nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-021-01029-0

Check for updates

OPEN

Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework

George Nicholson[1,2,10] ✉, Brieuc Lehmann[1,2,10] ✉, Tullia Padellini[2,3], Koen B. Pouwels[4,5], Radka Jersakova[2,6], James Lomax[2,6], Ruairidh E. King[2,7], Ann-Marie Mallon[2,7], Peter J. Diggle[2,8], Sylvia Richardson[2,9], Marta Blangiardo[2,3] and Chris Holmes[1,2,6,7] ✉

Global and national surveillance of SARS-CoV-2 epidemiology is mostly based on targeted schemes focused on testing individuals with symptoms. These tested groups are often unrepresentative of the wider population and exhibit test positivity rates that are biased upwards compared with the true population prevalence. Such data are routinely used to infer infection prevalence and the effective reproduction number, R_t, which affects public health policy. Here, we describe a causal framework that provides debiased fine-scale spatiotemporal estimates by combining targeted test counts with data from a randomized surveillance study in the United Kingdom called REACT. Our probabilistic model includes a bias parameter that captures the increased probability of an infected individual being tested, relative to a non-infected individual, and transforms observed test counts to debiased estimates of the true underlying local prevalence and R_t. We validated our approach on held-out REACT data over a 7-month period. Furthermore, our local estimates of R_t are indicative of 1-week- and 2-week-ahead changes in SARS-CoV-2-positive case numbers. We also observed increases in estimated local prevalence and R_t that reflect the spread of the Alpha and Delta variants. Our results illustrate how randomized surveys can augment targeted testing to improve statistical accuracy in monitoring the spread of emerging and ongoing infectious disease.

The spread of the new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the ensuing outbreaks of coronavirus disease 2019 (COVID-19) have placed a substantial burden on public health in the United Kingdom. As of 14 July 2021, the number of people recorded to have died in the United Kingdom within 28 days of a positive SARS-CoV-2 test was 128,530 (refs. [1,2]). In response to the ongoing epidemic, the UK government has implemented a number of non-pharmaceutical interventions to reduce the transmission of SARS-CoV-2, ranging from localized measures, such as the closures of bars and restaurants, to full national lockdowns[3]. The localized measures have been employed through a regional tier system, with lower tier local authorities (LTLAs) being placed under varying levels of restrictions according to data such as the number of positive polymerase chain reaction (PCR) tests returned there over a 7-day interval (or local weekly positive tests)[4]. Following a third national lockdown that began on the 6 January 2021, the United Kingdom has undergone a staged relaxation of restrictions, with lockdown rules ending on 19 July 2021 (ref. [5]).

In the United Kingdom, there are two major ongoing studies that undertake randomized survey testing to provide an insight into the prevalence of SARS-CoV-2. Since April 2020, the Office for National Statistics (ONS) COVID-19 Infection Survey (CIS) tests a random sample of people living in the community with longitudinal follow-up[6]. The survey is designed to be representative of the UK population, with individuals aged two years and over in private households randomly selected from address lists and previous ONS surveys, although it does not explicitly cover care homes, the sheltering population, student halls or individuals currently being hospitalized. The REal-time Assessment of Community Transmission (REACT) study is a second nationally representative prevalence survey of SARS-CoV-2 based on repeated cross-sectional samples from a representative subpopulation defined via (stratified) random sampling from the National Health Service patient register of England[7,8]. Importantly, both surveys recruit participants regardless of symptom status and are therefore able to largely avoid issues arising from ascertainment bias when estimating prevalence. The ONS CIS uses multilevel regression and post-stratification to account for any residual ascertainment effects due to non-response[6], whereas the REACT study uses survey weights for this purpose.

While randomized surveillance testing readily provides an accurate statistical estimate of prevalence of PCR positivity, precision can be low at finer spatiotemporal scales (for example, at the LTLA level), even in large studies such as the ONS CIS and REACT surveys. Our major goal here is to unlock the information in non-randomized testing under arbitrary, unknown ascertainment bias. Although we expect the methods to apply in a broad manner, here we focus on Pillar 1 and Pillar 2 (Pillar 1+2) PCR tests conducted in England between 31 May 2020 and 20 June 2021 (lateral

[1]University of Oxford, Oxford, UK. [2]The Alan Turing Institute and Royal Statistical Society Statistical Modelling and Machine Learning Laboratory, London, UK. [3]MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [4]Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [5]The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford, University of Oxford, Oxford, UK. [6]The Alan Turing Institute, London, UK. [7]MRC Harwell Institute, Harwell, UK. [8]CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK. [9]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. [10]These authors contributed equally: George Nicholson, Brieuc Lehmann. ✉e-mail: george.nicholson@stats.ox.ac.uk; b.lehmann@ucl.ac.uk; chris.holmes@stats.ox.ac.uk

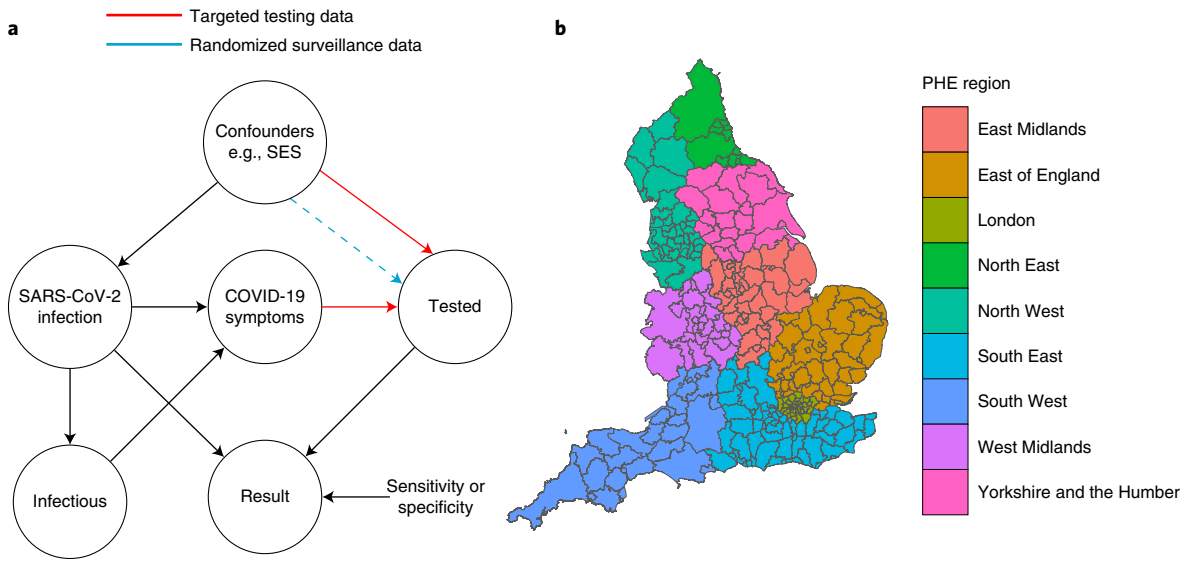


Fig. 1 | Causal diagram and spatial structure underlying the test count data. a, A DAG representing the causal models underlying SARS-CoV-2 swab testing data for targeted test-and-trace data (Pillar 1+2) and randomized surveillance data (for example, REACT). Randomization breaks the causal link between COVID-19 symptoms and swab testing. The nodes represent binary (yes/no) states for an individual in the relevant population. SES is shown as an example confounder (in addition to symptom status). The dashed line represents residual ascertainment effects stemming from non-ignorable non-response in the REACT study. **b**, A map of LTLAs in England and their corresponding PHE regions.

flow device (LFD) tests are not included; further details provided in Methods and Data availability). Pillar 1 tests refer to "all swab tests performed in Public Health England (PHE) labs and National Health Service (NHS) hospitals for those with a clinical need, and health and care workers"[9], and Pillar 2 tests comprise "swab testing for the wider population"[9]. Pillar 1+2 testing therefore has more capacity than the randomized programmes, but the protocol incurs ascertainment bias because those at increased risk of being infected are tested, such as frontline workers, contacts traced to a COVID-19 case or the subpopulation presenting with COVID-19 symptoms, such as loss of taste and smell[9]. Hence, raw prevalence estimates from Pillar 1+2 data (as a proportion of tested population) will tend to be biased upwards and cannot directly be used to estimate the unknown infection rate in a region. In contrast, as a proportion of the entire population, the bias is downwards as not all individuals with infection in the area are captured. Furthermore, the degree of upward bias may be influenced by overall testing capacity and uptake. In addition, the raw prevalence estimates tend not to capture asymptomatic infection, even though there is evidence to indicate that asymptomatic individuals can contribute to viral transmission[10,11].

Combining data from multiple surveillance schemes can improve estimates for prevalence. For example, Manzi et al.[12] incorporated information from multiple, biased, commercial surveys to provide more accurate and precise estimates of smoking prevalence in local authorities across the East of England. A number of geostatistical frameworks for infectious disease modelling based on multiple diagnostic tests have been developed[13–15]. These accommodate different sources of heterogeneity among the tests to deliver more reliable and precise inferences on disease prevalence.

To understand the ascertainment bias problem and to enable a statistical approach to correction, it is helpful to consider a simplified causal model[16,17] for Pillar 1+2 data. This is represented by a directed acyclic graph (DAG), shown in Fig. 1a, that charts the dependencies of an individual from infection status to test result. The circles indicate the binary (yes/no) states of an individual. The DAG characterizes the joint distribution of the major factors leading to the observed data. Throughout the paper, we use the term 'targeted testing data' to refer to data gathered under some ascertainment process distinct from (stratified) random sampling, with an exemplar being selection for testing of the subpopulation with COVID-19 symptoms, which comprises a sizeable proportion of Pillar 1+2 tests. There are several other potential confounders, exemplified in Fig. 1a by socioeconomic status (SES), which is a well-studied factor of both infection risk and access to healthcare and/or testing. The DAG explicitly characterizes statistically why we cannot directly use Pillar 1+2 data. The DAG also points to a potential solution that we pursue here: if the statistical dependencies as indicated by the arrows in Fig. 1a can be modelled, then we can correct for the ascertainment bias in Pillar 1+2 data.

In addition to prevalence, there are a number of epidemiological parameters that may be useful for informing localized non-pharmaceutical interventions. For example, one particular variable of interest is the (time-varying) effective reproductive number R_t, which is defined roughly as the average number of infections caused by an infectious individual. That is, when $R_t > 1$, the epidemic will continue to spread. The current pandemic has spurred the development of models that aim to incorporate multiple sources of data to estimate important epidemiological parameters. See Supplementary Table 1 for an overview of the methodological work most related to ours[18–25] (https://localcovid.info/), including a brief description of each method and what the data inputs and results outputted are; we also recommend refs. [26,27] for reviews, which have a particular focus on R_t.

Within this urgent and fast developing area of research, it is clearly important to define the aspects in which our method contributes. First, we have developed methods to infer unbiased local prevalence, I_t, from targeted testing data. This is important in its own right because being able to estimate local prevalence accurately from targeted testing data adds an important facet to existing COVID-19 monitoring capabilities. Here, we focus on weekly period prevalence and explicitly target the number of infectious individuals via a correction to the estimated PCR-positive numbers. Second, our method outputs bias-adjusted cross-sectional prevalence likelihoods $p(n_t \text{ of } N_t \mid I_t)$, where n_t and N_t are positive

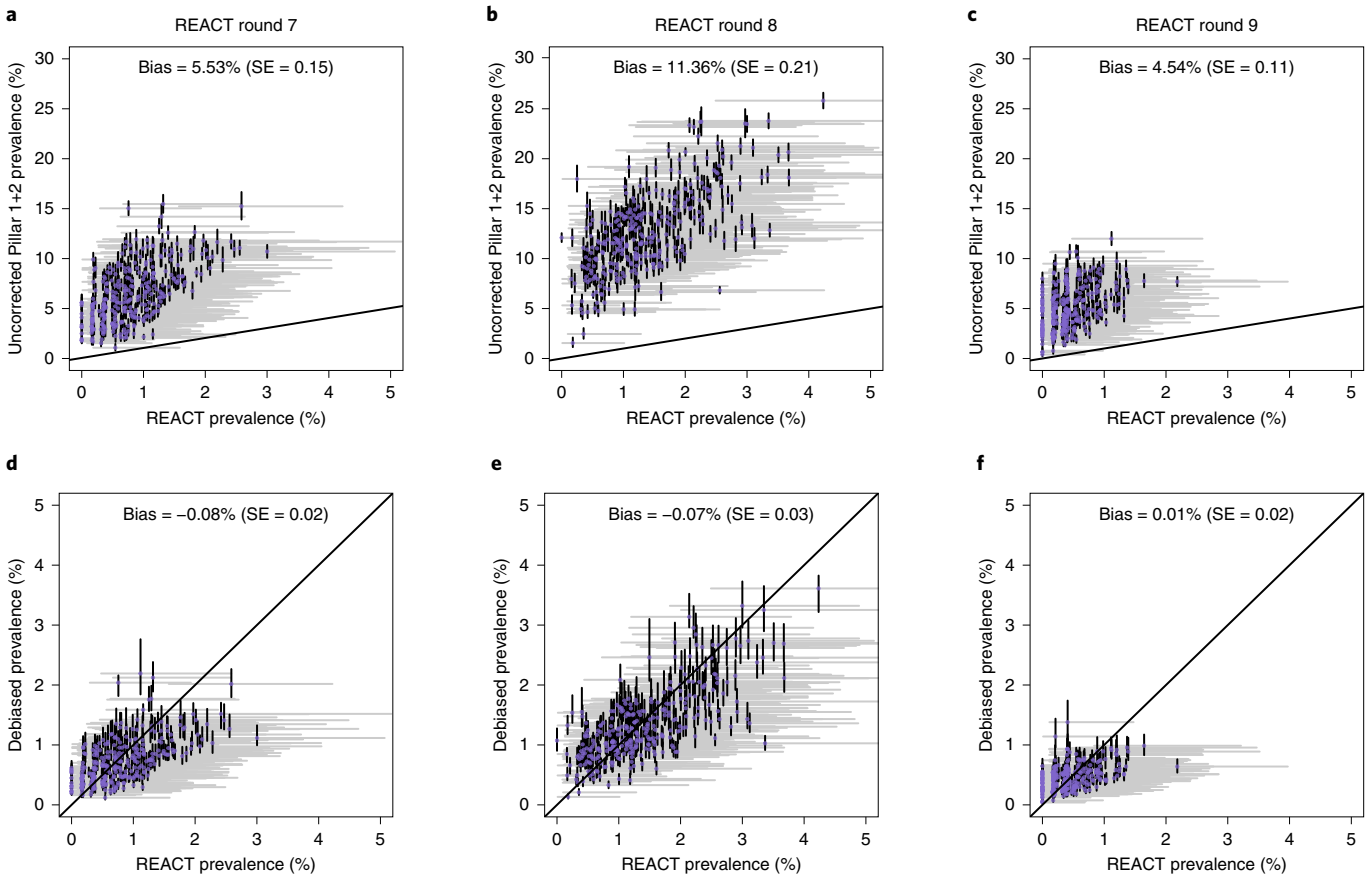


Fig. 2 | Uncorrected (top) and corrected (bottom) Pillar 1+2 prevalence estimates against REACT estimates. a–f, Uncorrected (raw positivity rates) and corrected (debiased) Pillar 1+2 PCR-positive prevalence estimates against (gold-standard) REACT estimates from randomized surveillance. Each point corresponds to a LTLA. Each scatter plot compares pillar 1+2 prevalence estimates against unbiased estimates from the REACT study. **a**,**d**, REACT round 7 data (13 November 2020 to 3 December 2020). **b**,**e**, Round 8 (6–22 January 2021. **c**,**f**, Round 9 (4–23 February 2021). Uncorrected results are shown in **a**–**c** and bias-corrected cross-sectional estimates in **d**–**f**. Horizontal grey lines are 95% exact binomial confidence intervals from the REACT data. The number of independent tests underlying each mean and (horizontal) credible intervals for the REACT data varied between 248 and 2,387. Vertical black lines in **a**–**c** are 95% exact binomial confidence intervals for from the raw, non-debiased Pillar 1+2 data. Vertical black lines in **d**–**f** are 95% posterior credible intervals from the debiased Pillar 1+2 data. The number of independent tests underlying each mean and (vertical) credible interval for the Pillar 1+2 data varied between 1,117 and 42,458. Neither set of prevalence estimates has been corrected for false positives or negatives. Note that in **d**–**f**, the credible interval widths are systematically tighter for the debiased Pillar 1+2 compared with the REACT data, which highlights the useful information content in debiased Pillar 1+2 data.

and total targeted test counts, respectively. This allows prevalence information from targeted data to be coherently embedded in a modular way into complex spatiotemporal epidemiological models, including those synthesizing multiple data types. We exemplify this by implementing a susceptible-infectious-recovered (SIR) model around our ascertainment model likelihood. Third, our local ascertainment model is based on targeted testing data alone with both the number of positive and total tests being modelled (n_t and N_t). This has two important benefits: spatiotemporal variation in testing uptake and capacity is explicitly conditioned on (via N_t), and differential test specificity and sensitivity can be naturally incorporated into our causal ascertainment model.

Results

Correcting for ascertainment bias in targeted testing data. Figure 2a–c displays the percentage of positive Pillar 1+2 tests (as a proportion of those tested) against accurate prevalence estimates from the REACT study, which shows a clear upward bias (each point corresponds to a single LTLA). Here, we introduce a bias-correction method that aims to provide accurate estimates of prevalence at the local level, as displayed in Fig. 2d–f, based on the posterior cross-sectional prevalence $p(I_t \,|\, n_t \text{ of } N_t)$.

With reference to the causal DAG in Fig. 1a, we define the essential bias parameter, δ, as

$$\delta := \log\left(\frac{\text{odds}(\text{ tested} \,|\, \text{infected})}{\text{odds}(\text{ tested} \,|\, \text{not infected})}\right) \quad (1)$$

that is, the log odds-ratio of being tested in the infected subpopulation versus in the non-infected subpopulation. Larger values of δ generally correspond to higher levels of ascertainment bias; that is, a higher chance of an individual with an infection being selected for testing relative to an individual without infection.

Our approach combines randomized surveillance data (REACT) and targeted surveillance data (Pillars 1+2) to infer δ at the coarse geographical level (PHE region; Fig. 1b). We then take forward this information by specifying a temporally smooth empirical Bayes (EB) prior on $\delta_{1:T}$, applied to each constituent local region (LTLA) in the local prevalence analyses. Figure 3a shows the resulting EB priors on δ. There is potentially more variation in δ across regions early and late in the sampling period (before September 2020 and after March 2021), although the prior credible intervals are broad and often overlapping. The data provide more information on δ between October 2020 and February 2021.

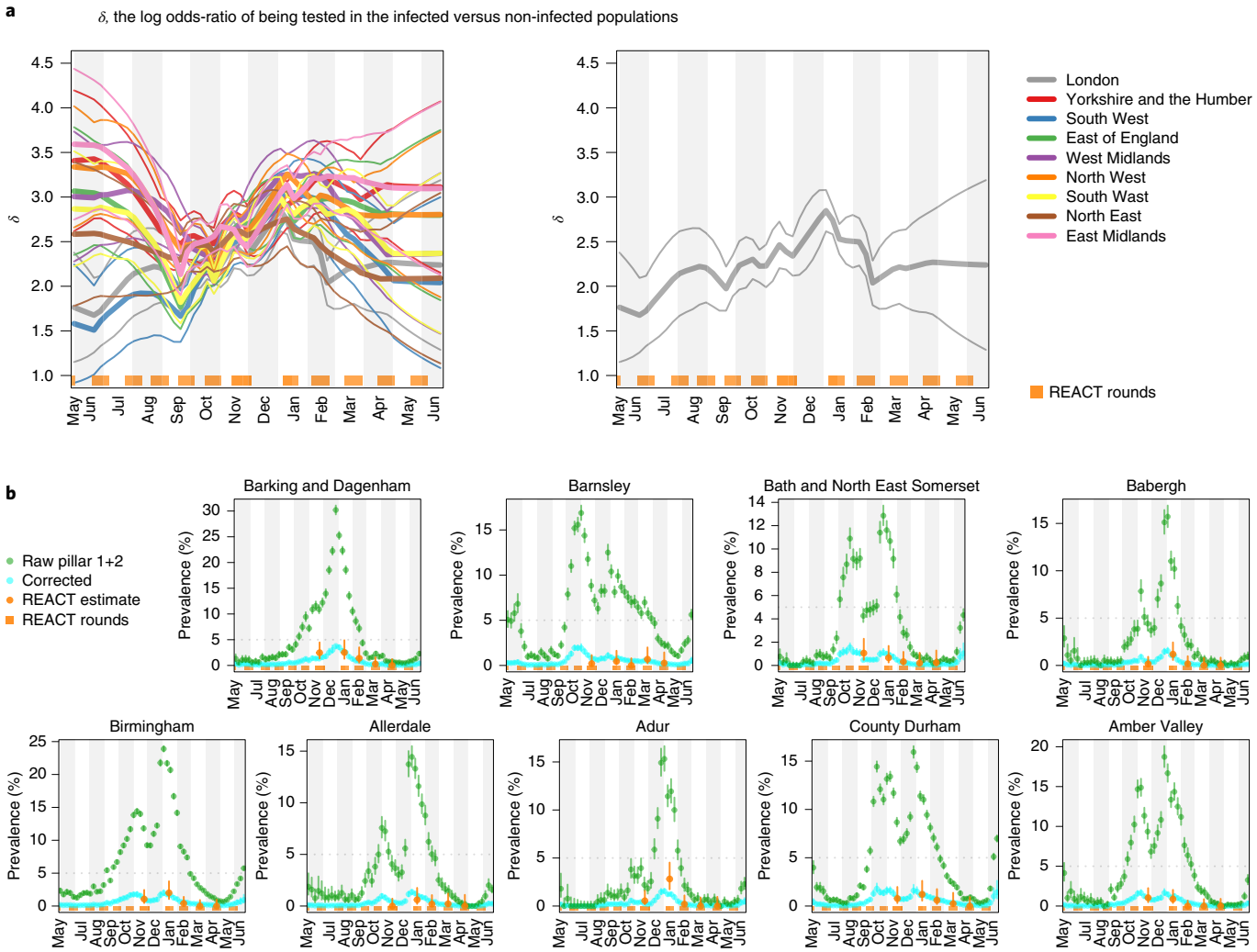


Fig. 3 | Ascertainment bias parameters and LTLA-level prevalence estimates. a, Smooth EB priors on bias parameters $\delta_{1:7}$. Left: heterogeneous bias across the nine PHE regions. Right: London only. The thick curves show the prior means and the narrow curves show 95% credible intervals. Note that δ is the log odds-ratio, so, for example, $\delta = 3$ implies that the odds of being tested are $e^3 \approx 20$ times higher in individuals with infection compared with individuals without infection. **b**, LTLA-level prevalence estimates: raw Pillar 1+2 estimates (that is, positivity rate), cross-sectionally corrected Pillar 1+2 and gold-standard REACT estimates. For each of the nine PHE regions, we present the constituent LTLA whose name is ranked top alphabetically. The number of independent tests underlying each (orange) mean and credible interval based on the REACT data varied between 288 and 620. The number of independent tests underlying each (green or cyan) mean and credible interval based on the Pillar 1+2 data varied between 390 and 43,650. The green symbols and error bars show the mean exact binomial 95% confidence intervals. The cyan symbols and error bars show posterior median and 95% credible intervals. The orange symbols and error bars show the mean and 95% exact binomial confidence intervals.

Cross-sectional local prevalence from targeted testing data. *Debiased likelihood for modular sharing of prevalence information.* Equipped with a coarse-scale (PHE-region level) EB prior on bias δ, we evaluated a fine-scale (LTLA-level) δ-marginalized likelihood of the form $p(n_t \text{ of } N_t | I_t, \hat{\nu}_t)$ as described in equation (17) in the Methods ("Cross-sectional inference on local prevalence"). This debiased prevalence likelihood can be readily exported and modularly incorporated into more complex models, as we illustrate below ("Longitudinal local prevalence and transmission").

Cross-sectional prevalence posterior. The δ-marginalized likelihood can be inputted directly into cross-sectional Bayesian inference, outputting the prevalence posterior $p(I_t | n_t \text{ of } N_t, \hat{\nu}_t)$ for each time point at which such count data are available. Figure 3b plots these cross-sectional prevalence posteriors beneath the raw counts for a subset of LTLAs across the nine PHE regions. REACT sampling periods are plotted at the base of each panel, and local prevalence estimates from REACT round 7 (November 2020) and round 8 (January 2021) are also superimposed. The corrected cross-sectional prevalence estimates are consistent with the gold-standard REACT estimates, but are more precise, as expected from Bayesian principles of data synthesis.

Longitudinal local prevalence and transmission. The cross-sectional debiased likelihood can be introduced modularly into a wide variety of downstream epidemiological models. We illustrate this by using the likelihood as an input to a simple SIR epidemic model (Methods, "Full Bayesian inference under a stochastic SIR epidemic model", and Extended Data Fig. 1). Figure 4a plots the estimated prevalence against R_t number at the most recent time point (the week of 20 June 2021), with each point corresponding to a single LTLA. The scatter plot provides a quick visual representation

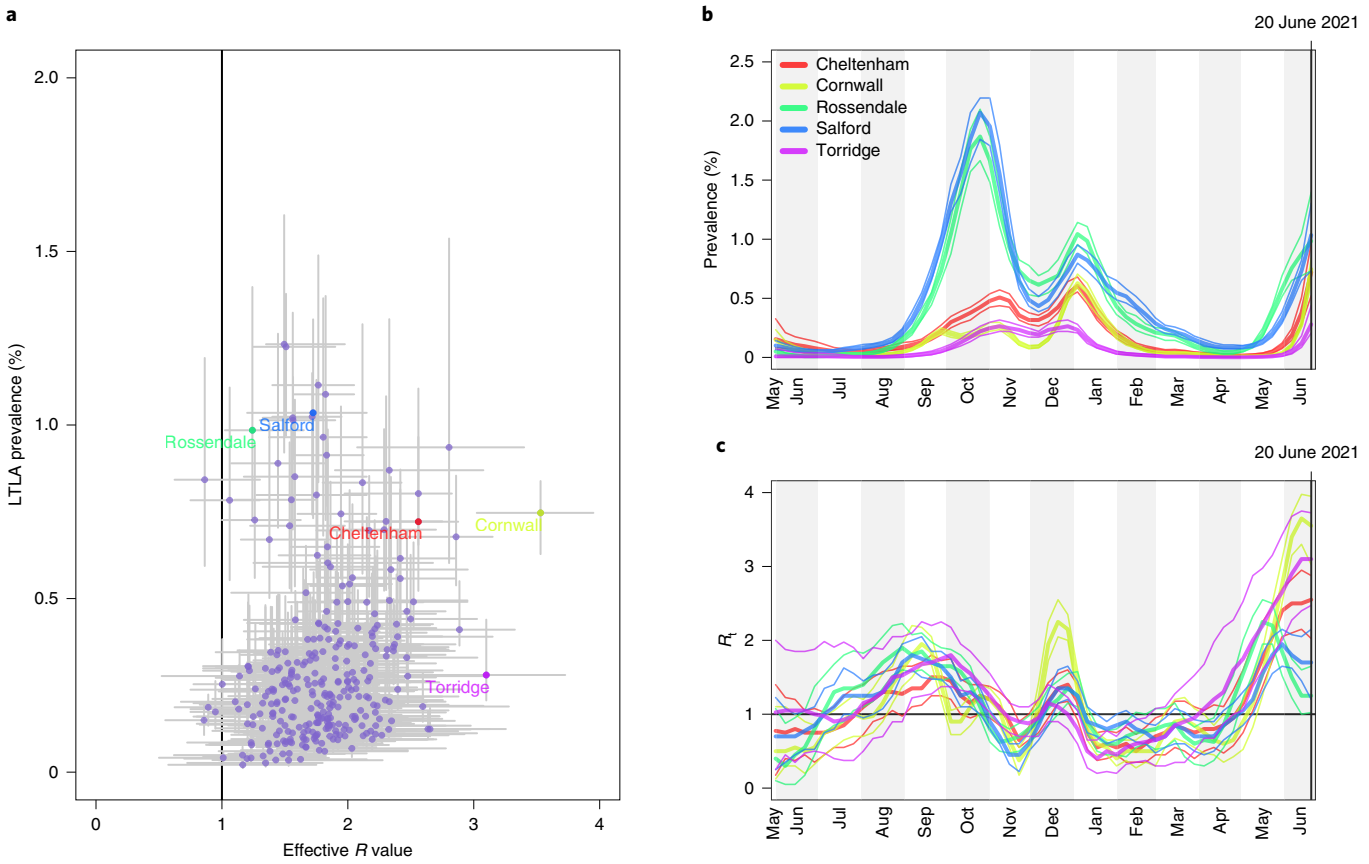


Fig. 4 | Outputs of the longitudinal local prevalence model. a, Scatterplot of prevalence against effective *R* number (each point corresponds to one LTLA) for the week of 20 June 2021. **b**, Longitudinal posteriors for prevalence at a selection of LTLAs. **c**, Longitudinal posteriors for R_t at a selection of LTLAs. The vertical line and horizontal line in **b** and **c**, respectively, indicate an effective reproduction number of $R_t = 1$; when $R_t > 1$, the number of cases occurring in a population will increase. In **a**, the symbols show posterior medians and the error bars show 95% credible intervals. In **b** and **c**, the thick lines show posterior medians and the narrow lines show 95% credible intervals.

of regions where transmission rates and/or prevalence are relatively high. To illustrate, we label five LTLAs with high prevalence and/or R_t estimates. The estimated longitudinal prevalence and R_t for this subset of LTLAs (Fig. 4b,c) can help further characterize the longitudinal dynamics of prevalence and transmission in the time interval leading up to 20 June 2021. In particular, the data show the estimated rate of change in prevalence and separately indicate whether R_t is increasing or decreasing.

Figure 5a displays the spatiotemporal local prevalence and Fig. 5b displays R_t, using a fortnightly sequence of maps, with each LTLA coloured according to its estimate prevalence or R_t. Zoom-in boxes display the local fine-scale structure for London.

Relating local prevalence and transmission to spread of the variants of concern. A striking feature of the maps in Fig. 5a is the increasing prevalence in London throughout November to December 2020. This is consistent with the known arrival of the Alpha variant of concern (VoC) 202012/01 (lineage B.1.1.7) that emerged in the South East of England in November 2020, and has been estimated to have a 43–90% higher reproduction number than pre-existing variants[28]. Similarly, the increase in R_t from May 2021 onwards is in accordance with the spread of the Delta VoC 21APR-02 (lineage B.1.617.2), which is estimated to have a reproduction number approximately 60% higher than that of the Alpha VoC[29].

Similar to a previous study[28], we characterized the relationship between the estimated local R_t and the frequency of Alpha VoC 202012/01, as approximated by the frequency of *S* gene target failure (SGTF) in Taqpath sequencing assays used during this time period[30]. Figure 6 illustrates the spatial distributions of the Alpha VoC 202012/01 against estimated prevalence and estimated R_t from mid-November 2020 to mid-December 2020. The increase in frequency of the VoC was initially isolated to the South East but then spread outwards, accompanied by a corresponding increase in both local estimated prevalence and R_t. We observe a strong positive association between the local VoC frequency and estimated local R_t, which are consistent with the increased transmissibility of this VoC identified in ref. [28].

We performed a similar analysis for the Delta VoC 21APR-02 using data provided by the Wellcome Sanger Institute's Covid-19 Genomics Initiative[31]. Extended Data Fig. 2 shows the spatial distributions of the Delta VoC 21APR-02 against estimated prevalence and estimated R_t from the end of April 2021 to the start of June 2021. We see that the Delta VoC becomes the dominant variant over the course of this time period, and in contrast to the Alpha VoC, the spread of the variant was not isolated to a single region of England. We again observe a strong positive association between the local VoC frequency and estimated local R_t. A simple linear regression of R_t against Delta frequency for the week of 23 May 2021 indicated an increase in transmissibility of 0.55 (0.39–0.71) due to the Delta VoC, which is in accordance with estimates obtained in ref. [29].

Accuracy validation using ultra-coarse and incomplete data to estimate δ. We assessed the performance of debiased fine-scale (LTLA-level) prevalence estimates by measuring how well they predict

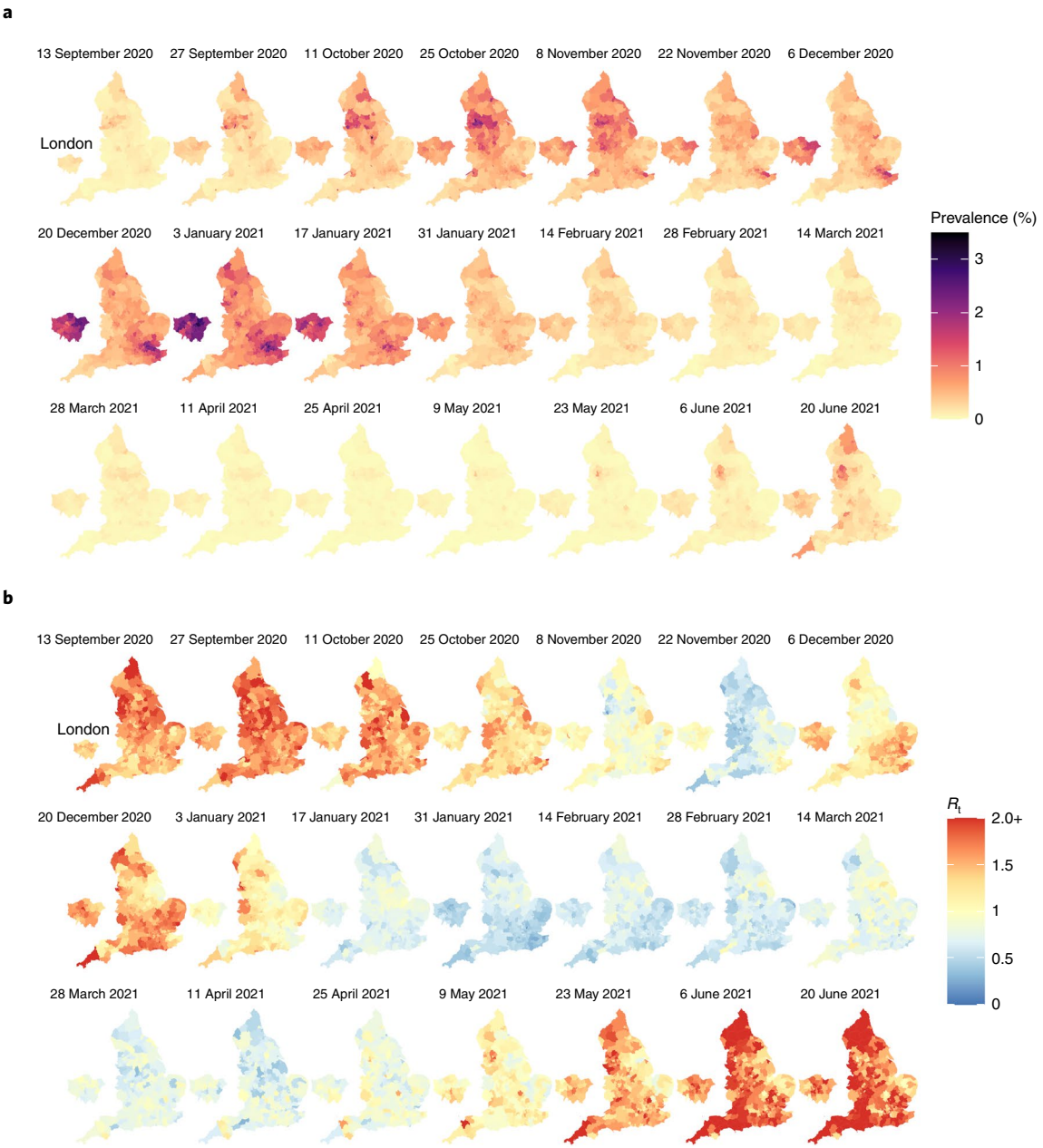


Fig. 5 | Maps of estimated prevalence and effective reproduction number. a, Fortnightly maps of estimated local prevalence in England from 13 September 2020 to 20 June 2021. **b**, Fortnightly maps of estimated local R_t in England from 13 September 2020 to 20 June 2021.

LTLA-level REACT data. The validation is best described in terms of coarse-scale REACT training data and contemporaneous fine-scale REACT test data. The training data inputted are REACT PHE-region-level and Pillar 1+2 LTLA-level positive (and number of) test counts for the week at the centre of the corresponding REACT round to be predicted. The test data are REACT LTLA-level positive (and number of) test counts aggregated across the relevant REACT sampling round. Figure 2 visually compares cross-sectional LTLA prevalence estimates from debiased targeted data (that is, based only on the training data) with accurate gold-standard estimates from REACT LTLA-level test data. The average estimated bias is reduced to low levels for comparisons with REACT round 7 (–0.08%, standard error (SE) = 0.02), round 8 (–0.07%, SE = 0.03) and round 9 (0.01%, SE = 0.02). Extended Data Fig. 3c,d displays analogous results for REACT rounds 10 and 11, with average estimated bias reduce to 0.03% (SE = 0.01) and 0% (SE = 0.01), respectively.

REACT and ONS CIS are among the most comprehensive randomized surveillance studies in the world. We have tried to assess how well the debiasing model might hold when we are faced with coarser-scale or more limited randomized testing data. First, to investigate the downstream effects of ultra-coarse-scale randomized surveillance data, we aggregated all REACT data to the national level, estimated the δ curve at this ultra-coarse national level and then took this δ forward to estimate local prevalence. We found that estimates retained a high level of accuracy (Extended Data Fig. 4g–i). Second, to examine the effects of a more limited randomized surveillance regime, we left out REACT round 8, re-estimated δ curves at the PHE-region level and used these to infer local prevalence. In this case, we lost precision in our prevalence estimates for omitted

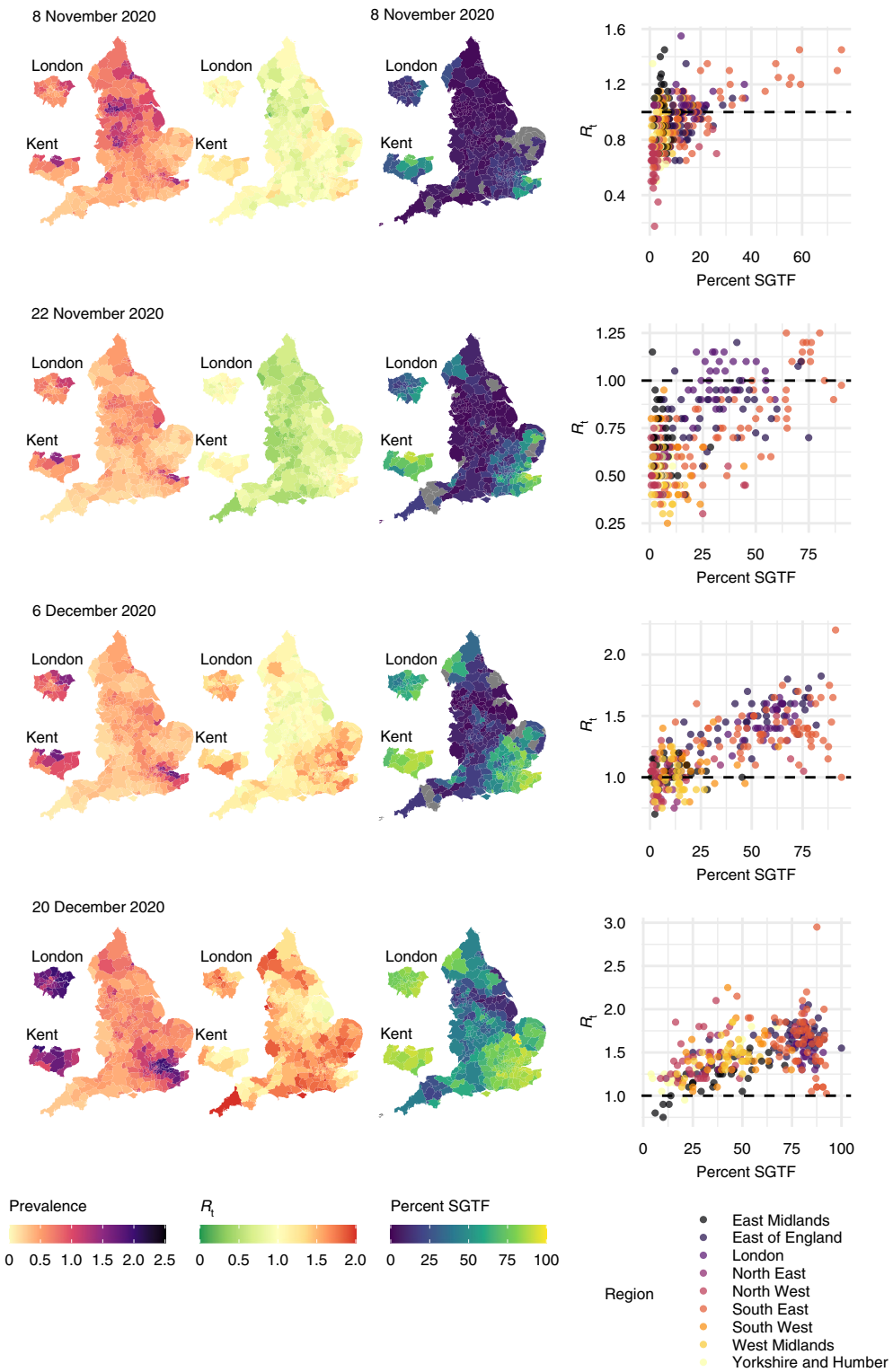


Fig. 6 | Maps of estimated prevalence, effective reproduction number and Alpha variant frequency. Maps of estimated local prevalence (left), estimated local R_t (middle) and frequency of SGTF (right), and scatter plots of SGTF frequency against estimated R_t (far right). Grey-coloured LTLAs denote missing data.

round 8, as we would expect, but the estimates remained highly accurate, with average bias of 0.05% (SE = 0.03; see also Extended Data Fig. 4j–l and compare vertical credible interval widths between Extended Data Fig. 4e and Extended Data Fig. 4k).

Predictive ability of R_t estimates. R_t measures whether the number of infectious individuals is increasing, $R_t > 1$, or decreasing, $R_t < 1$, in the population at time point t. Extended Data Fig. 5 compares LTLA R_t estimates with the future change in local case numbers. For validation purposes, here we are performed one-step-ahead at a time prediction and compared predictions with out-of-training-sample observed statistics (fold-change in raw case numbers from baseline). The results were stratified according to baseline case numbers, and we examined predictions 1 week and 2 weeks ahead. Each point

corresponds to an (LTLA, week) pair, and the results are for the period 18 October 2020 to 20 June 2021. Across each of the six scenarios presented, there is strong evidence of an association between R_t and future change in case numbers ($P < 2 \times 10^{-16}$). The strength of association between R_t and 1-week-ahead case numbers has Spearman's $\rho = 0.73$ for the high baseline case group (>500 cases per 100,000), which decreased to $\rho = 0.29$ in the low baseline group (≤200 cases per 100,000). The association remained strong when predicting caseloads 2 weeks ahead, with, for example, $\rho = 0.73$ (Spearman's) for the high baseline case group.

Comparison of effective reproduction number estimates from the debiasing approach with estimates from other studies. We extracted estimates of R_t based on our debiasing model likelihood implemented within a standard SIR model, illustrated in Extended Data Fig. 1. We compared the results to the local R_t estimates outputted by at the Imperial College COVID-19 website[32]. A cross-method comparison of longitudinal traces of R_t for a subset of LTLAs is shown in Extended Data Fig. 6. Encouragingly for both approaches, the estimates generally displayed good concordance, with credible intervals overlapping appropriately, despite being based on different data and models (Supplementary Table 1).

Discussion

The current standard practice internationally is to summarize SARS-CoV-2 infection rates by counting the number of individuals testing positive in a local area over a period of time, typically 1 week. The resulting statistic—cases per 100,000—is used to characterize and monitor the spatiotemporal state of an epidemic alongside other epidemiological measures such as R_t. Problematically, however, interpreting cases per 100,000 is not straightforward, as the data are subject to a number of unknown biasing influences such as (1) variation in testing capacity, (2) ascertainment bias on who is (self)-selected to be tested and (3) imperfect sensitivity and specificity of antigen tests. These factors, among others, make it difficult to quantify the true underlying local incidence or prevalence of SARS-CoV-2 infection, which places a burden on policymakers implicitly to adjust for such biases themselves. To address this problem, we developed an integrative causal model that can be used to debias raw case numbers and accurately estimate the number of individuals with infection in a local area.

The flexible statistical framework allows simultaneous and coherent incorporation of a number of important features. First, it corrects for ascertainment bias that result from preferential testing based on symptom status or on other confounders. This accounts for any variation in testing capacity by modelling the total number of tests conducted locally. Second, it can incorporate the use of different SARS-CoV-2 testing assays, such as LFD and PCR, including adjustment for particular sensitivity and specificity. Third, it infers the number of infectious individuals, while PCR tests may also pick up positive individuals at non-infectious stages. Finally, the model outputs week-specific debiased prevalence with uncertainty (via a marginal likelihood), which allows modular interoperability with other models. We illustrated this with a SIR epidemic model implementation that estimated local transmission rates while accounting for vaccine- and disease-induced immunity in the population. Our modelling work illustrates the benefits of having both a rolling randomized surveillance survey and targeted testing (for example, of frontline healthcare staff and symptomatic individuals). While targeted testing is routinely collected internationally, the United Kingdom has led the way in introducing regular national surveillance randomized surveys such as REACT[7,8] and ONS CIS[6]. Ongoing international pandemic preparedness can benefit from sampling designs that combine random sampling with targeted testing so that they can most powerfully complement and strengthen one another. Our model depends on the availability of randomized surveillance data. Future studies from other countries and collaborations with local experts will show and may further validate the breadth of utility of our debiasing framework and how it contributes towards global public health responses.

Since randomized surveillance data are currently rare internationally, there would be utility in extending the causal framework to address situations where targeted testing is accompanied by semi-randomized data with a well-known selection process (such as routine tests for healthcare workers, in care homes or regular testing at schools). Extending the current framework would begin with careful empirical exploration of the relationship between test positivity rate in such semi-randomized settings and comparable local prevalence (for example, in relevant age strata). The wealth of data available in the United Kingdom provides a good starting point for such exploratory work, which can be used to develop more complex causal models transferable to new semi-randomized contexts.

Methods

Ethics approval. The Alan Turing Institute Ethics Advisory Group provided guidelines for this study's procedures and advised that Health Research Authority approval is not required for this research.

Observational models for surveillance data. The primary target of inference is prevalence, I out of M, being the unknown number of infectious individuals at a particular time point in the local population of known size M. Our method estimates two types of prevalence: (1) the number of individuals that would test PCR positive ($\tilde{I}$) and (2) the number of individuals that are infectious (I). See below ("Focusing prevalence on the infectious subpopulation"), where we clarify the distinction between the PCR-positive and infectious subpopulations, and how we target the latter.

Temporal resolution of test count data. We applied the debiasing framework to test-count data aggregated into non-overlapping weeks. This has two clear advantages. First, by aggregating to weekly level data, we obviate the need to account for weekday effects that can be driven, for example, by logistical constraints or by individuals self-selecting to submit samples more readily on some weekdays than on others. Second, fitting a weekly model is computationally less intensive than fitting a model to daily test counts. The potential disadvantage of binning data by week is that high-frequency effects cannot be detected. Although it is possible in principle to adapt the framework to analyse daily testing data, we note that daily variation is likely to be confounded by weekday testing effects and so may be difficult to detect and interpret. Furthermore, while we use non-overlapping weekly data for model fitting, it is possible to output rolling weekly estimates, particularly to obtain as up-to-date prevalence estimates as are permitted by the data. However, we note that complete testing data are typically subject to a reporting lag of 4–5 days[33].

Randomized surveillance data, u of U. Suppose that out of a total U randomized surveillance (for example, REACT and ONS CIS) tests, we observe u positive tests. The randomized testing (for example, REACT and ONS CIS) likelihood is

$$\mathbb{P}(u \text{ of } U \mid \tilde{I}) = \text{Hypergeometric}(u \mid M, \tilde{I}, U), \quad (2)$$

and this allows direct, accurate statistical inference on $\tilde{I}$, the proportion of the population that would return a positive PCR test.

Focusing prevalence on the infectious subpopulation. PCR tests are sensitive and can detect the presence of SARS-CoV-2 both days before and weeks after an individual is infectious. It is usually desirable for prevalence to represent the proportion of a population that is infectious. We can obtain a likelihood for the number of infectious individuals I as follows:

$$\mathbb{P}(n \text{ of } N \mid I) = \int \mathbb{P}(n \text{ of } N \mid \tilde{I}) \mathbb{P}(\tilde{I} \mid I) \mathrm{d}\tilde{I}, \quad (3)$$

where I and $\tilde{I}$ are the number of infectious and PCR-positive individuals, respectively.

The conditional distribution $\mathbb{P}(\tilde{I} \mid I)$ can be specified on the basis of external knowledge of the average length of time spent PCR-positive versus infectious. Our approach to estimating this quantity imports information on the timing of COVID-19 transmission[34] and the interval of PCR positivity in individuals with SARS-CoV-2 infection[35]. More precisely, we specified the infectious time interval for an average individual with infection in the population to span the interval 1–11 days after infection (the empirical range of generation time from fig. 1A of ref. [34]). We then calculated the posterior probability of a positive PCR occurring 1–11 days after infection (fig. 1A of ref. [35]). We incorporated the effects of changing incidence

in the calculations; this is important because, for example, if incidence is rising steeply, the majority of people who would test PCR positive in the population are those that are relatively recently infected. Full details can be found in Supplementary Information "PCR positive to infectious mapping—method details".

Targeted surveillance data, n of N. In contrast to the randomized surveillance likelihood in equation (2), the targeted likelihood can be expressed in terms of the observation of n of N positive targeted (for example Pillar 1+2) tests as follows:

$$\begin{aligned}\mathbb{P}(n \text{ of } N \mid I, \delta, \nu) &= \text{Binomial}\left(n \mid I,\ \mathbb{P}(\text{ tested} \mid \text{infected})\right)\\ &\quad \times \text{Binomial}(N-n \mid M-I,\ \mathbb{P}(\text{ tested} \mid \text{not infected})),\end{aligned} \tag{4}$$

where $\mathbb{P}(\text{ tested} \mid \text{infected})$ and $\mathbb{P}(\text{tested} \mid \text{not infected})$ are the probabilities of an individual with infection (respectively, individual without infection) being tested.

Bias parameters, δ and ν. We introduce the following parameters:

$$\delta := \log\left(\frac{\text{odds}(\text{ tested} \mid \text{infected})}{\text{odds}(\text{ tested} \mid \text{not infected})}\right) \tag{5}$$

$$\nu := \log \text{odds}(\text{ tested} \mid \text{not infected}), \tag{6}$$

leading to the targeted swab testing likelihood being represented as

$$\begin{aligned}\mathbb{P}(n \text{ of } N \mid I, \delta, \nu) &= \text{Binomial}\left(n \mid I,\ \text{logit}^{-1}(\delta+\nu)\right)\\ &\quad \times \text{Binomial}(N-n \mid M-I,\ \text{logit}^{-1}\nu).\end{aligned} \tag{7}$$

The unknown parameter that requires special care to infer is δ, that is, the log odds-ratio of being tested in the infected subpopulation versus in the non-infected subpopulation. The other parameter, ν, is directly estimable from the targeted data: $\hat{\nu} := \text{logit}[(N-n)/M]$ is a precise estimator with little bias when prevalence is low.

Test sensitivity and specificity. The likelihood in equation (7) assumes a perfect antigen test. If the test procedure has false-positive rate α, and false-negative rate β, the targeted likelihood is instead

$$\mathbb{P}(n \text{ of } N \mid I, \delta, \nu) = \sum_{z=0}^{\min\{I,N\}} \mathbb{P}(z \text{ of } N \mid I, \delta, \nu)\mathbb{P}(n \mid z \text{ of } N), \tag{8}$$

where z denotes the unknown number of individuals who truly have an infection that were tested. The first term in the sum in equation (8) is obtained by substituting z in equation (7), while the second term is

$$\mathbb{P}(n \mid z \text{ of } N) = \sum_{n_\beta=\max\{0,z-n\}}^{\min\{z,N-n\}} \text{Binomial}(n_\beta \mid z,\ \beta)\ \text{Binomial}(n_\beta + n - z \mid N-z,\ \alpha), \tag{9}$$

with n_β denoting the number of false-negative test results. An analogous adjustment can be made to the randomized surveillance likelihood in equation (2).

Cross-sectional inference on local prevalence. We leveraged spatially coarse-scale randomized surveillance data to specify an EB prior on bias parameters $p(\delta)$ at coarse-scale (PHE region), and thereby accurately infer prevalence from targeted data at fine scale (LTLA j within PHE region J_j). We explicitly use the superscripts LTLA (j) in PHE region (J_j) in step 4 below, where notation from both coarse and fine scale appear together. All quantities in steps 1–3 are implicitly superscripted (J_j), but these are suppressed for notational clarity. For computational efficiency, we handle prevalence in a reduced-dimension space of bins as described in Supplementary Information section "Interval-based prevalence inference—set-up and assumptions". The method in detail is as follows:

1. Infer prevalence from unbiased testing data. At a coarse geographic level (PHE region J_j), estimate prevalence from randomized surveillance data u_t of U_t. Represent the posterior at time t in mass function

$$\hat{p}_t(I_t) := \mathbb{P}(I_t \mid u_t \text{ of } U_t) \tag{10}$$

where $\hat{p}_t : \{0, \ldots, M\} \to [0, 1]$ need only be available at a subset $t \in \mathcal{T} \subseteq \{1, \ldots, T\}$ of time points.

2. Learn δ_t from accurate prevalence. At a coarse geographic level, for each $t \in \mathcal{T}$, we estimate bias parameter δ_t by coupling biased data n_t of N_t with accurate prevalence information $\hat{p}_t$. With ν_t fixed at $\hat{\nu}_t := \text{logit}[(N_t - n_t)/M]$

$$p(\delta_t \mid n_t \text{ of } N_t, \hat{p}_t, \hat{\nu}_t) = \sum_{I_t} p(\delta_t \mid n_t \text{ of } N_t, I_t, \hat{\nu}_t)\hat{p}_t(I_t) \tag{11}$$

$$\approx \text{N}(\delta_t \mid \hat{\mu}_t,\ \hat{\sigma}_t^2) \tag{12}$$

where a moment-matched Gaussian approximation is performed in equation (12) (we assessed the reasonableness of this approximation using diagnostic plots (Supplementary Fig. 2)). The posterior density in the sum in equation (11), $p(\delta_t \mid n_t \text{ of } N_t, I_t, \hat{\nu}_t)$ is conjugate under a Beta(a,b) prior on $\text{logit}^{-1}(\nu_t + \delta_t) \equiv \mathbb{P}(\text{ tested} \mid \text{infected})$, and so can be evaluated as follows (where BetaCDF is the cumulative distribution function of the beta distribution):

$$\mathbb{P}(\delta_t \le \text{logit}(x) - \hat{\nu}_t \mid n_t \text{ of } N_t, I_t, \hat{\nu}_t) = \text{BetaCDF}(x \mid n_t + a, I_t - n_t + b). \tag{13}$$

3. Specify smooth EB prior on $\delta_{1:T}$. A smooth prior on $\delta_{1:T}$ is specified as follows:

$$p(\boldsymbol{\delta}) \propto \text{N}(\boldsymbol{\delta} \mid 0, \Sigma_\delta) \prod_{t\in\mathcal{T}} \text{N}(\delta_t \mid \hat{\mu}_t, \hat{\sigma}_t^2) \prod_{t\notin\mathcal{T}} \text{N}(\delta_t \mid 0, \sigma^2_{\text{flat}}) \tag{14}$$

where $\text{N}(\boldsymbol{\delta} \mid \mathbf{0}, \boldsymbol{\Sigma}_\delta)$ imparts a user-specified degree of longitudinal smoothness, thereby sharing information on δ across time points. Ignorance of δ_t, in the absence of random surveillance data, is encapsulated in a Gaussian with large variance σ^2_{flat}. A standard choice for $\text{N}(\boldsymbol{\delta} \mid \mathbf{0}, \boldsymbol{\Sigma}_\delta)$ corresponds to a stationary autoregressive, AR(1), process of the form

$$\delta_t = c + \psi\delta_{t-1} + \varepsilon_t \tag{15}$$

with a diffuse Gaussian prior $c \sim \text{N}(0, \sigma^2_{\text{flat}})$ and with smoothing tuned by $0 < \psi < 1$ and white noise variance σ^2_ε. The normalized form of the prior in equation (14) is

$$p(\boldsymbol{\delta}) = \text{N}\left(\boldsymbol{\delta} \mid (\boldsymbol{\Sigma}_\delta^{-1} + D^{-1})^{-1}D^{-1}\hat{\boldsymbol{\mu}},\ (\boldsymbol{\Sigma}_\delta^{-1} + D^{-1})^{-1}\right) \tag{16}$$

with ($\hat{\mu}$, diagonal matrix $D_{T\times T}$) having elements ($\hat{\mu}_t$, $\hat{\sigma}_t^2$) for $t \in \mathcal{T}$ and (0, σ^2_{flat}) for $t \notin \mathcal{T}$.

4. Infer cross-sectional local prevalence from biased testing data. At a fine-scale geographic level (LTLA j in PHE region J_j), having observed $n_t^{(j)}$ of $N_t^{(j)}$ positive test results (a subset of the $n_t^{(J_j)}$ of $N_t^{(J_j)}$ observed at the coarse-scale level above), we calculated the posterior for $I_t^{(j)}$ separately at each time point t as follows:

$$p(I_t^{(j)} \mid n_t^{(j)} \text{ of } N_t^{(j)}) \propto p(I_t^{(j)})p(n_t^{(j)} \text{ of } N_t^{(j)} \mid I_t^{(j)}, \hat{\nu}_t^{(j)}) \tag{17}$$

$$= p(I_t^{(j)}) \int_{\delta_t^{(J_j)}} p(n_t^{(j)} \text{ of } N_t^{(j)} \mid I_t^{(j)}, \hat{\nu}_t^{(j)}, \delta_t^{(J_j)}) p(\delta_t^{(J_j)})\, d\delta_t^{(J_j)} \tag{18}$$

where $\hat{\nu}_t^{(j)} := \text{logit}[(N_t^{(j)} - n_t^{(j)})/M_t^{(j)}]$, the likelihood in the integral in equation (18) is available in equation (7), and the prior $p(\delta_t^{(J_j)})$ is time point t's marginal Gaussian from equation (16).

Debiasing LFD tests with PCR surveillance (or vice versa). The methods can be adapted in a straightforward manner to the situation in which the randomized surveillance study uses a different assay to the targeted testing. For a concrete example, we could use REACT PCR prevalence posterior $\hat{p}_t(\tilde{I}_t)$ from equation (10) to debias Pillar 1+2 LFD test data n_t of N_t. Equation (11) can be adjusted to estimate the ascertainment bias δ pertaining to LFD data as follows:

$$p(\delta_t \mid n_t \text{ of } N_t, \hat{p}_t, \hat{\nu}_t) = \sum_{\bar{I}_t} \{p(\delta_t \mid n_t \text{ of } N_t, \bar{I}_t, \hat{\nu}_t) \sum_{\tilde{I}_t} \mathbb{P}(\bar{I}_t \mid \tilde{I}_t)\hat{p}_t(\tilde{I}_t)\}, \tag{19}$$

where $\bar{I}_t$ and $\tilde{I}_t$ are the unobserved LFD- and PCR-positive prevalence, respectively, and the conditional distribution $\mathbb{P}(\bar{I}_t \mid \tilde{I}_t)$ can be estimated on the basis of external knowledge of the average length of time spent PCR-positive versus LFD-positive, analogously to as described in above in "Focusing prevalence on the infectious subpopulation". The remaining computations, from equation (12) onwards, are unchanged, with the outputted fine-scale marginal likelihood $p(n_t^{(j)} \text{ of } N_t^{(j)} \mid I_t^{(j)}, \hat{\nu}_t^{(j)})$ in equation (17) to be interpreted as targeting the local LFD-positive prevalence $\bar{I}_t^{(j)}$.

Full Bayesian inference under a stochastic SIR epidemic model. The cross-sectional analysis described above in "Cross-sectional inference on local prevalence" generates the δ-marginalized likelihood, $p(n_t^{(j)} \text{ of } N_t^{(j)} \mid I_t^{(j)}, \hat{\nu}_t)$ in equation (17), at each time point for which targeted data are available. These likelihoods can be used as input for longitudinal models to obtain better prevalence estimates and to infer epidemiological parameters such as R_t.

We illustrate this via a Bayesian implementation of a stochastic epidemic model whereby individuals become immune through population vaccination and/or exposure to COVID-19 (Supplementary Fig. 1). We incorporate known population vaccination counts into a standard discrete time Markov chain SIR model (ref. [36], chapter 3). Details of the transition probability calculations are given in the Supplementary Information sections "SIR model details" and "SIR model—discussion, assumptions and caveats".

Priors on R, I and R$^+$. We place priors on I, R^+ measured as a proportion of the population; this proportion then gets mapped to prevalence intervals on subpopulation counts as described in "Interval-based prevalence inference—set-up and assumptions" in the Supplementary Information. Specifically, we use truncated, discretized Gaussian distributions on the proportion of the population who are immune and infectious. For example, on the number of infectious individuals I_t at each time point t, we specify the prior (suitably normalized over its support)

$$\mathbb{P}(I_t = j) \propto \int_{(j-1)/M}^{j/M} \mathrm{N}\left(x|\mu_I, \sigma_I^2\right) \mathrm{d}x \text{ for } j/M \in [p_{\min}, \ldots, p_{\max}], \quad (20)$$

with an example weakly informative hyperparameter setting being $\mu_I = 0.5\%$, $\sigma_I = 1\%$, $p_{\min} = 0\%$, $p_{\max} = 4\%$. To ensure meaningful inference on $R^+_{1:T}$, we placed an informative prior that reflects the state of knowledge of the immune population size. We did this using an informative truncated Gaussian prior on R^+_1 and noninformative priors on $R^+_{2:T}$. We placed a noninformative uniform prior on each R_t, for example a Uniform(0.5, 2.5).

Markov chain Monte Carlo sampling implementation. We performed inference under the model represented in the DAG in Supplementary Fig. 1. The likelihood is marginalized with respect to δ, and we used Markov chain Monte Carlo to draw samples from the posterior

$$p(I, R^+, \mathcal{R} \mid n, N).$$

We sampled $\mathcal{R}$ and (I, R^+) using separate Gibbs updates. For sampling (I, R^+), we represented the joint full conditional as

$$p(I, R^+|\mathcal{R}, n, N) = p(I \mid \mathcal{R}, n, N)p(R^+|I), \quad (21)$$

sampling I^{new} from $p(I \mid \mathcal{R}, n, N)$, and then $R^{+\text{new}}$ from $p(R^+ \mid I^{\text{new}})$.

Sampling from p(I | R, n, N). The sampling distribution on prevalence can be expressed as

$$\begin{aligned} p(I \mid \mathcal{R}, n, N) &\propto p(n, N \mid I, \mathcal{R})p(I \mid \mathcal{R}) \\ &= p(n_1, N_1 \mid I_1)p(I_1) \prod_{t=2}^{T} p(n_t \text{ of } N_t \mid I_t)p(I_t \mid I_{t-1}, \mathcal{R}_{t-1}), \end{aligned} \quad (22)$$

which is a hidden Markov model with emission probabilities taken from the δ-marginalized likelihood in equation (18), and transition probabilities taken from equation (37) (Supplementary Information).

Sampling from p(R$^+$*|I).* We expressed the full conditional for $\Delta R^+_{1:T}$ as

$$\mathbb{P}(R^+_{1:T} \mid I_{1:T}) \propto \mathbb{P}(R^+_1 \mid V_1) \prod_{t=2}^{T} \mathbb{P}(R^+_t \mid R^+_{t-1}, I_{t-1}, \Delta V_t)$$

and sampled the $\Delta R^+_{1:T}$ sequentially, with $\mathbb{P}(R^+_t \mid R^+_{t-1}, I_{t-1}, \Delta V_t)$ available in equation (39) (Supplementary Information).

Sampling from p(R | I). The prior joint distribution of $\mathcal{R}_{1:T}$ was modelled using a random walk as follows:

$$\mathcal{R}_t \sim \mathrm{N}(\mathcal{R}_{t-1}, \sigma^2_{\mathcal{R}}), \quad (23)$$

where $\sigma^2_{\mathcal{R}}$ is a user-specified smoothness parameter.

The update involves sampling from

$$p(\mathcal{R} \mid I) = p(\mathcal{R}_1) \prod_{t=2}^{T-1} p(\mathcal{R}_t \mid \mathcal{R}_{t-1}) \prod_{t=2}^{T} p(I_t \mid I_{t-1}, \mathcal{R}_{t-1}). \quad (24)$$

We discretized the space of R_t into an evenly spaced grid and sample from the hidden Markov model defined in equation (24)[37]. The transition probabilities are given by equation (23) (suitably normalized over the discrete R_t space) and the emission probabilities given by equation (37) (Supplementary Information).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data underlying the Alpha VoC 202012/01 analysis were accessed via the UK Health Security Agency Data Science Hub (DaSH) data platform; they are not publicly available and can only be accessed using approved UK government email domains such as @test-and-trace.nhs.uk. For the remainder of the results presented here, the data are publicly available. Randomized surveillance data comes from the REACT study[7,8] (https://github.com/mrc-ide/reactidd/tree/master/inst/extdata).

From REACT, we create weekly test counts at the spatially coarse-scale level (PHE region) and, for validation purposes but not model fitting, use round-aggregated counts at the fine-scale level (LTLA), for rounds 7–11. The combined weekly Pillar 1+2 data are publicly available for download (https://www.gov.uk/government/publications/nhs-test-and-trace-england-statistics-14-january-to-20-january-2021; note that LFD results are not included in these weekly summaries). We downloaded R_t estimates outputted by the Imperial College team's Epidemia model[38,39] from https://imperialcollegelondon.github.io/covid19local/downloads/UK_hotspot_Rt_estimates.csv on 13 October 2021, and we provide a copy of that downloaded file in our Zenodo repository at https://doi.org/10.5281/zenodo.5784718.

Code availability

The R scripts[40] used to generate the results in this manuscript are available in the following Git repository: https://github.com/alan-turing-institute/jbc-turing-rss-testdebiasing.

Received: 18 October 2021; Accepted: 18 November 2021;
Published online: 31 December 2021

References

1. *PHE Data Series on Deaths in People with COVID-19: Technical Summary—12 August Update* (Public Health England, 2020).
2. The Official UK Government Website for Data and Insights on Coronavirus (COVID-19) (GOV.UK, accessed 15 February 2021); https://coronavirus.data.gov.uk
3. Summary of Effectiveness and Harms of NPIs. *Scientific Advisory Group for Emergencies* (21 September 2020); https://www.gov.uk/government/publications/ summary-of-the-effectiveness-and-harms-of-different- non-pharmaceutical-interventions-16-september-2020
4. Prime Minister Announces New local COVID Alert Levels. *Prime Minister's Office, 10 Downing Street* (12 October 2020); https://www.gov.uk/government/news/ prime-minister-announces-new-local- covid-alert-levels
5. COVID-19 Response—Spring 2021 (Summary). *Cabinet Office* (22 February 2021); https://www.gov.uk/government/ publications/covid-19-response-spring-2021/ covid-19-response-spring-2021-summary
6. Pouwels, K. B. et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *Lancet Public Health* **6**, e30–e38 (2021).
7. Riley, S. et al. Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study. Preprint at *medRxiv* https://doi.org/10.1101/2020.07.10.20150524 (2020).
8. Chadeau-Hyam, M. et al. REACT-1 study round 14: High and increasing prevalence of SARS-CoV-2 infection among school-aged children during September 2021 and vaccine effectiveness against infection in England. Preprint at *medRxiv* https://doi.org/10.1101/2021.10.14.21264965 (2021).
9. COVID-19 Testing Data: Methodology Note. *Department of Health and Social Care* (21 August 2020); https://www.gov.uk/government/publications /coronavirus-covid-19-testing-data-methodology /covid-19-testing-data-methodology-note
10. Byambasuren, O. et al. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *Off. J. Assoc. Med. Microbiol. Infect. Dis. Can.* **5**, 223–234 (2020).
11. Subramanian, R., He, Q. & Pascual, M. Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proc. Natl Acad. Sci. USA* **118**, e2019716118 (2021).
12. Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J. & Thompson, S. G. Modelling bias in combining small area prevalence estimates from multiple surveys. *J. R. Stat. Soc. Ser. A* **174**, 31–50 (2011).
13. Giorgi, E., Sesay, S. S. S., Terlouw, D. & Diggle, P. J. Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *J. R. Soc. Stat. Soc. Ser. A* **178**, 445–464 (2015).
14. Amoah, B., Diggle, P. J. & Giorgi, E. A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. *Biometrics* **76**, 158–170 (2020).
15. Crainiceanu, C. M., Diggle, P. J. & Rowlingson, B. Bivariate binomial spatial modeling of loa loa prevalence in tropical africa. *J. Am. Stat. Assoc.* **103**, 21–37 (2008).
16. Pearl, J. *Causality* (Cambridge Univ. Press, 2009).
17. Hernán, M. A. & Robins, J. M. *Causal Inference: What if* (Chapman & Hall/CRC, 2010).
18. Birrell, P., Blake, J., van Leeuwen, E., Gent, N. & De Angelis, D. Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave. *Philos. Trans. R. Soc. B Biol. Sci.* https://doi.org/10.1098/rstb.2020.0279 (2021).
19. Irons, N. J. & Raftery, A. E. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proc. Natl Acad. Sci. USA* **118**, e2103272118 (2021).

20. Teh, Y. W. et al. *Efficient Bayesian inference of instantaneous reproduction numbers at fine spatial scales, with an application to mapping and nowcasting the Covid-19 epidemic in British local authorities* (UK Local Covid Map, 2021); https://localcovid.info/assets/docs/localcovid-writeup.pdf
21. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate timevarying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).
22. Flaxman, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
23. Jewell, C., Read, J., Roberts, G., Rowlington, B. & Suter, C. *Bayesian stoschastic model-based forecasting for spatial Covid-19 risk in England*. Technical Concept Note (GitHub, 2020); https://github.com/chrism0dwk/covid19uk/blob/master/doc/lancs_space_model_concept.pdf
24. Colman, E., Enright, J., Puspitarani, G. A. & Kao, R. R. Estimating the proportion of SARS-CoV-2 infections reported through diagnostic testing. Preprint at *medRxiv* https://doi.org/10.1101/2021.02.09.21251411 (2021).
25. Abbott, S. et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Technical Report. Preprint at *Wellcome Open Research* https://doi.org/10.12688/wellcomeopenres.16006.2 (2020).
26. Anderson, R. et al. Reproduction number (*R*) and growth rate (*r*) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. *Royal Society* https://royalsociety.org/-/media/policy/projects/set-c/set-covid-19-R-estimates.pdf (2020).
27. Funk, S. et al. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. Preprint at *medRxiv* https://doi.org/10.1101/2020.11.11.20220962 (2020).
28. Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).
29. Campbell, F. et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance* **26**, 2100509 (2021).
30. Investigation of Novel SARS-COV-2 Variants of Concern: Technical Briefings. *Public Health England*; www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant- of-concern-20201201 (2020).
31. *Lineage Counts by Local Authority and Week* for England; https://covid19.sanger.ac.uk/downloads (Wellcome Sanger Institute COVID-19 Genomics Surveillance, 2021).
32. *COVID-19 United Kingdom*; https://imperialcollegelondon.github. io/covid19local/#map (Imperial College London, 2021).
33. Jersakova, R. et al. Bayesian imputation of COVID-19 positive test counts for nowcasting under reporting lag. Preprint at https://arxiv.org/abs/2103.12661 (2021).
34. Ferretti, L. et al. The timing of COVID-19 transmission. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.04.20188516 (2020).
35. Hellewell, J. et al. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. BMC Med. 19, https://doi.org/10.1186/s12916-021-01982-x (2021).
36. Brauer, F., van den Driessche, P. & Wu, J. *Mathematical Epidemiology*. *Mathematical Biosciences Subseries* (Springer, 2008).
37. Scott, S. L. Bayesian methods for hidden Markov models: recursive computing in the 21st century. *J. Am. Stat. Assoc.* **97**, 337–351 (2002).
38. Mishra, S. et al. A COVID-19 model for local authorities of the United Kingdom. Preprint at *medRxiv* https://doi.org/10.1101/2020.11.24.20236661 (2020).
39. Scott, J. A. et al. epidemia: modeling of epidemics using hierarchical Bayesian models. R package version 1.0.0 https://imperialcollegelondon.github.io/epidemia/ (2020).
40. R Core Team. R: A Language and Environment for Statistical Computing; https://www.R-project.org/ (R Foundation for Statistical Computing, 2021).

Acknowledgements

B.L. was supported by the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme (grant number EP/R018561/1) and gratefully acknowledges funding from Jesus College, Oxford. K.B.P. is 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 and the Huo Family Foundation. S.R. is supported by MRC programme grant MC_UU_00002/10, The Alan Turing Institute grant TU/B/000092, and the EPSRC Bayes4Health programme grant EP/R018561/1. M.B. acknowledges partial support from the MRC Centre for Environment and Health, which is currently funded by the Medical Research Council MR/S019669/1. G.N. and C.H. acknowledge support from the Medical Research Council Programme Leaders award MC_UP_A390_1107. C.H. acknowledges support from The Alan Turing Institute, Health Data Research, UK, and the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme grant. Infrastructure support for the Department of Epidemiology and Biostatistics is also provided by the NIHR Imperial BRC. Authors at the Alan Turing Institute and Royal Statistical Society Statistical Modelling and Machine Learning Laboratory gratefully acknowledge funding from the Joint Biosecurity Centre, a part of NHS Test and Trace within the Department for Health and Social Care. The computational aspects of this research were supported by the Wellcome Trust Core Award grant number 203141/Z/16/Z (to B.L.) and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, the Department of Health, the Joint Biosecurity Centre or PHE.

Author contributions

G.N., B.L. and C.H. conceived and designed the research. G.N., B.L., T.P., R.J., J.L., R.E.K. and A.-M.M. acquired, analysed or interpreted the data. G.N., B.L., R.J. and J.L. created new software used in the work. G.N., B.L., R.J., T.P., K.B.P., P.J.D., S.R., M.B. and C.H. wrote the paper.

Competing interests

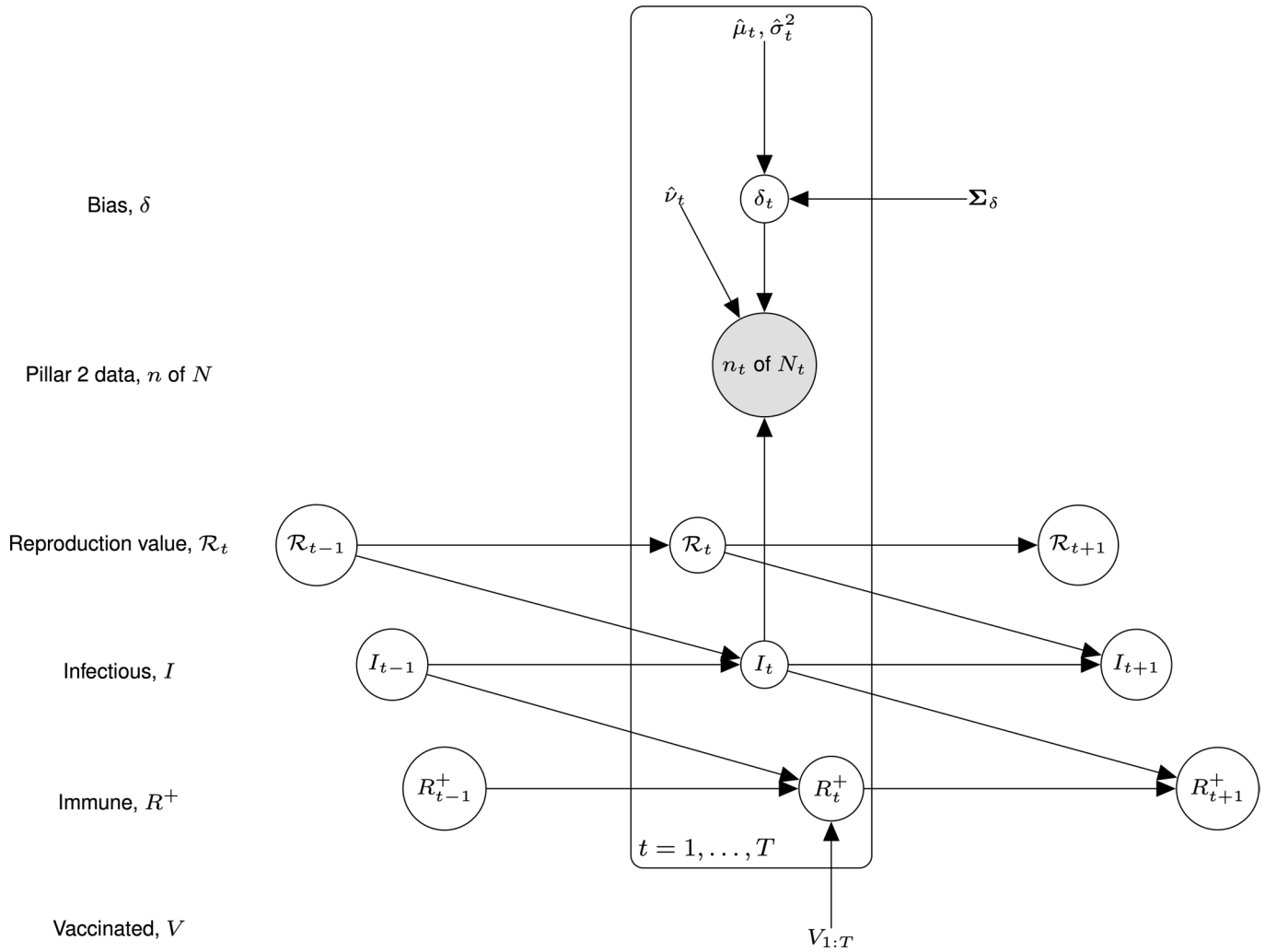
The authors declare no competing interests.

Additional information

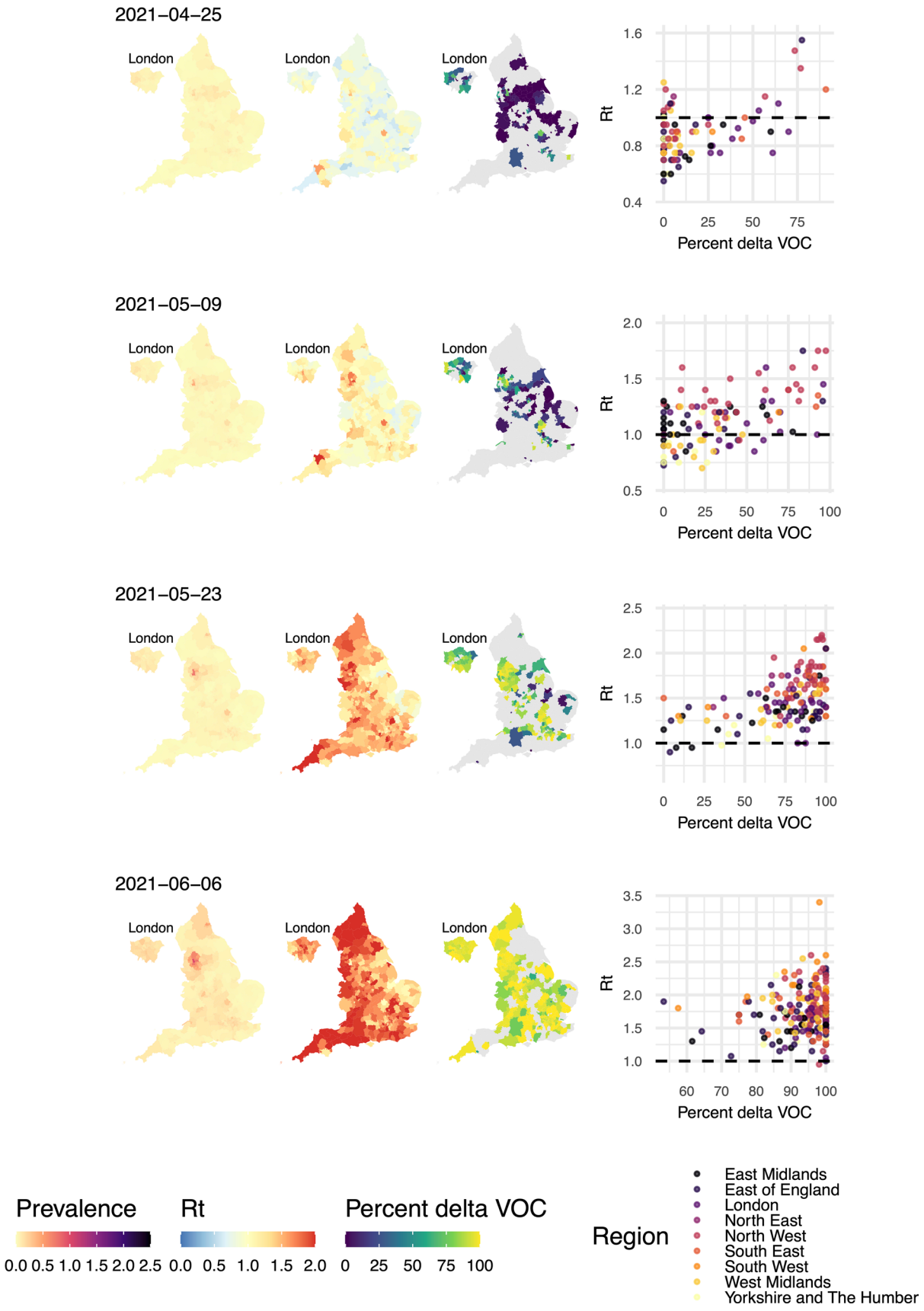
Extended data is available for this paper at https://doi.org/10.1038/s41564-021-01029-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-021-01029-0.

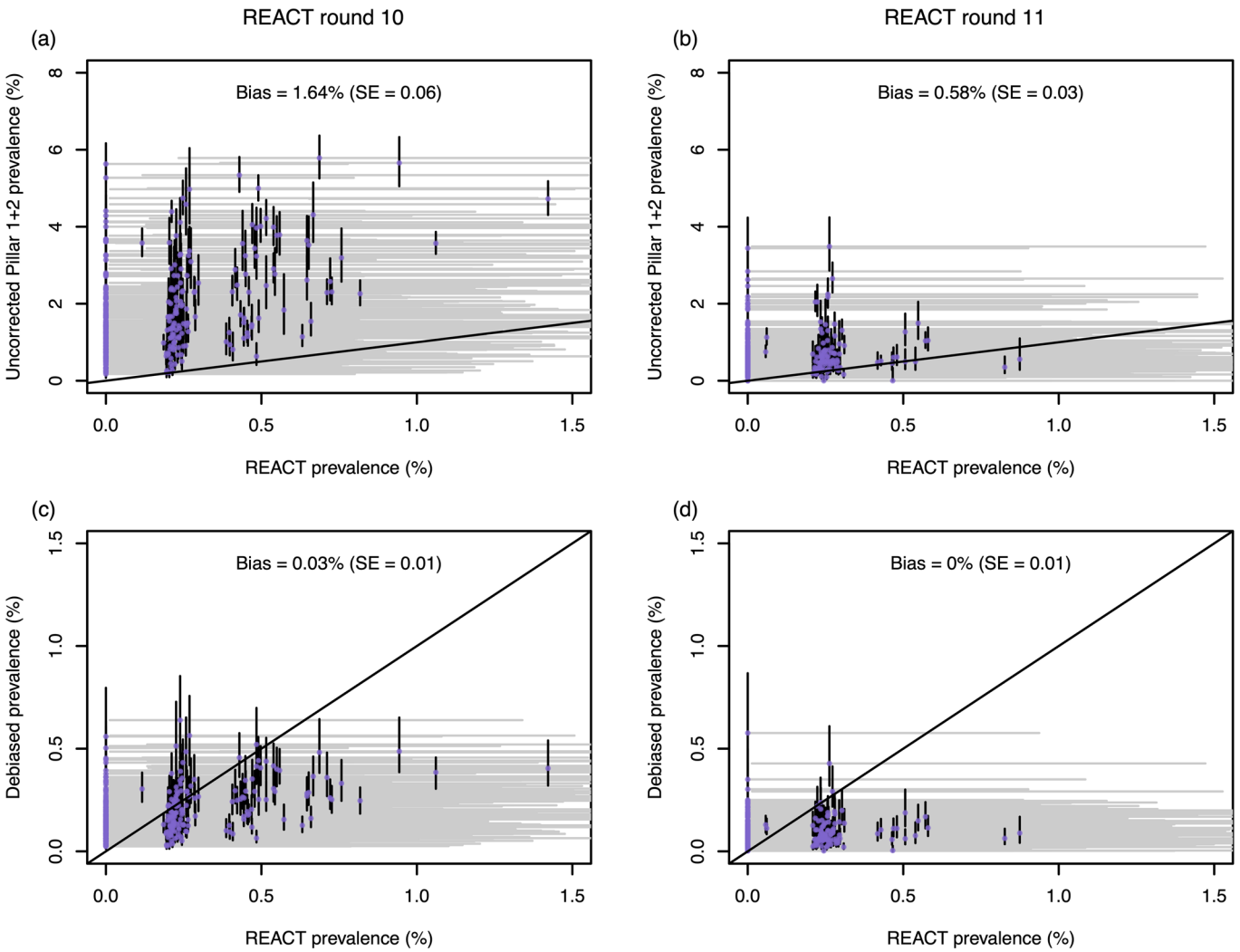
Correspondence and requests for materials should be addressed to George Nicholson, Brieuc Lehmann or Chris Holmes.



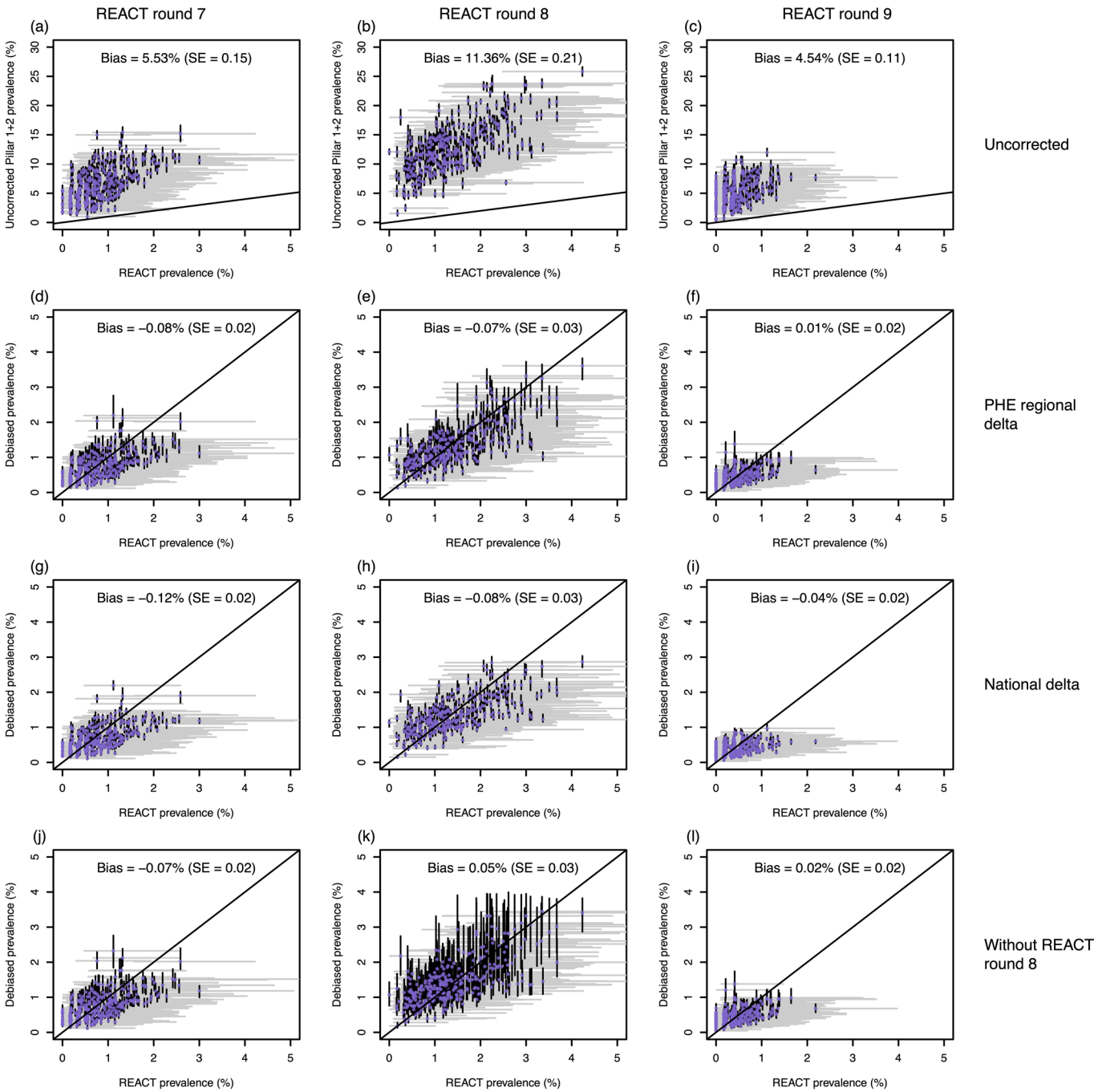
Extended Data Fig. 1 | Longitudinal model DAG for SIR epidemic model at local level (for example LTLA). Directed paths characterise conditional probability distributions, in contrast to the paths showing transitions between model compartments in Supplementary Fig. 1. Inference is for a region, for example an LTLA, based only on targeted test data collected in this region, n_t of N_t. A prior on δ_t parameterized ($\hat{\mu}_t$, $\hat{\sigma}_t^2$) brings information on the Pillar 1+2 ascertainment bias learned from randomized surveillance testing data available for the PHE region in which the LTLA lies. The $T \times T$ covariance matrix $\boldsymbol{\Sigma}_\delta$ ***imparts temporal smoothness on*** $\delta_{1:T}$. Effective reproduction numbers are denoted $\mathcal{R}_{1:T}$, number of infectious individuals by $I_{1:T}$, and the number of immune individuals by $R^+_{1:T}$.



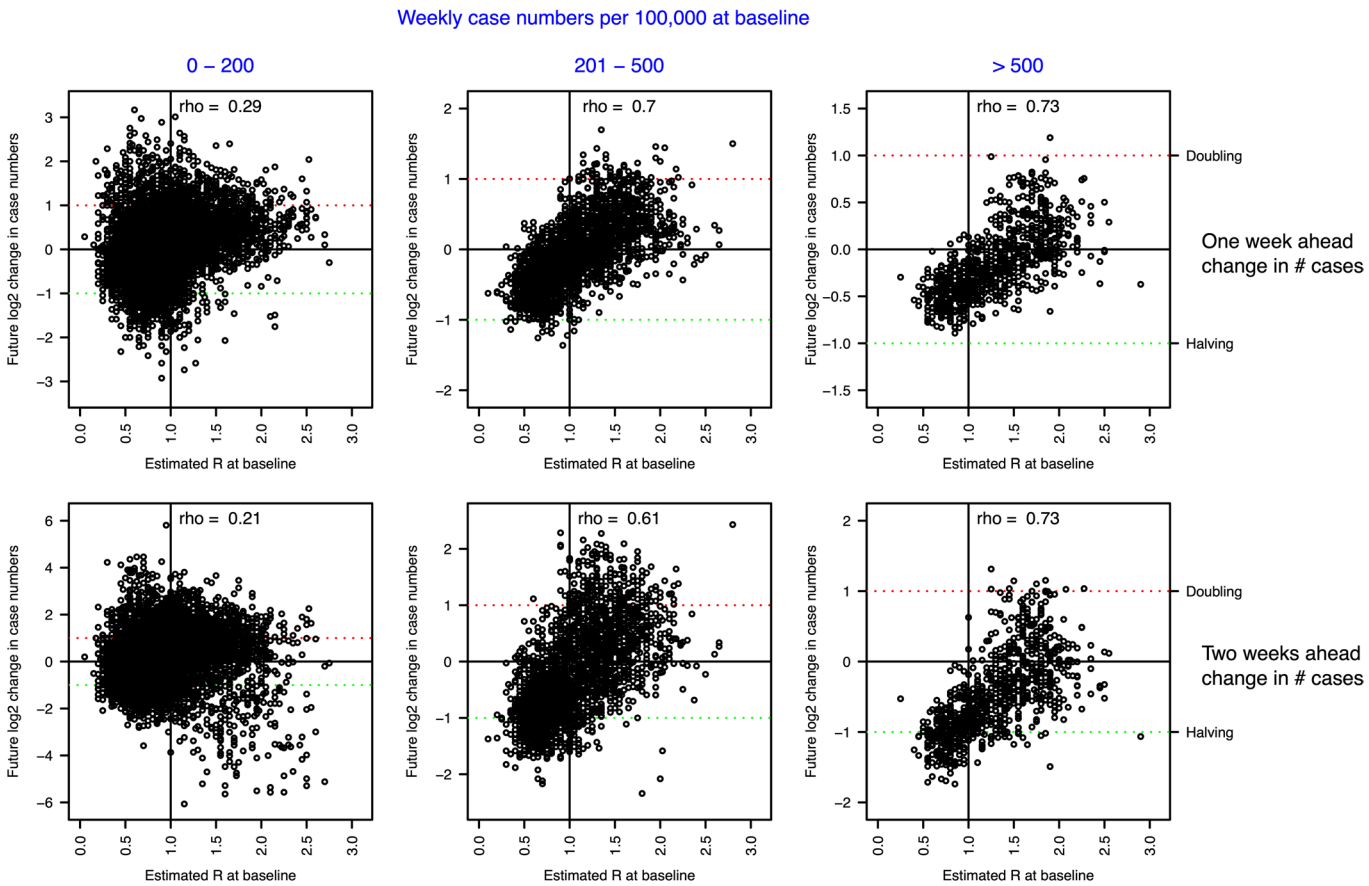
Extended Data Fig. 2 | Maps of estimated local prevalence (left), estimated local R_t (middle), and frequency of the delta variant (right), and scatter plot of Delta variant frequency against estimated R_t. Grey-coloured areas denote where the total number of variant sequencing assays performed (across all variants) is less than 10; in these cases the delta variant frequency estimates are omitted due to having high standard error.



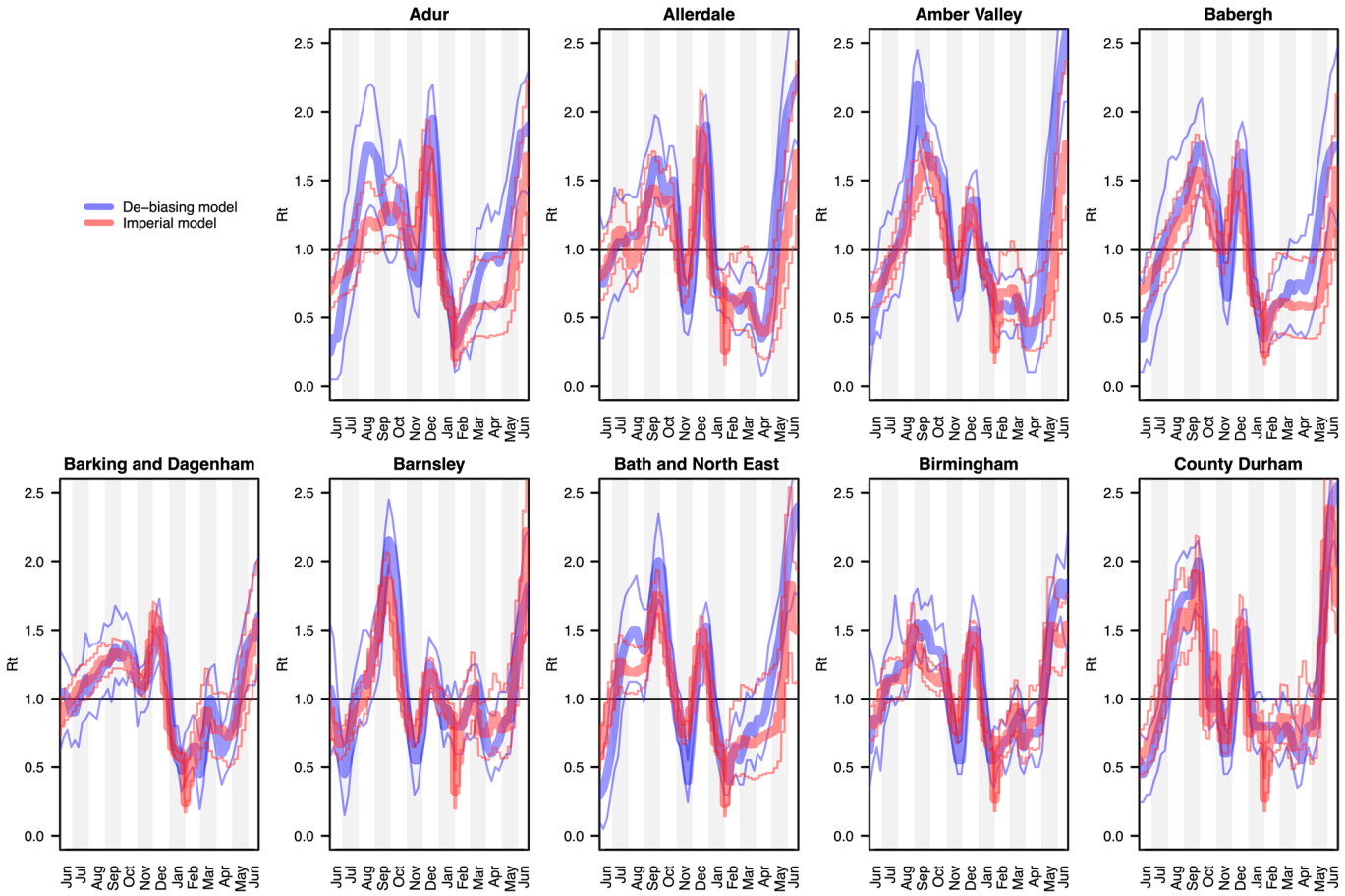
Extended Data Fig. 3 | Uncorrected (raw positivity rates) and corrected (debiased) Pillar 1+2 PCR-positive prevalence estimates against (gold-standard) REACT estimates from randomised surveillance for REACT rounds 10 and 11. Each point corresponds to an LTLA. Each scatter plot compares Pillar 1+2 prevalence estimates against unbiased estimates from the REACT study. Panels (a,c) show REACT round 10 data (11th Mar - 30th Mar 2021), and panels (b,d) show round 11 (15th Apr - 3rd May 2021). Uncorrected results are shown in panels (a-b) and bias-corrected cross-sectional estimates in (c-d). Horizontal grey lines are 95% exact binomial confidence intervals from the REACT data. Vertical black lines in panels (a-b) are 95% exact binomial confidence intervals from the raw, non-debiased Pillar 1+2 data. Vertical black lines in panels (c-d) are 95% posterior credible intervals from the debiased Pillar 1+2 data. Neither set of prevalence estimates has been corrected for false positives/negatives. Note that in panels (c-d), the CI widths are systematically tighter for the debiased Pillar 1+2 compared to the REACT data, pointing to the useful information content in debiased Pillar 1+2 data. The number of independent tests underlying each mean and (horizontal) CI for the REACT data varied between 289 and 1,894. The number of independent tests underlying each mean and (vertical) CI for the Pillar 1+2 data varied between 977 and 29,998.



Extended Data Fig. 4 | Uncorrected (raw positivity rates) and corrected (debiased) Pillar 1+2 PCR-positive prevalence estimates against (gold-standard) REACT estimates from *limited* randomised surveillance. Each point corresponds to an LTLA. Each scatter plot compares Pillar 1+2 prevalence estimates against unbiased estimates from the REACT study. Left to right the columns of panels show results from REACT round 7 (13th Nov - 3rd Dec 2020), round 8 (6th-22nd Jan 2021), and round 9 (4th-23rd Feb 2021). On the vertical axes: (a-c) show uncorrected test positivity rates; (d-f) show bias-corrected prevalence estimates; (g-i) show bias-corrected prevalence estimates where the bias δ was estimated at the ultra-coarse national level; and (j-l) show bias-corrected prevalence estimates where data from REACT round 8 was omitted, in order to assess the impact of a more limited randomised surveillance regime. Horizontal grey lines are 95% exact binomial confidence intervals from the REACT data. Vertical black lines in (a-c) are 95% exact binomial confidence intervals from the raw, non-debiased Pillar 1+2 data. Vertical black lines in panels (d-l) are 95% posterior credible intervals from the debiased Pillar 1+2 data. Neither set of prevalence estimates has been corrected for false positives/negatives. The number of independent tests underlying each mean and (horizontal) CI for the REACT data varied between 248 and 2,387. The number of independent tests underlying each mean and (vertical) CI for the Pillar 1+2 data varied between 1,117 and 42,458.



Extended Data Fig. 5 | Predicting future change in case numbers from current estimated $\mathcal{R}_t$. Each point corresponds to an (LTLA, week) pair, predicting future case numbers in the LTLA using $\mathcal{R}_t$ for that week. Future case numbers are represented by forward-in-time $\log_2$ fold change $\log_2(n_{t+k}/n_t)$. Case data underlying the plot are from the period 2020-10-18 - 2021-06-20. Note the number of points in each column differs based on how many LTLA-week pairs have baseline case numbers in the intervals in blue shown at the top of the plot.



Extended Data Fig. 6 | Comparison of $\mathcal{R}_t$ estimates between de-biasing model and Imperial model. For each of the nine PHE regions, we present the constituent LTLA whose name is ranked top alphabetically.

nature portfolio

Corresponding author(s): Prof. Chris Holmes

Last updated by author(s): Nov 11, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The R script used to download the data from public repositories prior to analysis in this manuscript is available at: https://github.com/alan-turing-institute/jbc-turing-rss-testdebiasing/scripts/00_download_data.R
Data analysis	The R scripts used to generate the results in this manuscript are available at: https://github.com/alan-turing-institute/jbc-turing-rss-testdebiasing

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

With the exception of the Alpha VOC 202012/01 analysis, all data underlying the results presented here are publicly available. Randomised surveillance data comes from the REACT study (data downloaded from https://github.com/mrc-ide/reactidd/tree/master/inst/extdata). From REACT, we aggregate weekly test counts at the spatially coarse-scale level (PHE region) and, for validation purposes but not model fitting, use round-aggregated counts at the fine-scale level (LTLA), for rounds 7 to 11. The combined weekly Pillar 1 and Pillar 2 data are publicly available for download (https://www.gov.uk/government/publications/nhs-test-and-trace-england-statistics-14-january-to-20-january-2021). Note that lateral flow test results are not included in the these weekly summaries. We downloaded Rt estimates

outputted by the Imperial College team's Epidemia model from https://imperialcollegelondon.github.io/covid19local/downloads/UK_hotspot_Rt_estimates.csv on 2021-10-13, and we provide a copy of that downloaded file in our Zenodo repository at <URL to be provided>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Methods for studying COVID-19 epidemiology by analysing community testing data (quantitative count data, comprising the number positive and total number tested)
Research sample	We collectively analyse two types of community testing data: The REal-time Assessment of Community Transmission (REACT) study is a nationally representative prevalence survey of SARS-CoV-2 based on repeated cross-sectional samples from a representative subpopulation defined via (stratified) random sampling from England's National Health Service patient register. Pillar 1 and Pillar 2 PCR test data form the main part of the UK government's national antigen testing strategy. Pillar 1 tests refer to "all swab tests performed in Public Health England (PHE) labs and National Health Service (NHS) hospitals for those with a clinical need, and health and care workers", and Pillar 2 comprises "swab testing for the wider population". The Pillar 1+2 research sample is dynamic, and is not representative of the population as a whole. In particular, Pillar 1+2 is enriched for NHS workers, individuals with symptoms, and those who self-select the testing, all of which can affect the demographic representation in the targeted Pillar 1+2 individuals. The REACT research sample is however designed to be nationally representative by stratified random sampling. We are able to use these REACT data to account for the ascertainment bias implicit in Pillar 1+2 data. For Pillar 1+2, the age and sex demographic breakdowns for the number of tests conducted are available here: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1007158/Demographic_LA_tables_Week_60.ods As an example, for the week commencing 2021-01-07, the age breakdown was as follows: 0-9yr: 83,864 (4%); 10-19yr: 127,482 (6%); 20-29yr: 394,029 (17%); 30-39yr: 406,058 (18%); 40-49yr: 364,772 (16%); 50-59yr: 397,950 (18%); 60-69yr: 223,665 (10%); 70-79yr: 119,226 (5%); 80-89yr: 93,704 (4%); 90+yr: 49,908 (2%). The sex split was Males: 886,970 (39%); Females: 1,373,688 (61%). For REACT, the age-stratified breakdown for round 8 is available here https://github.com/mrc-ide/reactidd/blob/master/inst/extdata/region_age_week_aggregated/round_8_go.csv. As an example, for REACT round 8, spanning 6th-22nd Jan 2021, the age breakdown is 5-12yr: 11,545 (7%); 13-17yr: 8,842 (5%); 18-24yr: 6,614 (4%); 25-34yr: 14,715 (9%); 35-44yr: 21,357 (13%); 45-54yr: 27,583 (17%); 55-64yr: 31,665 (19%); 65+yr: 43,246 (26%). Sex split metadata for REACT study are not publicly available. The rationale for choosing Pillar 1+2 positive and total counts is that these data are the most highly publicised case counts in England and they are regularly made publicly available, which allows ready reproducibility and ongoing implementation of our methodology. The rationale for choosing the REACT study is that we require randomised surveillance data in order to correct for ascertainment bias. REACT is one of the two major randomised surveillance studies, along with the office for National statistics COVID-19 infection survey, ONS CIS. We specifically use REACT data here because they are regularly made publicly available in raw form (positive and total counts) at the PHE region level, which is compatible with our downstream statistical modelling, and which allows transparency and reproducibility of our findings.
Sampling strategy	For the REACT data, participants were included in the tested group through stratified random sampling. For the Pillar 1+2 data, however, there is strong ascertainment bias, since infected individuals are more likely to be chosen for testing (e.g. frontline workers, symptomatics). In our paper we correct for this ascertainment bias in the Pillar 1+2 data by designing a causal model and thereby adjusting for the bias to obtain accurate estimates of local prevalence. We have used two datasets in our study: Pillar 1+2 and REACT. These have the advantage of being both publicly available (allowing reproducibility of our results), and we demonstrate in the paper that they are sufficient for us to develop, illustrate and, importantly, to validate our methodology (See section "Accuracy validation using ultra-coarse and incomplete data to estimate delta")
Data collection	In the REACT study, participants self-gathered throat and nose swab samples; no researcher was typically present during sample collection and other members of the public could have been present; participants were not blinded to the study hypothesis. Samples were then sent by post to a pre-specified laboratory for processing. For the Pillar 1+2 data, throat and nose swabs were gathered in various ways. For home testing, participants self-gathered throat and nose swab samples; no researcher was typically present during sample collection and other members of the public could have been present; participants were not blinded to the study hypothesis. In the case of regional or local test sites, or mobile testing units, swabs were gathered by a trained healthcare professional while the participant was seated in a motor vehicle; usually only the researcher and participant were present; neither the researcher nor the participant were blinded to the study hypothesis.
Timing	Between 31st May 2020 and 20th June 2021
Data exclusions	No data were excluded from the analysis

Non-participation	Over the course of the REACT study, there has been a considerable number of individuals who (i) are invited to participate but decline; or (ii) dropped out at some point during the study. We do not have access to these data.
Randomization	For the REACT data, participants were included in the tested group through stratified random sampling. For the Pillar 1+2 data there is a strong ascertainment effect since infected individuals are more likely to be chosen for testing (e.g. frontline workers, COVID-19 symptomatic individuals). In our paper we correct for this ascertainment bias in the Pillar 1+2 data by specifying a causal model to adjust for the bias and to obtain accurate estimates of local prevalence. We did not directly control for confounding covariates in our analysis of the Pillar 1+2 data; instead, we indirectly controlled for any potential confounders by estimating the marginal causal probabilities of being tested in the infected and non-infected groups (see Figure 1(a), in which socio-economic status is an example confounder).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	The REACT study approached a random sample of the population in England aged five years and above, using the National Health Service (NHS) records and invited them to join the study. There is some residual degree of non-response bias in the REACT study despite it being a randomised study; we illustrate this residual ascertainment effect in Figure 1(a) with a dashed blue arrow denoting non-response bias associated with socioeconomic status (SES). If metadata on SES were available, then this bias could be mitigated; we did not have access to SES data, and the likely impact on our results is that the debiased prevalence is mildly unduly weighted towards those strata in the population which are more likely to respond to REACT's invitation to participate. Individuals tested in the Pillar 1+2 data were recruited, self-selected or selected according to place of work, according to a number of potential criteria (e.g. NHS workers, those with COVID-19 symptoms); i.e. the Pillar 1+2 data harbour self-selection and other biases -- it is the purpose of the current work to adjust collectively for these biases.
Ethics oversight	The Alan Turing Institute Ethics Advisory Group

Note that full information on the approval of the study protocol must also be provided in the manuscript.