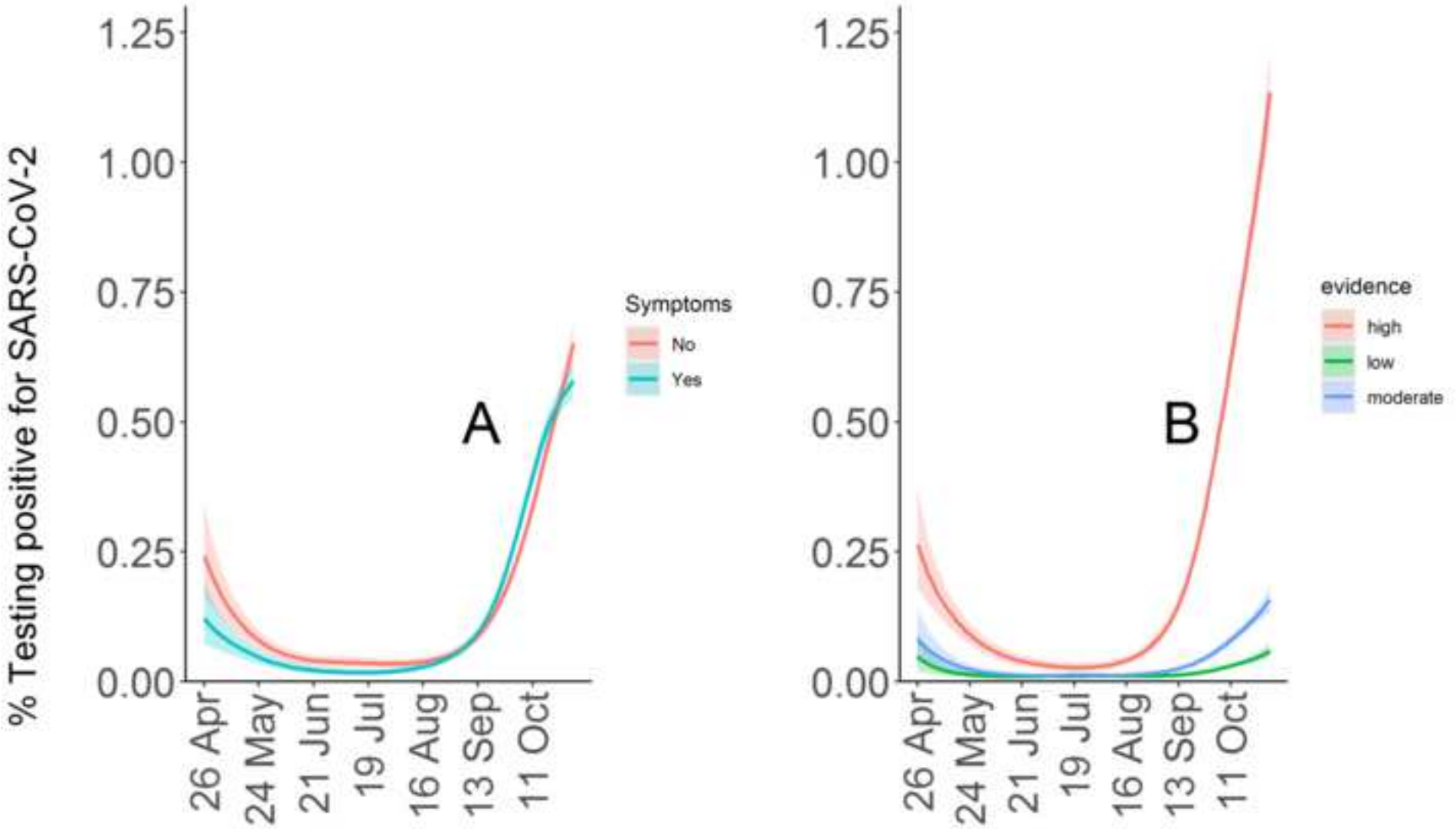


Figure 3

Click here to access/download;Figure;age_smooth_regions.jpeg

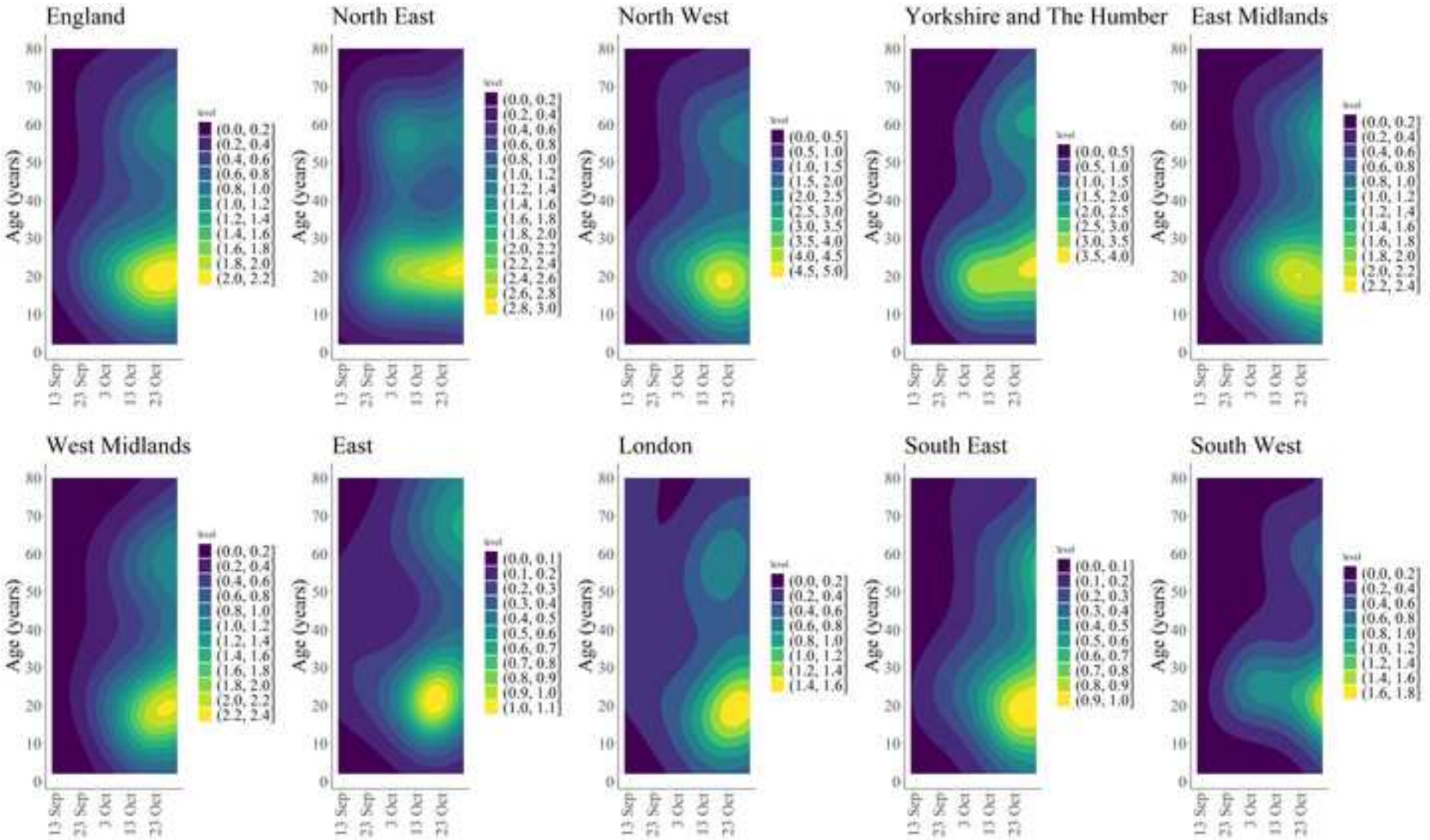


Figure 4

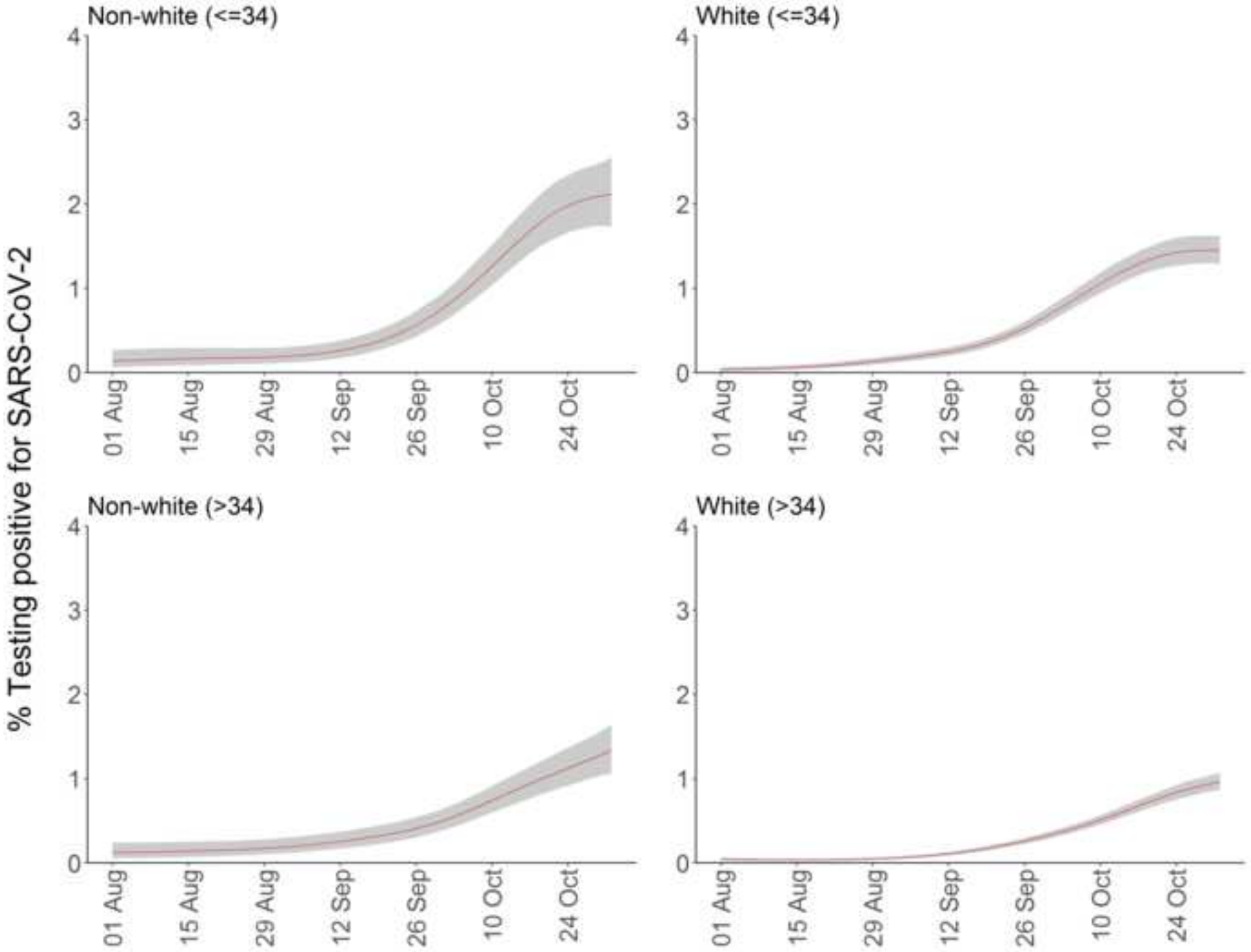
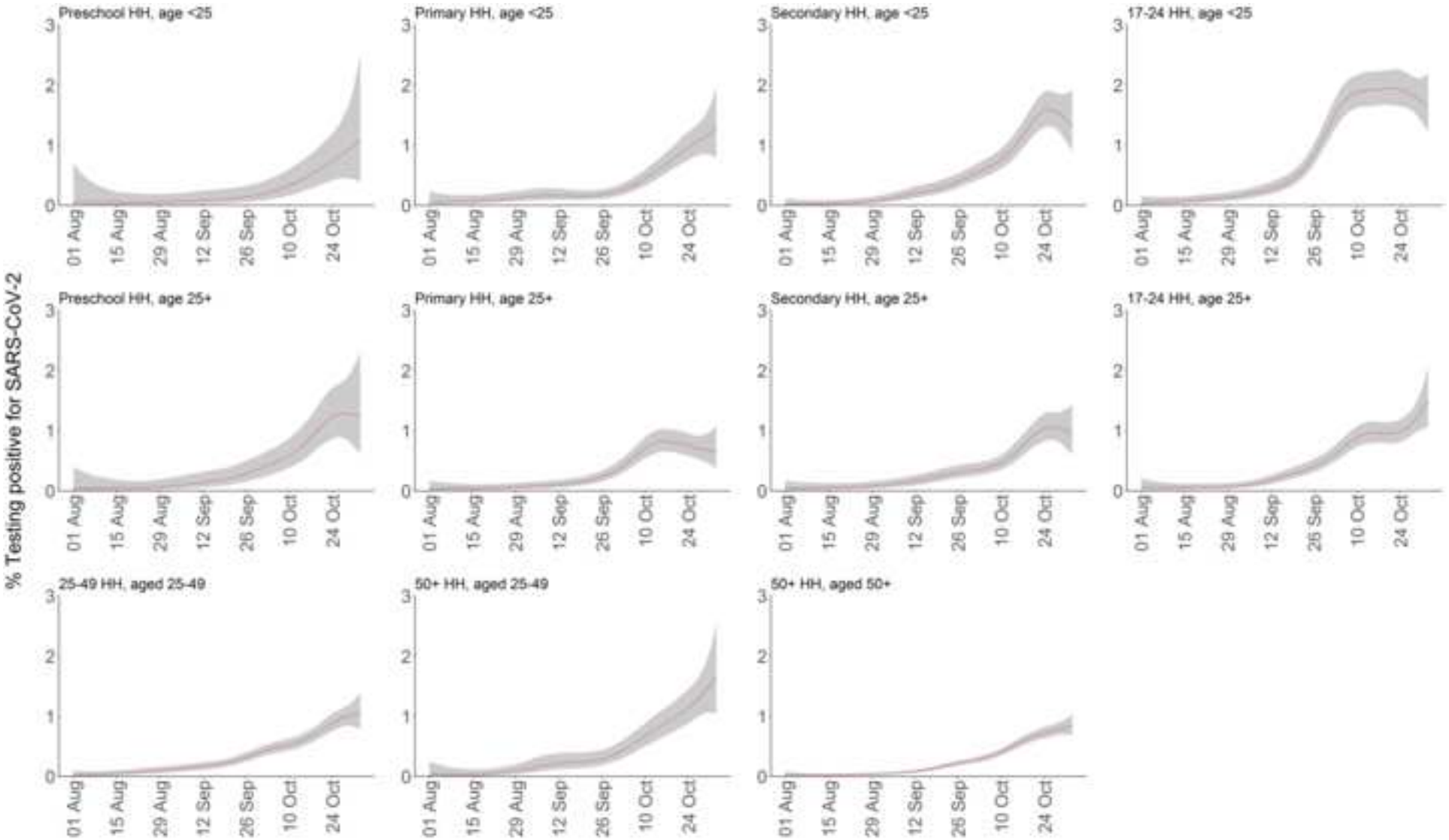


Figure 5



Community prevalence of SARS-CoV-2 in England during April to November 2020: Results from the ONS Coronavirus Infection Survey

Koen B Pouwels, Thomas House, Emma Pritchard, Julie V Robotham, Paul J Birrell, Andrew Gelman, Karina-Doris Vihta, Nikola Bowers, Ian Boreham, Heledd Thomas, James Lewis, Iain Bell, John I Bell, John N Newton, Jeremy Farrar, Ian Diamond, Pete Benton, Ann Sarah Walker, and the COVID-19 Infection Survey team

1 Sampling design

The following information on the sampling design can also be found here: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/covid19infectionsurveypilotmethodsandfurtherinformation#study-design-sampling

At the start of the study at the end of April, the sample for the survey was drawn mainly from the Annual Population Survey (APS), which consists collectively of those who successfully completed the last wave of the Labour Force Survey (LFS) or local LFS boost, and who had consented to future contact regarding research.

Around 38,000 households respond to the LFS each quarter and it is the largest regular household survey in the UK. The sampling frame for the LFS is the Postal Address File of small users, which contains approximately 26 million addresses. Only private households are included in the sample. People living in care homes, other communal establishments and hospitals are not included. Only private households in England are included in the current paper.

We initially invited about 20,000 households to take part, anticipating that this would result in approximately 21,000 individuals from approximately 10,000 households participating. Since the end of May, additional households have been invited to take part in the survey each week (roughly 5,000 a week).

At the start of the study, all respondents to the COVID-19 Infection Survey were individuals who have previously participated in an Office for National Statistics (ONS) social survey, which means the number of ineligible addresses in the sample is substantially reduced. To take part, invited households opted into the survey by contacting IQVIA, a company working on behalf of the ONS, to arrange a visit.

Since the end of July, we have further expanded the survey to invite a random sample of households from AddressBase, which is a commercially available list of addresses maintained by the Ordnance Survey. In line with our plans to increase our overall sample size, we prioritised areas under government local restriction because of an outbreak of the coronavirus (COVID-19). We have invited 40,000 extra households from 14 local authorities within selected local authorities in Greater Manchester, Lancashire and West Yorkshire to participate in this study. We also boosted our sample in London, inviting 50,000 extra households to increase the household involvement rates in this area.

In August, we announced our plans to further expand the study with the aim of increasing from 28,000 people tested per fortnight in England to 150,000 people tested per fortnight by October until March 2021. A random sample of households from AddressBase was invited for this expansion.

The number of participants enrolled in each month alongside the number of subsequent visits is shown in Table S1.

2 Models estimated in the paper

Dynamic MRP

The regression model that was used for the dynamic multilevel model and post-stratification (MRP) analysis was a Bayesian multilevel generalised additive model (GAMM) with a complementary loglog link implemented using the rstanarm package.[1-2] Sex was modelled as a fixed effect as it has only 2 levels, while age (5 levels) and region (12 levels) were modelled as random effects. This model was implemented using the following syntax:

```
stan_gamm4(result ~ s(time, by=region, k=10) + sex,
random = ~(1| age) + (1|region),
family = binomial(link="cloglog"),
data = data, iter = 3000, cores = 4,
prior = normal(0,0.5), prior_covariance
= decov(shape = 1, scale = 1),
prior_smooth=normal(location=4),
control=list(adapt_delta=0.95))
```

Associations between variables and testing positive

To assess whether particular subgroups are more likely to test positive for SARS-CoV-2 viral RNA we performed an additional analysis including variables on which we did not post-stratify. We used the same model as for the dynamic MRP but in working aged individuals only (16-74 years inclusive as defined in the Labour Force Survey) with these additional variables included as fixed covariates and age modelled as a continuous variable, using a thin-plate spline, instead of a categorical variable. Associated results can be found in Table S1.

2 Epidemiological interpretation of the complementary log-log link function when focusing on associations between variables and testing positive for the presence of SARS-CoV-2 RNA.

Our regression model operates at the individual level; in particular, we assume that there are n swabs taken, and the i-th of these is associated with time t_i, English region e_i, and a vector of other covariates x_i. These covariates are as detailed in the main text: age; work etc. The probability of the i-th swab being positive is then given by a generalised linear model (generalised additive model).

$$\pi_i = p(x_i, e_i, t_i) = g^{-1}\left(s_{e_i}(t_i) + \beta \cdot x_i + \zeta_{e_i}\right). \quad (1)$$

Here, g is the link function of the GAMM, s_{ei} is the time smoother for the region e, β is the vector of regression coefficients, and ζ_e is the random effect for region e, with these effects assumed i.i.d. with $\zeta_e \sim N(0, \sigma^2)$. The likelihood function for this model given observations $y_i = 1$ for a positive swab and $y_i = 0$ for negative, is then

$$L = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \quad (2)$$

Now suppose that the individual i acquires infection at a rate $\lambda_i(t)$, known as the *force of infection* in infectious disease modelling. The Chapman-Kolmogorov equation, using dots for time derivatives, is:

$$\dot{\pi}_i(t) = (1 - \pi_i)\lambda_i(t). \quad (3)$$

This has solution

$$\pi_i = 1 - \exp(-\int_{u=0}^{t} \lambda_i(u)du). \quad (4)$$

If we choose a complementary log-log link function

$$g(x) = \log(-\log(1 - x)) \quad (5)$$

in (1), and assume that

$$\lambda_i(t) = \emptyset_i \lambda_{e_i}(t). \quad (6)$$

in (5), then we get

$$\pi_i = 1 - \exp\left(-\emptyset \int_{u=0}^{t} \lambda_{e_i}(u)du\right) = 1 - \exp\left(-exp(\beta \cdot x_i)\, exp\left(s_{e_i}(t_i)\right) exp\left(\zeta_{e_i}\right)\right). \quad (7)$$

This implies that

$$\emptyset_i \propto \prod_a e^{\beta_a x_{ia}}, \tag{8}$$

and so we quote the value of exp(β_a) for the a-th covariate, with 1 as reference, since this is interpretable as the relative exposure to infectious risk.

3 Figures and tables

Table S1. Enrollment by month and % of participants with follow-up visits

Recruitment month	Number of participants with a first visit	Number of participants with at least 2 visits (%)	Number of participants with at least 3 visits (%)	Number of participants with at least 4 visits (%)	Number of participants with at least 5 visits (%)
April	2,440	2,421 (99%)	2,411 (99%)	2,398 (98%)	2,376 (97%)
May	19,575	19,377 (99%)	19,267 (98%)	19,156 (98%)	18,804 (97%)
June	15,987	15,811 (99%)	15,737 (98%)	15,608 (98%)	15,288 (96%)
July	16,194	16,041 (99%)	15,896 (98%)	15,567 (96%)	14,784 (91%)
August	36,395	35,755 (98%)	34,977 (96%)	32,917 (90%)	27,499 (76%)
September	94,120	90,112 (96%)	84,601 (90%)	70,090 (75%)	40,606 (43%)
October	93,064	70,918 (76%)	42,437 (46%)	15,977 (17%)	2,648 (3%)

Table S2. Characteristics of participants

	Individuals (n=280,327) N (%)[a]	Samples (n=1,119,170) N (%)[a]	Private households in England (n = 54,524,766) N (%)
Age (years)			
2-11	22,317 (8.0%)	93,125 (7.8%)	6,934,338 (12.7%)
12-16	14,692 (5.2%)	61,789 (5.2%)	3,293,437 (6.0%)
17-24	16,563 (5.9%)	64,418 (5.4%)	4,933,547 (9.1%)
25-34	28,619 (10.2%)	112,812 (9.5%)	7,526,664 (13.8%)
35-49	55,226 (19.7%)	230,480 (19.3%)	10,813,332 (19.8%)
50-69	92,629 (33.0%)	403,707 (33.9%)	13,621,021 (25.0%)
70+	50,281 (17.9%)	224,839 (18.9%)	7,402,427 (13.6%)
Sex			
Male	133,371 (47.6%)	566,252 (47.5%)	26,959,721 (49.4%)
Female	146,956 (52.4%)	624,918 (52.5%)	27,565,045 (50.6%)
Region			
North East	13,157 (4.7%)	59,616 (5.0%)	2,575,735 (4.7%)
North West	39,938 (14.2%)	174,819 (14.7%)	7,083,473 (13.0%)
Yorkshire and The Humber	27,894 (10.0%)	123,105 (10.3%)	5,315,017 (9.8%)
East Midlands	21,830 (7.8%)	98,399 (8.3%)	4,692,054 (8.6%)
West Midlands	25,327 (9.0%)	114,562 (9.6%)	5,764,018 (10.6%)
East of England	33,274 (11.9%)	134,410 (11.3%)	6,046,125 (11.1%)
London	51,113 (18.2%)	181,079 (15.2%)	8,719,114 (16.0%)
South East	41,781 (14.9%)	186,857 (15.7%)	8,855,708 (16.2%)
South West	26,013 (9.3%)	118,323 (9.9%)	5,473,522 (10.0%)
Ethnicity[b]			
White	257,413 (91.8%)	1,102,405 (92.5%)	82.6%[b]
Asian	11,748 (4.2%)	46,381 (3.9%)	9.9%[b]
Black	2,715 (1.0%)	10,790 (0.9%)	3.8%[b]
Mixed	5,115 (1.8%)	20,433 (1.7%)	2.7%[b]
Other	2,763 (1.0%)	10,178 (0.9%)	1.2%[b]
Not reported	573 (0.2%)	983 (0.08%)	NA
Household size[b]			
1	44,914 (16.0%)	118,122 (15.8%)	16%[c]
2	119,050 (42.5%)	509,332 (42.8%)	33%[c]
3	45,273 (16.2%)	189,739 (15.9%)	19%[c]
4	49,259 (17.6%)	211,404 (17.7%)	21%[c]
5+	21,831 (7.7%)	92,573 (7.8%)	12%[c]

[a] The methods that we used – Bayesian dynamic multilevel regression and poststratification – adjust for residual non-representativeness in terms of age, sex, and region.

[b] Ethnicity distribution in the target population was estimated by combining ONS estimates of age-sex-region specific number of individuals living in private households with estimates of the ethnicity distribution within those same age-sex-region categories in the overall population (including those not living in private households) obtained from the Ethpop database: Wohland P, Burkitt M, Norman P, Rees P, Boden P and Durham H, ETHPOP Database, ESRC Follow on Fund "Ethnic group population trends". www.ethpop.org. Date of extraction 21 09 2020.

[c] Household size distribution is provided for the overall population (including those not living in private households) in absence of data on household size of the target population. Therefore the 'reference data' may have too many large household sizes (e.g. student halls).

Table S3. Risk factors for testing positive for SARS-CoV-2 between 26 April and 28 June 2020.

Factor	**Number of visits in sample (number positive)**	**Relative exposure to SARS-CoV-2 (95% CrI)[a]**
Male	44,308 (53)	Ref.
Female	49,308 (57)	0.84 (0.57 - 1.25)
Work location		
Working from home	22,392 (11)	Ref.
Working outside of your home	20,621 (45)	2.47 (1.40 - 4.55)
Both	4,370 (5)	1.43 (0.53 - 3.54)
Not applicable	46,233 (49)	2.09 (1.20 - 3.75)
Job with direct contact with patients or care home residents		
Non-patient facing	89,643 (87)	Ref.
Patient facing	3,973 (23)	4.06 (2.37 - 6.72)
Non-resident facing	92,690 (105)	Ref.
Care home resident facing	926 (5)	2.35 (0.85 - 5.27)
Ethnicity		
White	88,793 (93)	Ref.
Asian	2,514 (8)	1.89 (0.87 - 3.64)
Black	788 (2)	1.04 (0.28 - 3.07)
Mixed	1,068 (0)	0.46 (0.09 - 1.84)
Other	453 (7)	7.50 (2.86 - 16.50)
Household size		
1	13,096 (16)	Ref.
2	40,426 (28)	0.62 (0.35 - 1.09)
3	17,125 (33)	1.51 (0.83 - 2.73)
4	16,704 (21)	1.36 (0.68 - 2.63
5 or more	6,261 (12)	1.25 (0.51 - 2.88)
Number of children in household		
0	64,917 (70)	Ref.
1	11,206 (22)	1.01 (0.59 - 1.73)
2	10,083 (8)	0.44 (0.20 - 0.94)
3 or more	2,593 (8)	1.65 (0.64 - 4.17)

[a]A relative exposure of 1 is the reference value (no effect).

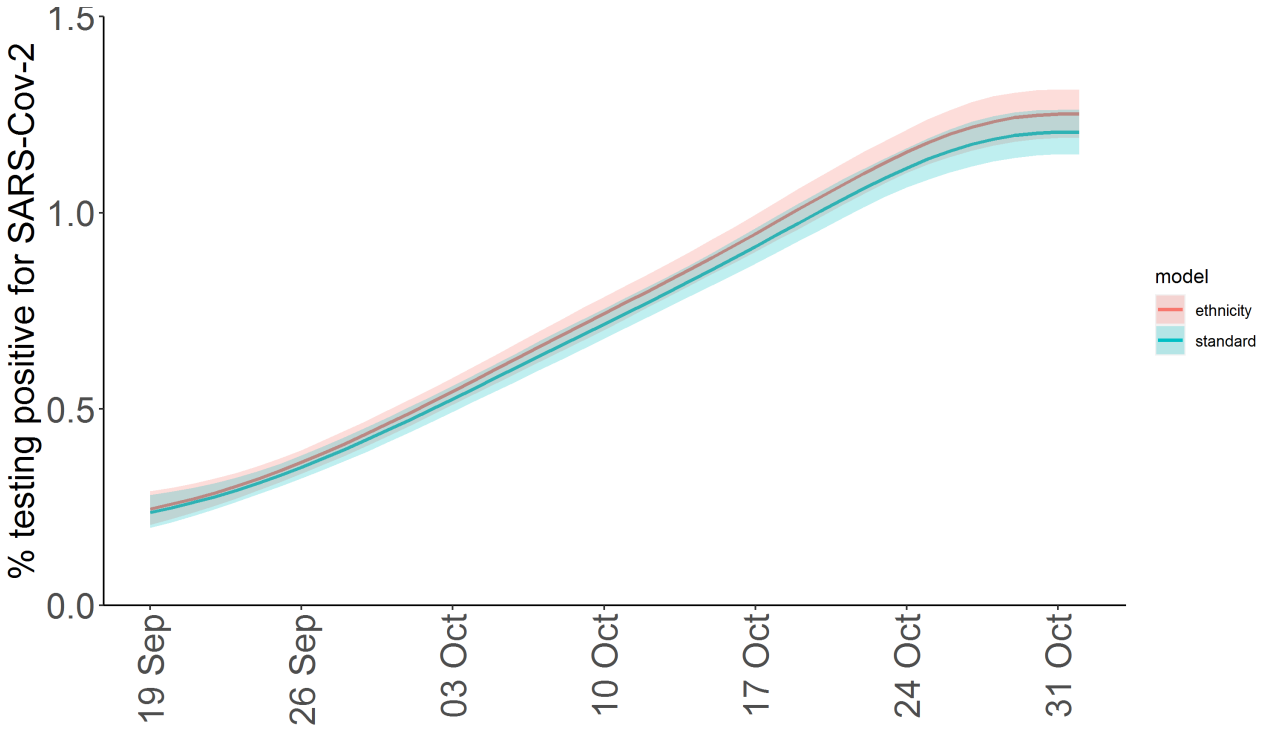


Figure S1. Comparison of multi-level regression with post-stratification model without accounting for ethnicity (standard) and with post-stratifcation for ethnicity (ethnicity). The shaded area falls within the 95% credible intervals.

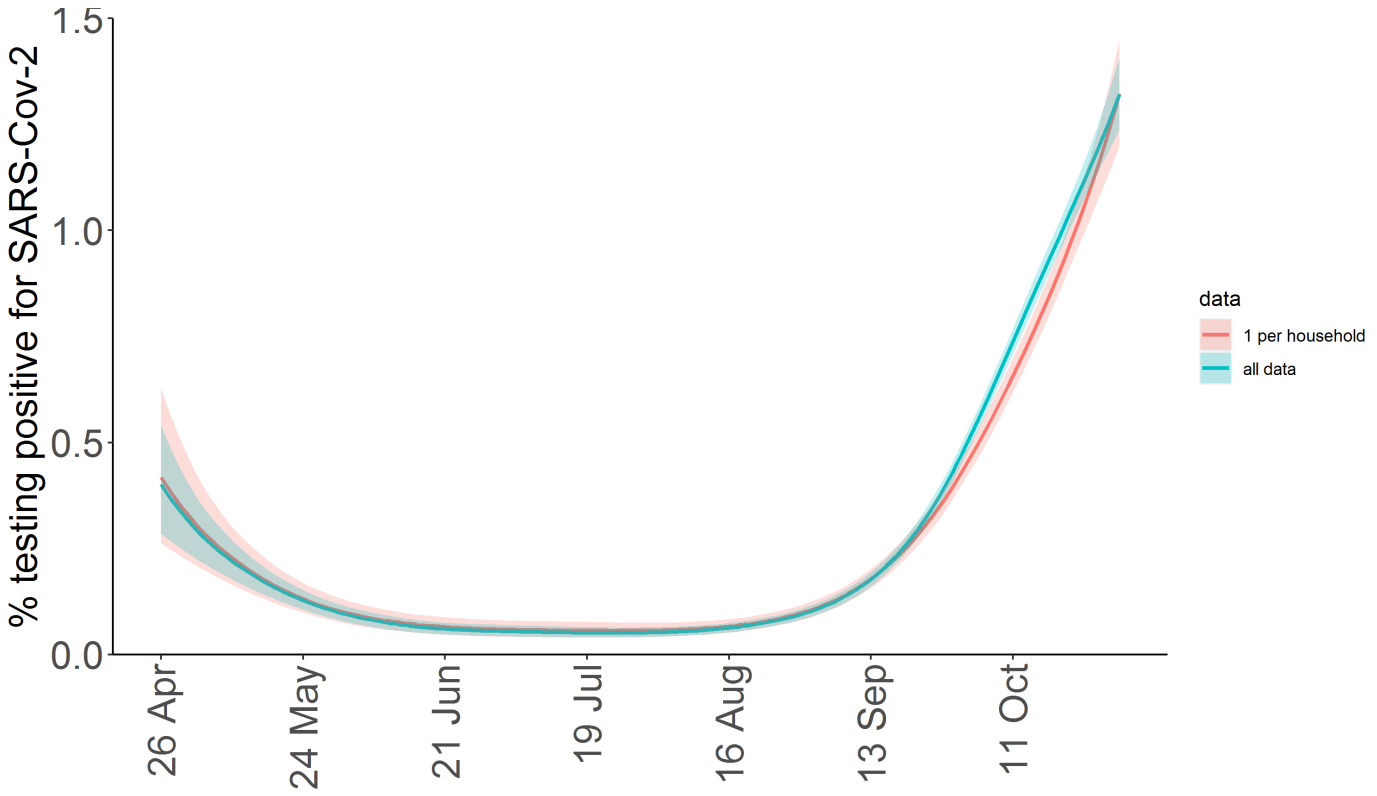


Figure S2. Comparison of using all data for the multi-level regression with post-stratification for estimation of the percentage of inhabitants testing positive over time versus use one randomly selected person per household. The shaded area falls within the 95% credible intervals.

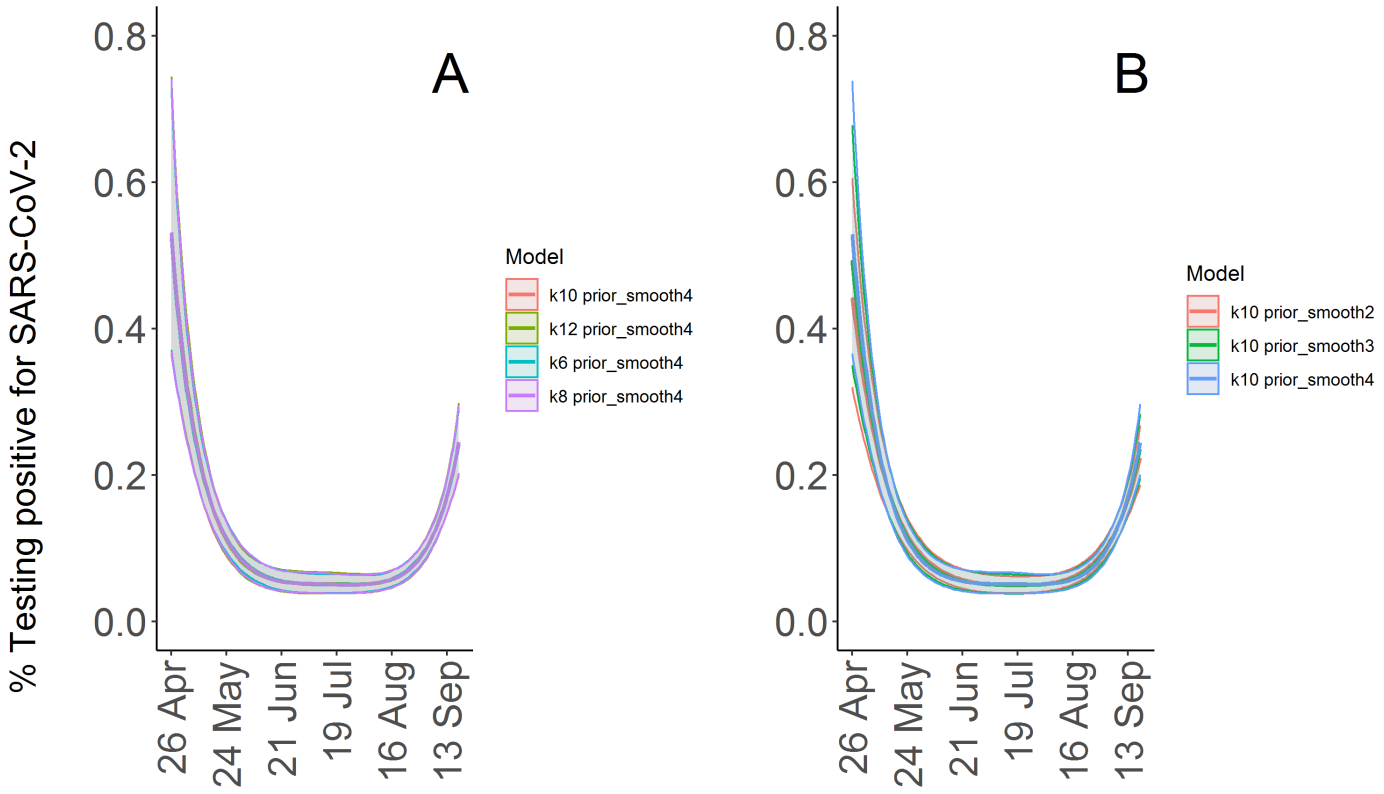


Figure S3. Comparison of models using different values for k (6, 8, 10, and 12, higher values resulted in divergent transitions) (A) and for the prior of the standard deviation of the smooth (normal prior with location 2, 3, and 4) for estimating the percentage of the population living in private households testing positive for SARS-CoV-2 with and without reporting symptoms.

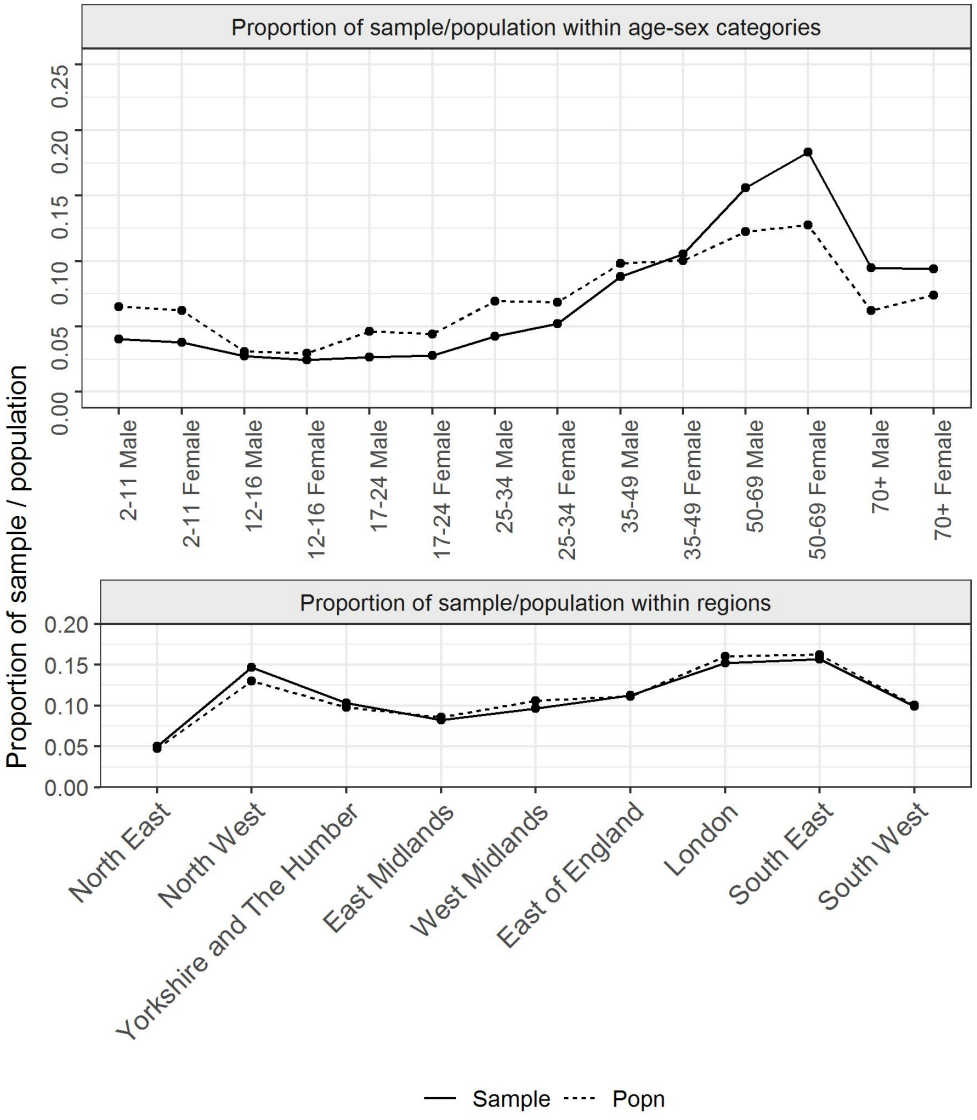


Figure S4. Representativeness of sample in terms of age and sex. Proportion of sample (solid line) and population in England (dashed line) within age- and sex-categories and regions.

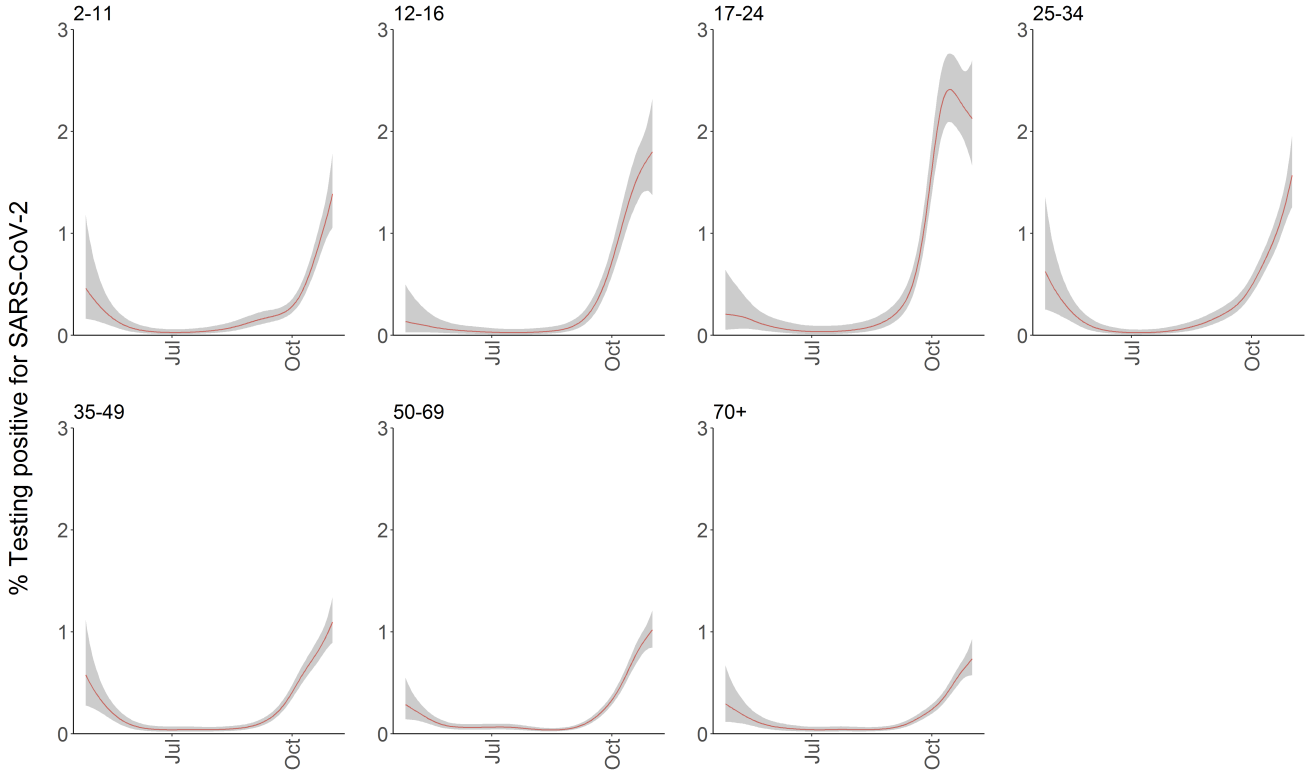


Figure S5. Percentage of population within age (in years) subgroups testing positive for SARS-CoV-2. Estimates are from a multilevel Bayesian GAM without post-stratification. The shaded area falls within the 95% credible intervals.

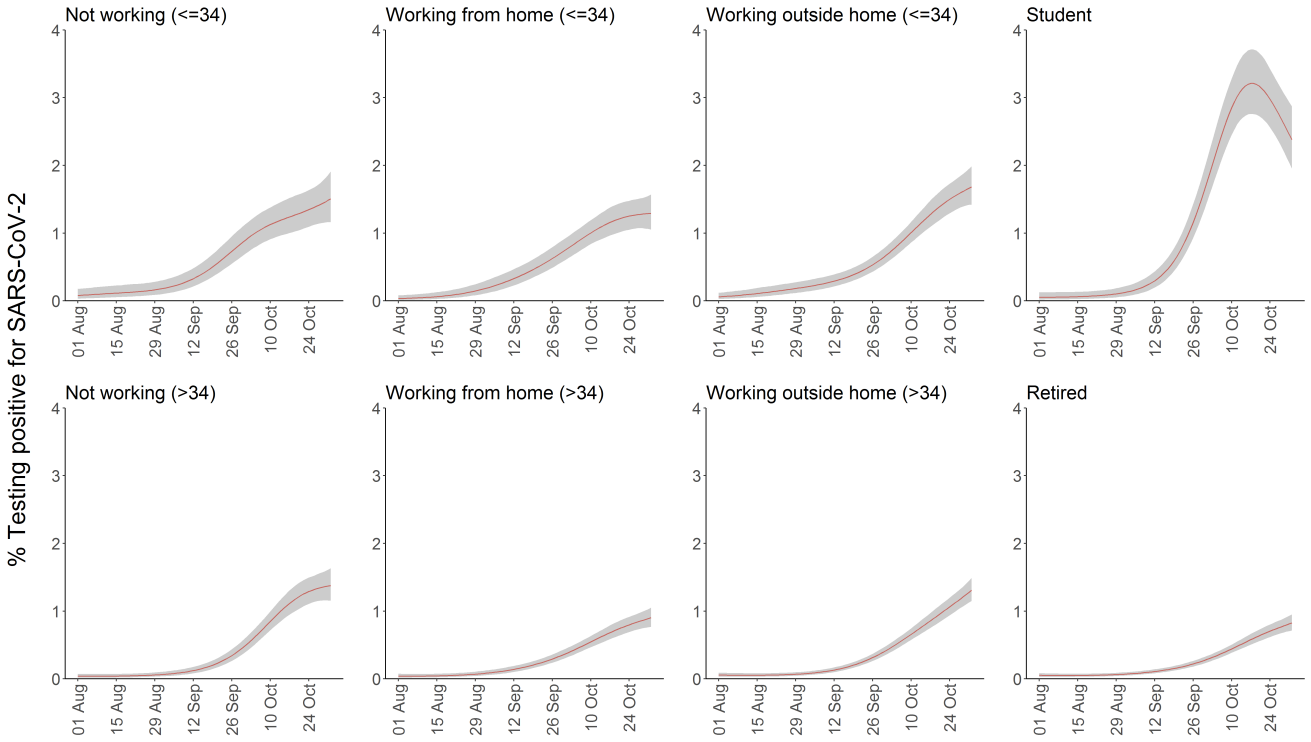


Figure S6. Percentage of population testing positive for SARS-CoV-2 by work location. Estimates are from a multilevel Bayesian GAM without post-stratification. The shaded area falls within the 95% credible intervals.

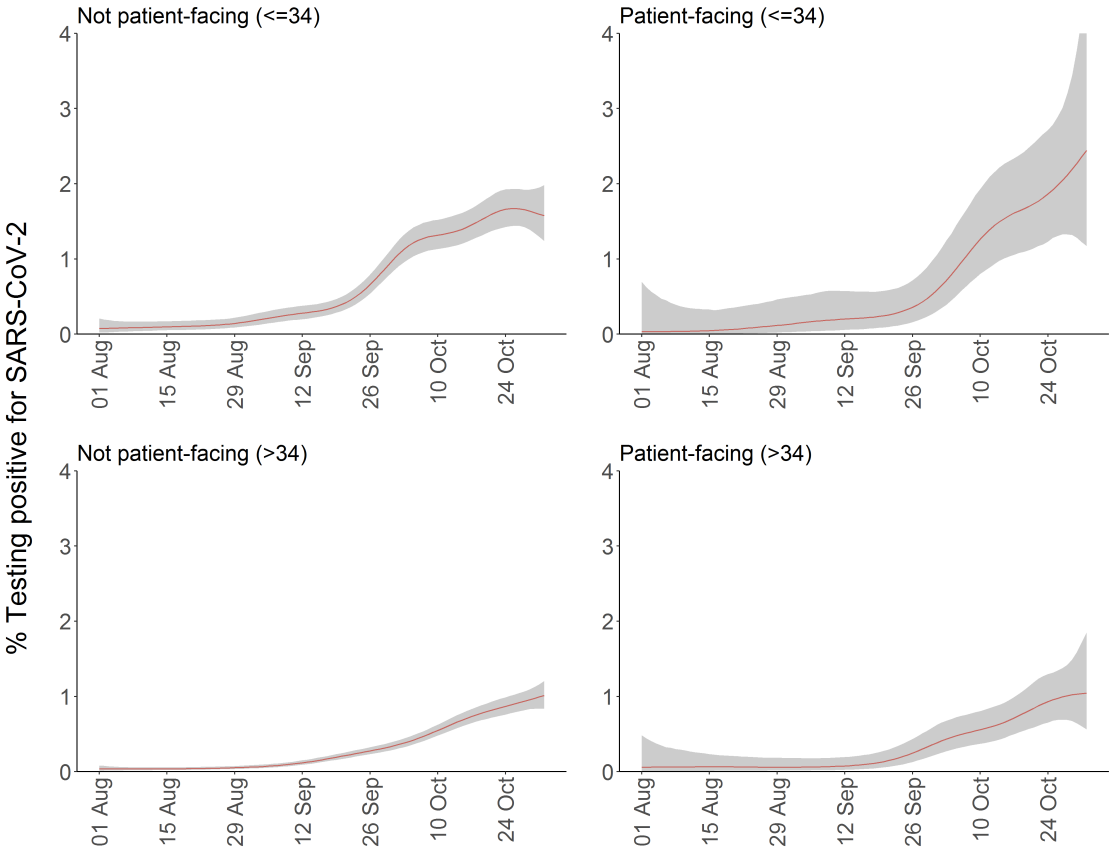


Figure S7. Percentage of population testing positive for SARS-CoV-2 by having a patient-facing role or not. Estimates are from a multilevel Bayesian GAM without post-stratification. The shaded area falls within the 95% credible intervals.

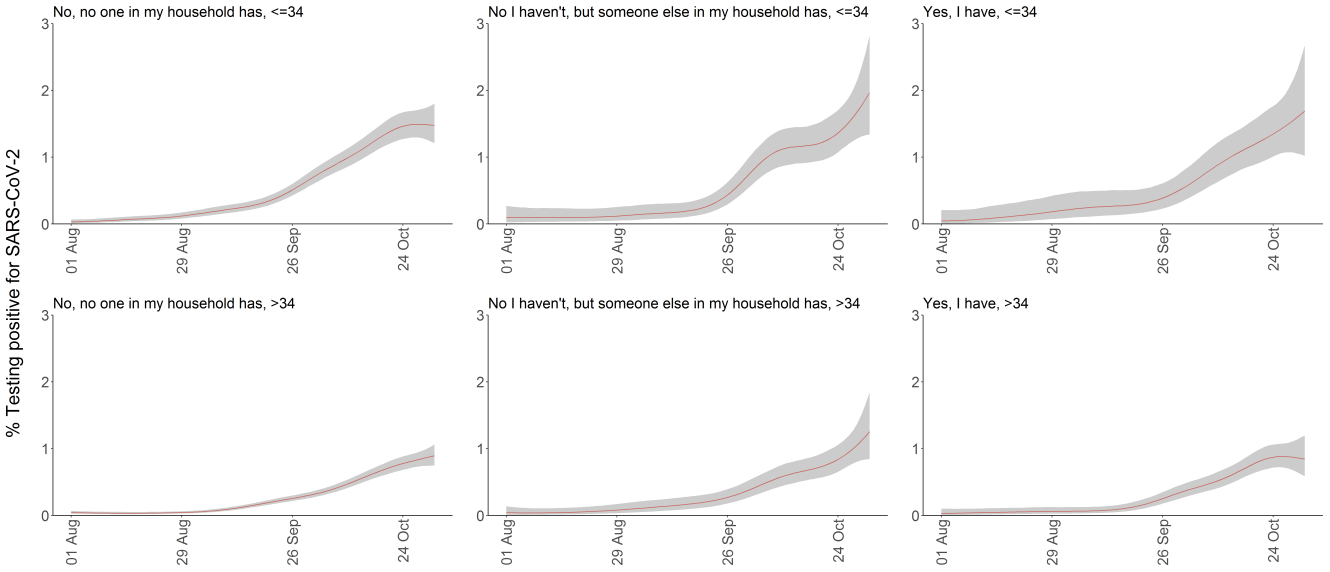


Figure S8. Percentage of population testing positive for SARS-CoV-2 by contact with hospital. Estimates are from a multilevel Bayesian GAM without post-stratification. The shaded area falls within the 95% credible intervals.

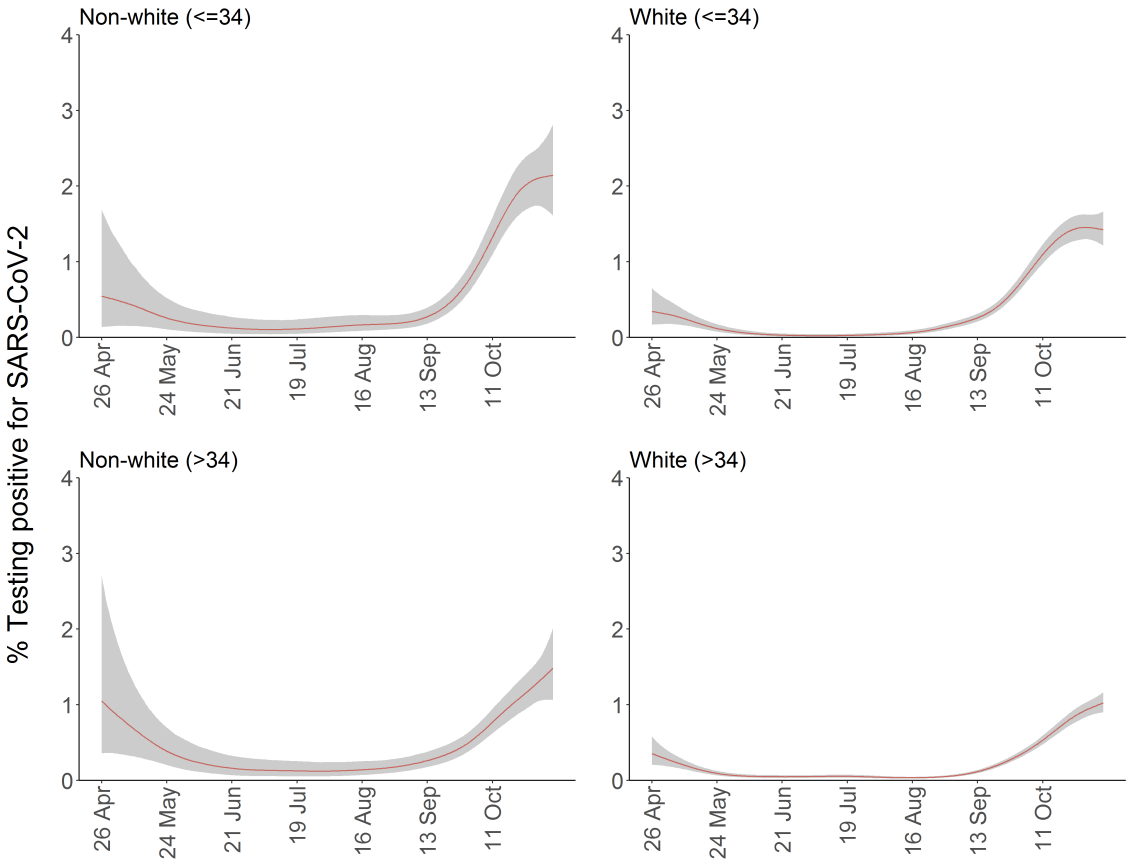


Figure S9. Percentage of population testing positive for SARS-CoV-2 by ethnicity (white / non-white). Estimates are from a multilevel Bayesian GAM without post-stratification. The shaded area falls within the 95% credible intervals.

7 References

1. https://cran.r-project.org/web/packages/rstanarm/vignettes/mrp.html
2. http://www.stat.columbia.edu/~gelman/research/unpublished/MRT(1).pdf
3. https://www.jstatsoft.org/article/view/v080i01

8 Rstanarm code used for the regression model results from table S1 in the main paper and full details on all models considered

Model 1: Analysis from Table S1

Model 2: Model 1 + random intercept for household.

Model 3: Model 1 + hospital/care home contact added to the model. The relative exposure for contact with hospital and contact with care home reported in Table 1 come from this model. As these questions were only added since 8 May 2020, this regression is restricted to subset of data starting at 8 May.

Model 4: A reduced model with less covariates than model 1.

	Model 1: Analysis from main text (table 1)	**Model 2: model 1 + random intercept for household**	**Model 3: model 1 + hospital/carehome contact**	**Model 4: reduced model**
priors for covariables	norm(0,1)	norm(0,1)	norm(0,1)	norm(0,1)
prior for intercept	norm(0,10)	norm(0,10)	norm(0,10)	norm(0,10)
prior for covariance matrix (decov function in rstanarm)	gamma(1,1)	gamma(1,1)	gamma(1,1)	gamma(1,1)
regression formula	stan_gamm4(result ~sex + s(study_day, by=region, k=5) + s(age, k=5) + work_location + patient_facing + resident_facing + ethnicity + householdsize + numchild, random= ~ (1 | region), family=binomial(link="cloglog"), data=adults, iter=3000, cores=4, prior=normal(0,1), control=list(adapt_delta=0.95))	stan_gamm4(result ~sex + s(study_day, by=region, k=5) + s(age, k=5) + work_location + patient_facing + resident_facing + ethnicity + householdsize + numchild, random= ~ (1 | region/household_id), family=binomial(link="cloglog"), data=adults, iter=3000, cores=4, prior=normal(0,1), control=list(adapt_delta=0.95))	stan_gamm4(result ~sex + s(study_day, by=region, k=5) + s(age, k=5) + work_location + patient_facing + resident_facing + ethnicity + householdsize + numchild + contact_hospital + contact_carehome, random= ~ (1 | region), family=binomial(link="cloglog"), data=adults, iter=3000, cores=4, prior=normal(0,1), control=list(adapt_delta=0.95))	stan_gamm4(result ~sex + s(study_day, by=region, k=5) + s(age, k=5) + work_location + patient_facing + resident_facing, random= ~ (1 | region), family=binomial(link="cloglog"), data=adults, iter=3000, cores=4, prior=normal(0,1), control=list(adapt_delta=0.95))
	Relative exposure (95% CrI)	Relative exposure (95% CrI)	Relative exposure (95% CrI)	Relative exposure (95% CrI)
Intercept	0.00045 (0.0002 to 0.00092)	0.00002 (0.000004 to 0.00005)	0.0005 (0.002 to 0.0011)	0.00043 (0.0002 to 0.00079)
Female	0.84 (0.57 to 1.25)	0.89 (0.59 to 1.35)	0.77 (0.49 to 1.19)	0.82(0.55 to 1.22)
work location				
Working outside of your home	2.47 (1.40 to 4.55)	2.34 (1.24 to 4.51)	2.11 (1.09 to 4.28)	2.73 (1.55 to 4.96)
Both (working from home and working outside of your home)	1.43 (0.53 - 3.54)	1.54 (0.55 to 4.05)	0.88 (0.24 to 2.60)	1.52 (0.54 to 3.76)
Not applicable	2.09 (1.20 to 3.75)	2.07 (1.14 to 4.03)	2.16 (1.17 to 4.11)	2.08 (1.18 to 3.78)
Patient/resident facing				
Patient-facing	4.06 (2.37 to 6.72)	4.73 (2.38 to 9.29)	3.76 (1.92 to 7.02)	4.24 (2.47 to 6.99)
Resident-facing	2.35 (0.85 to 5.27)	1.79 (0.57 to 5.06)	0.91 (0.19 to 3.17)	2.39 (0.88 to 5.52)
Ethnicity				
Asian	1.89 (0.87 to 3.64)	1.56 (0.50 to 4.47)	1.45 (0.56 to 3.30)	
Black	1.04 (0.28 to 3.07)	1.38 (0.32 to 5.46)	1.30 (0.31 to 3.97)	
Mixed	0.46 (0.09 to 1.84)	0.54 (0.09 to 2.46)	0.49 (0.09 to 1.98)	
Other	7.5 (2.86 to 16.50)	3.41 (0.74 to 13.09)	6.59 (2.14 to 16.07)	
Household size				
2	0.62 (0.35 to 1.09)	0.64 (0.32 to 1.30)	0.47 (0.24 to 0.91)	
3	1.51 (0.83 to 2.73)	1.28 (0.59 to 2.88)	1.61 (0.83 to 3.11)	
4	1.36 (0.68 to 2.63)	1.38 (0.59 to 3.36)	1.58 (0.76 to 3.21)	
5 or more	1.25 (0.51 to 2.88)	1.39 (0.47 to 4.10)	1.36 (0.49 to 3.52)	
Number of children in household				
1	1.01 (0.59 to 1.73)	1.12 (0.52 to 2.39)	0.80 (0.43 to 1.46)	
2	0.44 (0.20 to 0.94)	0.52 (0.19 to 1.38)	0.40 (0.16 to 0.89)	
3 or more	1.65 (0.64 to 4.17)	1.61 (0.46 to 5.19)	1.38 (0.47 to 3.83)	
Contact with hospital				
Yes, I have			2.18 (1.09 to 4.18)	
No I haven't, but someone else in my household has			1.99 (0.86 to 4.13)	
Contact with care home				
Yes, I have			0.77 (0.18 to 2.79)	
No I haven't, but someone else in my household has			0.48 (0.10 to 1.88)	

8 STROBE checklist

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No.	Recommendation	Page No.
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	4-5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	4-5
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	4-5
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	5-7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	4-7
Bias	9	Describe any efforts to address potential sources of bias	6-7
Study size	10	Explain how the study size was arrived at	4-5

Continued on next page

Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	5-6
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	5-7
		(*b*) Describe any methods used to examine subgroups and interactions	6-7
		(*c*) Explain how missing data were addressed	5-6
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	NA
		(*e*) Describe any sensitivity analyses	6-7
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	7
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Figure S3, Table S2
		(b) Indicate number of participants with missing data for each variable of interest	5-6
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	7
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	7, Figure 1
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	
		Cross-sectional study—Report numbers of outcome events or summary measures	7
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	7-8, Table S3
		(*b*) Report category boundaries when continuous variables were categorized	Table S3
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA

Continued on next page

Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	7-8
Discussion			
Key results	18	Summarise key results with reference to study objectives	9
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	10-11
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	9-11
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.