Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of Covid-19 in England

Katharine Sherratt, Sam Abbott*, Sophie R Meakin, Joel Hellewell, James D Munday, Nikos Bosse, CMMID Covid-19 working group, Mark Jit, Sebastian Funk*

Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine

** Equal contributors*

Abstract

The time-varying reproduction number (R_t: the average number secondary infections caused by each infected person) may be used to assess changes in transmission potential during an epidemic. While new infections are not usually observed directly, they can be estimated from data. However, data may be delayed and potentially biased. We investigated the sensitivity of R_t estimates to different data sources representing Covid-19 in England, and we explored how this sensitivity could track epidemic dynamics in population sub-groups.

We sourced public data on test-positive cases, hospital admissions, and deaths with confirmed Covid-19 in seven regions of England over March through August 2020. We estimated R_t using a model that mapped unobserved infections to each data source. We then compared differences in R_t with the demographic and social context of surveillance data over time.

Our estimates of transmission potential varied for each data source, with the relative inconsistency of estimates varying across regions and over time. R_t estimates based on hospital admissions and deaths were more spatio-temporally synchronous than when compared to estimates from all test-positives. We found these differences may be linked to biased representations of subpopulations in each data source. These included spatially clustered testing, and where outbreaks in hospitals, care homes, and young age groups reflected the link between age and severity of disease.

We highlight that policy makers could better target interventions by considering the source populations of R_t estimates. Further work should clarify the best way to combine and interpret R_t estimates from different data sources based on the desired use.

Keywords

covid-19, SARS-CoV-2, surveillance, bias, transmission, time-varying reproduction number

Background

Within six months of its emergence in late 2019, the novel coronavirus SARS-CoV-2 had caused over six million cases of disease (Covid-19) worldwide (1). Its rapid initial spread and high death rate prompted global policy interventions to prevent continued transmission, with widespread temporary bans on social interaction outside the household (2). Introducing and adjusting such policy measures depends on a judgement in balancing continued transmission potential with the multidimensional consequences of interventions. It is therefore critical to inform the implementation of policy measures with a clear and timely understanding of ongoing epidemic dynamics (3,4).

In principle, transmission could be tracked by directly recording all new infections. In practice, real-time monitoring of the Covid-19 epidemic relies on surveillance of indicators that are subject to different levels of bias and delay. In England, widely available surveillance data across the population includes: 1) the number of positive tests, biased by changing test availability and practice, and delayed by the time from infection to symptom onset (if testing is symptom-based), from symptom onset to a decision to be tested and from test to test result; 2) the number of new hospital admissions, biased by differential severity that triggers care seeking and hospitalisation, and additionally delayed by the time to develop severe diseases; and 3) the number of new deaths due to Covid-19, biased by differential risk of death and the exact definition of a Covid-19 death, and further delayed by the time to death.

Each of these indicators provides a different view on the epidemic and therefore contains potentially useful information. However, any interpretation of their behaviour needs to reflect these biases and lags and is best done in combination with the other indicators. One approach that allows this in a principled manner is to use the different data sets to separately track the time-varying reproduction number, R_t, the average number of secondary infections generated by each new infected person (5). Because R_t quantifies changes in infection levels, it is independent of the level of overall ascertainment as long as this does not change over time or is explicitly accounted for (6). At the same time, the underlying observations in each data source may result from different lags from infection to observation. However, if these delays are correctly specified then transmission behaviour over time can be consistently compared via estimates of R_t.

Different methods exist to estimate the time-varying reproduction number, and in the UK a number of mathematical and statistical methods have been used to produce estimates used to inform policy (7–9). Empirical estimates of R_t can be achieved by estimating time-varying patterns in transmission events from mapping to a directly observed time-series indicator of infection such as reported symptomatic cases. This can be based on the probabilistic assignment of transmission pairs (10), the exponential growth rate (11), or the renewal equation (12,13). Alternatively, R_t can be estimated via mechanistic models which explicitly compartmentalise the disease transmission cycle into stages from susceptible through exposed, infectious, and recovered (14,15). This can include accounting for varying population structures and context-specific biases in observation processes, before fitting to a source of observed cases. Across all methods, key parameters include the time after an infection to the onset of symptoms in the infecting and infected, and the source of data used as a reference point for earlier transmission events (16,17).

In this study, we used a modelling framework based on the renewal equation, adjusting for delays in observation to estimate regional and national reproduction numbers of SARS-Cov-2 across England. The same method was repeated for each of three sources of data that are available in real time. After assessing differences in R_t estimates by data source, we explored why this variation may exist. We compared the divergence between R_t estimates with spatio-temporal variation in case detection, and the proportion at risk of severe disease, represented by the age distribution of test positive cases and hospital admissions and the proportion of deaths in care homes.

Methods

Data management

Three sources of data provided the basis for our R_t estimates. Time-series case data were available by specimen date of test. This was a de-duplicated dataset of Covid-19 positive tests notified from all National Health Service (NHS) settings (Pillar One of the UK Government's testing strategy)(18) and by commercial partners in community settings outside of healthcare (Pillar Two). Hospital admissions were also available by date of admission if a patient had tested positive prior to admission, or by the day preceding diagnosis if they were tested after admission. Death data were available by date of death and included only those which occurred within 28 days of a positive Covid-19 test in any setting. All data were publicly available and taken from the UK government source (19,20), and were aggregated to the seven English regions used by the NHS.

To provide context for R_t estimates, we sourced weekly data on regional and national test positivity (percentage positive tests of all tests conducted) from Public Health England (21), available as weekly average percentages from 10th May. From the same source, we also identified the age distributions of cases admitted to hospital and all test-positive cases. Hospital admissions by age were available as age bands with rates per 100,000, so we used regional population data from 2019 (22) to approximate the raw count. We separately sourced daily data on the number of deaths in care homes by region from March, available from 12th April (23). Care homes are defined as supported living facilities (residential homes, nursing homes, rehabilitation units and assisted living units). Data were available by date of notification, which included an average 2-3 day lag after the date of death. We also drew on a database which tracked Covid-19 UK policy updates by date and area (24).

R_t estimation

We estimated R_t using EpiNow2 version 1.2.0, an open-source package in R (13,25,26). This package implements a Bayesian latent variable approach using the probabilistic programming language Stan (27). To initialise the model, infections were imputed prior to the first observed case using a log linear model with priors based on the first week of observed cases. This means that the initial observations both inform the initial parameters and are then also fit, which makes the initial Rt estimates less reliable than later estimates. This was a pragmatic choice to allow the model to be identifiable when only estimating part of the observed epidemic. We explored other parameterisations, but these suffered from poor model identification. For each subsequent time step with observed cases, new infections were imputed using the sum of previous modelled infections weighted by the generation time probability mass function, and combined with an estimate of R_t, to give the prevalence at time t (12). The generation time was assumed to follow a gamma distribution that was fixed over time but varied between samples, with priors drawn from the literature for the mean and standard deviation (28).

These infection trajectories were mapped to reported case counts (D_t) by convolving over an incubation period distribution and report delay distribution (ξ). We assumed a negative binomial observation model for observed reported case counts (C_t), with overdispersion □ using an exponential prior with mean 1 and mean D_t. We combined this with a multiplicative day of the week effect (ω(tmod7)) with an independent effect for each day of the week. We controlled temporal variation using an approximate Gaussian process (29) with a squared exponential kernel (GP).

In mathematical notation:

$$R_t \sim R_{t-1} \times \mathrm{GP}$$

$$I_t = R_t \sum_{\tau} w_\tau I_{t-\tau}$$

$$D_t = \sum_{\tau} \xi_\tau I_{t-\tau}$$

$$C_t \sim \mathrm{NB}(D_t{}^{\omega_{(t \bmod 7)}}, \phi)$$

The length scale and magnitude of the kernel were estimated during model fitting. We used an inverse gamma prior for the length scale, optimising shape and scale values to give a distribution with 98% of the density between 2 and 21 days, and the prior on the magnitude was standard normal. Each region was fitted independently using Markov-chain Monte Carlo (MCMC). Eight chains were used with a warmup of 1,000 samples and 2,000 samples post warmup. Convergence was assessed using the R hat diagnostic.

We used a gamma distributed generation time with mean 3.6 days (standard deviation (SD) 0.7), and SD of 3.1 days (SD 0.8), sourced from (28). Instead of the incubation period used in the original study (which was based on fewer data points), we refitted using a log-normal incubation period with a mean of 5.2 days (SD 1.1) and SD of 1.52 days (SD 1.1)(30). This incubation period was also used to convolve from unobserved infections to unobserved symptom onsets (or a corresponding viral load in asymptomatic cases) in the model. When fitting the model, the time interval distributions had independent priors placed on the mean and standard deviation of their respective log-normal distributions.

We estimated both the delay from onset to positive test (either in the community or in hospital) and the delay from onset to death as log-normal distributions using a subsampled Bayesian bootstrapping approach (with 100 subsamples each using 250 samples) from given data on these delays. Our delay from date of onset to date of positive test (either in the community or in hospital) was taken from a publicly available linelist of international cases (31). We removed countries with outlying delays (Mexico and the Philippines). The resulting delay data had a mean of 4.4 days and standard deviation (SD) 5.6. Delays for hospital admissions and test positives were treated as having the same delay from infection to onset and observation. For the delay from onset to death we used data taken from a large observational UK study (32). We re-extracted the delay from confidential raw data, with a mean delay of 14.3 days (SD 9.5). There was insufficient data available on the various reporting delays to estimate spatially- or temporally-varying delays, so they were considered to be static over the course of the epidemic, although we discuss the effects of this assumption. We have also discussed this approach more extensively in (25).

Comparison of R_t estimates

We compared R_t estimates by data source, plotting each by region over time. To avoid the first epidemic wave obscuring visual differences, all plots were limited to the earliest date that any R_t estimate for England crossed below 1 after the peak. We also identified the time at which each R_t estimate fell below 1, the local minima and maxima of median R_t estimates, and the number of times in the time-series that each R_t estimate crossed its own median, before comparing these across regions and against the total count of the raw data.

We investigated correlations between R_t estimates and the demographic and social context of transmission. We used linear regression to assess whether the level of raw data count influenced oscillations in R_t. We assessed the influence of local outbreaks using test positivity. We used a 5% threshold for positivity as the level at which testing is either insufficient to keep pace with widespread community transmission (33), or where outbreaks have already been detected and tests targeted to those more likely to be positive. We plotted this against raw data and R_t, and also used linear regression to test the association. We interpreted results in light of known outbreaks and policy

changes. We plotted and qualitatively assessed variation in R_t estimates against the age distribution of cases over time, and similarly explored patterns in R_t estimates against the qualitative proportion of cases to all deaths. The latter was not assessed quantitatively due to differences in reference dates (23). With the exception of fitting the delay from onset to death (held confidentially), code and data to reproduce this analysis is available (34).

Results

[Figure 1]

Across England, the Covid-19 epidemic peaked at 4,798 reported test-positive cases (on the 22nd April), 3,099 admissions (1st April), and 975 deaths (8th April) per day (Figure 1A). Following the peak, a declining trend continued for daily counts of admissions and deaths, while daily case counts from all reported test-positive cases increased from July and had more than tripled by August (from 571 on 30th June to 1,929 on 1st September). Regions followed similar patterns over time to national trends. However, in the North East and Yorkshire, Midlands, and North West, incidence of test-positive cases did not decline to near the count of admissions as in other regions, and also saw a small temporary increase during the overall rise in case counts in early August.

[Figure 2]

Following the initial epidemic peak in mid-March, the date at which R_t estimates crossed below 1 varied by both data source and geography (Figure 1B, Figure 2). The first region to cross into a declining epidemic was London, on the 26th March according to an R_t estimated from deaths (where the lower 90% CrI crossed below 1 on the 24th and the upper CrI on the 28th March). However, the data source used to estimate R_t was as important as any regional variation in estimating the earliest date of epidemic decline. R_t estimated from hospital admissions gave the earliest estimate of a declining epidemic, while using all test-positive cases to estimate R_t took the longest time to reach a declining epidemic, in all but one region (East of England). This difference by data source varied by up to 21 days in the North East and Yorkshire, where hospital admissions gave a median R_t estimate under 1 on the 1st April (90%CrIs 31st March, 2nd April), but the median R_t estimate from test-positive cases crossed 1 on only the 22nd April (90%CrIs 1st April, 25th April).

When not undergoing a clear state change, R_t estimates from all data sources oscillates, with oscillations damped when R_t estimates were transitioning to new levels. In England and all NHS regions, test-positive cases showed evidence of larger damped oscillations from July when a state change occurred to R_t over 1. In England, R_t estimates from test-positive cases increased from 0.99 (90%CrI 0.94-1.04) on 30th June to 1.37 (90%CrI 1.31-1.1.44) on 27th August. Meanwhile, the timing and duration of oscillations did not align between R_t estimates (Figure 1B). In some regions, the difference between R_t estimates was consistent over time, such as between R_t from admissions and deaths in the South East. In other regions such as the Midlands this was not the case, with the divergence between the R_t estimates from test-positive cases, admissions, and deaths each varying over time. R_t estimates from test-positive cases were the most likely to differ from estimates derived from other data sources across all regions. Across all regions, R_t estimates from deaths had slower damped oscillations compared to estimates from test-positive cases or hospital admissions. However, oscillations in R_t estimates did not appear to be linked to the level of raw data counts in each source (SI Figure 2).

More rapid oscillations in R_t estimates from test-positive cases appeared to be linked to targeted testing of case clusters, seen in high test positivity (Table SI2). Both the North East and Yorkshire and the Midlands saw more frequent oscillations in R_t estimates from test-positive cases than other regions. The R_t estimates from cases crossed its own median 10 times over the time-series in both regions, while in all other NHS regions this averaged 6 times, and oscillations in R_t estimates from

cases also had a shorter duration in the North East and Yorkshire and the Midlands compared to other regions (Table SI1). Across all regions, 84% of weeks with over 5% positivity (N=19) were in the North East and Yorkshire and the Midlands (Figure 2A). In these regions, positivity peaked on the week of 9th May at 14% and 12% respectively, and overall averaged 6% (95%CI 4.4-7.6%) and 5.9% (95%CI 4.6-7.2%, weeks of 10th May to 22nd August) respectively. High test positivity is likely to have resulted from targeted testing among known local outbreaks in these regions. In the Midlands, these included local restrictions and increased testing across Leicester and in a Luton factory (restrictions between 4th and 25th July (35)). In Yorkshire case clusters were detected with local restrictions in Bradford, Calderdale, and Kirklees (with restrictions from 5 August (36)).

In England, a divergence between R_t estimated from cases versus R_t estimated from deaths and admissions coincided with a decline in the age distribution among all test-positive cases in England to a younger population (Figure SI2A). From mid-April to June, national estimates of R_t from test-positive cases remained around the same level as those from admissions or deaths, while after this, cases diverged to a higher steady state (Figure 1A). On the 23rd May, the median R_t estimated from cases matched that of deaths at 0.83 (both with 90% CrIs 0.78-0.89), but this was followed by a 78 day period before the two estimates were again comparable, on 8th August. Over this period the median R_t estimate from cases was on average 14% higher (95%CI 12-15%). Meanwhile, the share of test-positive cases under age 50 increased from under one-quarter of cases in the week of 28th March (24%, N=16,185), to accounting for nearly three-quarters of cases by 22nd August (77%, N=6,733). While the percentage of test-positive cases aged 20-49 increased consistently from April to August, the 0-19 age group experienced a rapid increase over mid-May through July, increasing by a mean 1% each week over May 9th through August 1st (from 4% of 18,774 cases to 14.8% of 5,017 cases).

Similarly, R_t estimates from admissions in England oscillated over June through July, potentially linked to the age distribution of hospital admissions. From 0.92 (90%CrI 0.87-0.98) on the 11th June, R_t estimated from admissions fell to 0.8 (90%CI 0.75-0.85)) on the 27th June. In contrast, this transition was not observed in the R_t estimate based on test-positive cases (Figure 1A). Older age groups dominated Covid-19 hospital admissions, where 0-44 years never accounted for more than 12.8% of hospital-based cases (a maximum in the week of 22nd August, N=690; Figure SI2B). While the proportion of hospital admissions aged 75+ remained steady over May through mid-June, this proportion appeared to oscillate over July through August (standard deviation of weekly percentage at 6.1 over June-August, compared to 5.4 in months March-May). These variations were not seen in the proportion aged 70+ in the test-positive case data, which saw a continuous decline from 30% at the start of June to 7% by August.

R_t estimated from either admissions or deaths experienced near-synchronous local peaks across regions over April and May. We compared this R_t estimated from deaths with its source data and a separate regional dataset of deaths in care homes. In the South East and South West, the R_t estimates from deaths rose over April, with a peak in early May. In the South West, the median R_t estimate from deaths increased by 0.04 from 22nd April to 7th May (from 0.8 (90%CrI 0.72-0.88) to 0.84 (90%CrI 0.76-0.95)); and by 0.06 from the 17th April to 4th May in the South East (from 0.82 (90%CrI 0.77-0.9) to 0.88 (90%CrI 0.72-0.88)). In both these regions, this early May peak in R_t estimates from deaths coincided with similarly rising R_t estimates from hospital admissions, while the reverse trend was seen in R_t estimates from cases. In all regions, care home deaths peaked over the 22nd-29th April (by date of notification; Figure SI3). This was later than regional peaks in the raw count of all deaths in any setting (which peaked between the 8th-16th April, by date of death), even accounting for a 2-3 day reporting lag. This meant that the proportion of deaths from care homes varied over time, where in the South East and South West, deaths in care homes appeared to account for nearly all deaths for at least the period mid-May to July.

Discussion

We estimated the time-varying reproduction number for Covid-19 over March through August across England and English NHS regions, using test-positive cases, hospital admissions, and deaths with confirmed Covid-19. Our estimates of transmission potential varied for each of these sources of infections, and the divergence between estimates from each data source was not consistent within or across regions over time, although estimates based on hospital admissions and deaths were more spatio-temporally synchronous than compared to estimates from cases. We compared differences in R_t estimates to the extent and context of transmission and found that the difference between R_t estimated from cases, admissions, and deaths may be linked to uneven rates of testing, the changing age distribution of cases, and outbreaks in care home populations.

R_t estimates varied by data source, and the extent of variation itself differed by region and over time. Following the initial epidemic peak in mid-March, the date at which R_t estimates crossed below 1 varied by both data source and geography, following which R_t estimates from all data sources varied when not undergoing a clear state change. The differences in these oscillations by data source may indicate different underlying causes. This implies that each data source was influenced differently by changes in subpopulations over time.

Increasingly rapid oscillations in R_t estimates from test-positive cases were associated with higher test-positivity rates. Increasing test-positivity rates could be an indication of inconsistent community testing, with the observation of an initial rise in transmission amplified by expanded testing and local interventions where a cluster of new, mild cases has been identified (18). This targeted testing may drive regionally localised instability in case detection and resulting R_t estimates but may not reflect changes in underlying transmission. This is a limitation of monitoring epidemic dynamics using test-positive surveillance data in areas where testing rates vary across the population and over time. This also suggests that R_t estimates from admissions may be more reliable than that from all test-positive cases for indicating the relative intensity of an epidemic over time (37).

We hypothesised that variations in R_t estimates were also related to changes in the age distribution of cases over time, because age is associated with severity (38,39). If each data source represented a different sample of this age-severity gradient, and transmission also varied by age or severity, R_t estimates from each source would diverge. Early in the epidemic, tests were largely limited to hospital settings, and disproportionately represented healthcare workers compared to the general population. This sampling bias would be reflected in the R_t from test-positive cases. The early peak in R_t could then represent a substantial separate route of transmission in healthcare settings, in a wave of nosocomial infections (40). If healthcare workers were less susceptible to severe disease than those older than working age, an early peak in R_t estimated from test-positive cases would not have been represented in R_t estimated from hospital admissions or deaths. Meanwhile, either hospital admissions or deaths data would be more representative of sampling a separate route of transmission among the general population. If infections spread through the general population later than nosocomial infections, then the timing of peaks in R_t estimates from each data source would not have matched.

From late spring, outbreaks in care homes may have contributed to a divergence between Rt estimates from test positive cases and other data sources. All regions saw a near-synchronous local peak in R_t estimated from hospital admissions over spring, which was not seen in R_t estimated from test-positive cases. This may have reflected the known widespread regional outbreaks in care homes. The care home population is on average older and more clinically vulnerable than the general population, while also being less likely to appear for community testing (41,42). Increased transmission in care homes would then be seen in an increased R_t from hospital admissions, but not observed in an R_t from test-positive cases.

Similarly, the age-severity gradient may have impacted transmission estimates later in the epidemic when community testing became more widely available. We found that from June onwards, R_t

estimates from all test-positive cases appeared to increasingly diverge away from R_t estimates from admissions and deaths, transitioning into a separate, higher, steady state. This was followed by the observed age distribution of all test-positive cases becoming increasingly younger, while the age distribution of admissions remained approximately level. Because of the severity gradient, this suggested the R_t estimates from all test-positive cases and admissions were more biased by the relative proportion of younger cases and older cases respectively than the R_t estimates from admissions or deaths.

Our analysis was limited where data or modelling assumptions did not reflect underlying differences in transmission. R_t estimates can become increasingly uncertain and unstable with lower case counts. Further, estimated unobserved infections were mapped to reported cases or deaths using two delay distributions: the time from infection to test in the community or hospital, and a longer delay from infection to death. Mis-specification of the priors would have created bias in the temporal distribution of all resulting R_t estimates, with estimated dates of infection and R_t incorrectly shifted too much or little in time compared to the true infection curve, and decreased accuracy of R_t estimates (43).

We used the same distribution priors for both delays after symptom onset to positive test, and to hospital admission. This may be inaccurate where cases with mild symptoms take longer to present for testing than severe cases presenting for hospital admission, or vice versa. The difference between the two delays over time may also have varied, with a possible decrease in delay to reported tests when mass community testing became available over the summer. This would have had a differential impact on the accuracy of R_t estimates over time in either direction, which could explain some of the oscillations in R_t estimates from test-positive case data compared to hospital admissions. We had no data over time on delays from symptom onset to reporting in each data source with which to test this hypothesis. However, we have mitigated some of the impact of this by using a sub-sampled bootstrap of the available delay data when estimating the delay distribution priors. This inflated the uncertainty of these priors in line with the hypothesis that they varied over time. This adjustment may be conservative if the delay distributions are stable over time.

Spatial dependence in delay distributions may also have contributed to their mis-specification and increased uncertainty in R_t estimates. We observed that the variation in R_t estimates from admissions and deaths often showed comparable levels and patterns in oscillations over time but were out of phase with each other. This may have been due to using data sources from different populations for each delay estimate. To estimate the delay between symptom onset to either a positive test or hospitalisation, we used a linelist of all patients publicly reported globally, which had a mean delay 5.4 days (SD 5.6). This varied only slightly from an early estimate in the UK epidemic, where the delay from onset to hospitalisation had a mean 5.14 days (SD 4.2) in confidential Public Health England (FF100) data (44). Meanwhile, the same global public linelist contained few records with delay from onset to death, with mean 11.4 (SD 16.5). We compared this to confidential UK data from an observational study which had mean delay 14.3 days (SD 9.5) (32).

Comparing each type and source of delay, we judged the benefits of using open data to outweigh the minor observed spatial variation of the delay from onset to test or admission, although at the expense of increased uncertainty. However, we judged the difference in delay from onset to death in the UK compared to public (international) data was sufficiently meaningful to justify using confidential UK data in order to maintain accuracy of the R_t estimate from deaths. The difference in geographic source of delay distributions should not have substantially altered our conclusions about discrepancies between central estimates of R_t from either test-positives or admissions, compared to R_t estimated from deaths. However, using the international public linelist for the delay to test or admission may have introduced additional uncertainty around the respective R_t estimates, compared to greater accuracy (reduced uncertainty) in estimates of R_t from deaths based on a UK-specific delay distribution.

The data sources themselves may also have been inaccurate or biased, which would change the representation of the population we have assumed here. For example, we excluded data from other nations of the UK (Wales, Scotland and Northern Ireland) in our analysis, as these differed in both availability over time and in data collection and reporting practices (19,45). English regional data may also contain bias where new parts of the population might be under focus for testing efforts, or the population characteristics of hospital admissions from Covid-19 may have changed over time with changes in clinical criteria or hospital capacity for admission. This would mean that an R_t estimate from these data sources would represent different source populations over time, limiting our ability to reliably compare against R_t estimates from other data sources. Where possible we highlighted this by comparing R_t estimates to known biases and changes in case detection and reporting.

Our approach is unable to make strong causal conclusions about varying transmission, and assumptions about sampling and the representation of subpopulations remain implicit. Alternatively, varying epidemics in subpopulations could have been addressed with mechanistic models that explicitly consider transmission in different settings and are fitted to multiple data sources. However, these require additional assumptions, detailed data to parameterise, and may be time-consuming to develop. In the absence of data, the number of assumptions required for these models can introduce inherent structural biases. Our approach contains few structural assumptions and therefore may be more robust when data are sparse, or information is required in real-time.

We conclude that when estimating R_t, the choice of data source should be guided by the policy context in which the estimates will be used and interpreted. This work highlights that there is no clear superior choice of data source, while R_t estimates are sensitive to assumptions about the underlying population of each data source. This means that both producers and users of R_t estimates should understand relevant biases in the data source's population sampling strategy, such as by community case detection or patient severity, before drawing conclusions about transmission in the population as a whole.

We also recommend presenting concurrent R_t estimates jointly, rather than pooling estimates of R_t from different data sources. Pooling estimates would both suffer from unclear weighting and lose useful information about variation in subpopulation transmission. Although the reconstruction of the underlying transmission process from the reporting processes is robust, it is unclear how weights would be assigned based on likelihood to estimates from different data sources. Further, the variation in concurrent R_t estimates provides more information about population transmission than any single estimate, when considered in light of the sampling biases of each data source. This additional information can be useful to identify transmission intensity by subpopulation where access to high quality disaggregated data may not be available in real time. While this can be difficult to interpret without specific knowledge of population structure and dynamics, this information would be lost altogether in a single or pooled estimate of R_t. In contrast, if policy were to be based on either a single or an averaged R_t estimate, it would be unclear what any recommendation should be and for whom.

Future work could explore systematic differences in the influence of data sources on R_t estimate by extending the comparison of R_t by data source to other countries or infectious diseases. Additionally, work should also clarify the potential for comparing R_t estimates in real-time tracking of outbreaks and explore the inconsistencies in case detection over time and space, where a cluster of cases leads to a highly localised expansion of community testing, creating an uneven spatial bias in transmission estimates. These findings may be used to improve R_t estimation and identify findings of use for epidemic control. Based on the work presented here we now provide R_t estimates, updated each day, for test positive cases, admissions, and deaths in each NHS region and in England. Our estimates are visualised on our website, are available for download, and are produced using publicly accessible code (46,47).

Tracking differences by data source can improve understanding of variation in testing bias in data collection, highlight outbreaks in new subpopulations and indicate differential rates of transmission among vulnerable populations, and clarify the strengths and limitations of each data source. Our approach can quickly identify such patterns in developing epidemics that might require further investigation and early policy intervention. Our method is simple to deploy and scale over time and space using existing open-source tools, and all code and estimates used in this work are available to be used or re-purposed by others.

Acknowledgements

The following authors were part of the Centre for Mathematical Modelling of Infectious Disease COVID-19 Working Group. Each contributed in processing, cleaning and interpretation of data, interpreted findings, contributed to the manuscript, and approved the work for publication: Fiona Yueqian Sun, C Julian Villabona-Arenas, Emily S Nightingale, Alicia Showering, Gwenan M Knight, Yang Liu, Kaja Abbas, Akira Endo, Alicia Rosello, Rachel Lowe, Matthew Quaife, Amy Gimma, Oliver Brady, Nicholas G. Davies, Anna Vassal, W John Edmunds, Jack Williams, Simon R Procter, Rosalind M Eggo, Yung-Wai Desmond Chan, Rosanna C Barnard, Georgia R Gore-Langton, Naomi R Waterlow, Charlie Diamond, Timothy W Russell, Graham Medley, Katherine E. Atkins, Kiesha Prem, David Simons, Megan Auzenbergs, Damien C Tully, Christopher I Jarvis, Kevin van Zandvoort,Carl A B Pearson, Thibaut Jombart, Anna M Foss, Adam J Kucharski, Billy J Quilty, Hamish P Gibbs, Samuel Clifford, Petra Klepac.

Funding

The following funding sources are acknowledged as providing funding for the named authors. Wellcome Trust (210758/Z/18/Z: JDM, JH, KS, NIB, SA, SFunk, SRM). This research was partly funded by the Bill & Melinda Gates Foundation (INV-003174: MJ). This project has received funding from the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688: MJ). The following funding sources are acknowledged as providing funding for the working group authors. Alan Turing Institute (AE). BBSRC LIDP (BB/M009513/1: DS). This research was partly funded by the Bill & Melinda Gates Foundation (INV-001754: MQ; INV-003174: KP, MJ, YL; NTD Modelling Consortium OPP1184344: CABP, GFM; OPP1180644: SRP; OPP1183986: ESN; OPP1191821: KO'R, MA). BMGF (OPP1157270: KA). Foreign, Commonwealth and Development Office (FCDO)/Wellcome Trust (Epidemic Preparedness Coronavirus research programme 221303/Z/20/Z: CABP, KvZ). DTRA (HDTRA1-18-1-0051: JWR). Elrha R2HC/UK FCDO/Wellcome Trust/This research was partly funded by the National Institute for Health Research (NIHR) using UK aid from the UK Government to support global health research. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care (KvZ). ERC Starting Grant (#757699: JCE, MQ, RMGJH). This project has received funding from the European Union's Horizon 2020 research and innovation programme - project EpiPose (101003688: KP, MJ, PK, RCB, WJE, YL). This research was partly funded by the Global Challenges Research Fund (GCRF) project 'RECAP' managed through RCUK and ESRC (ES/P010873/1: AG, CIJ, TJ). HDR UK (MR/S003975/1: RME). MRC (MR/N013638/1: NRW). Nakajima Foundation (AE). NIHR (16/136/46: BJQ; 16/137/109: BJQ, CD, FYS, MJ, YL; Health Protection Research Unit for Immunisation NIHR200929: NGD; Health Protection Research Unit for Modelling Methodology HPRU-2012-10096: TJ; NIHR200929: FGS, MJ; PR-OD-1017-20002: AR, WJE). Royal Society (Dorothy Hodgkin Fellowship: RL; RP\EA\180004: PK). UK DHSC/UK Aid/NIHR (ITCRZ 03010: HPG). UK MRC (LID DTP MR/N013638/1: GRGL, QJL; MC_PC_19065: AG, NGD, RME, SC, TJ, WJE, YL; MR/P014658/1: GMK). Authors of this research receive funding from UK Public Health Rapid Support Team funded by the United Kingdom Department of Health and Social Care (TJ). Wellcome Trust (206250/Z/17/Z: AJK, TWR; 206471/Z/17/Z: OJB; 208812/Z/17/Z: SC; 208812/Z/17/Z: SFlasche). No funding (AKD, AMF, AS, CJVA, DCT, JW, KEA, SH, YJ, YWDC)

References

1. European Centre for Disease Prevention and Control. COVID-19 situation update worldwide, as of 6 June 2020 [Internet]. 2020. Available from: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases
2. World Health Organisation. Strengthening and adjusting public health measures throughout the COVID-19 transition phases. Policy considerations for the WHO European Region [Internet]. WHO Regional Office for Europe; 2020 May. Available from: http://www.euro.who.int/en/countries/hungary/publications/strengthening-and-adjusting-public-health-measures-throughout-the-covid-19-transition-phases.-policy-considerations-for-the-who-european-region,-24-april-2020
3. HM Government. Our Plan to Rebuild: The UK Government's COVID-19 recovery strategy [Internet]. 2020 May. (CP:239). Available from: https://www.gov.uk/government/publications/our-plan-to-rebuild-the-uk-governments-covid-19-recovery-strategy
4. Michael Parker. Ethics and value judgements involved in developing policy for lifting physical distancing measures [Internet]. 2020 Apr. (SAGE 30). Available from: https://www.gov.uk/government/publications/ethics-and-value-judgements-involved-in-developing-policy-for-lifting-physical-distancing-measures-29-april-2020
5. Thompson RN. Epidemiological models are important tools for guiding COVID-19 interventions. BMC Med. 2020 May 25;18(1):152.
6. Pitzer VE, Chitwood M, Havumaki J, Menzies NA, Perniciaro S, Warren JL, et al. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. medRxiv [Internet]. 2020 Apr;2020.04.20.20073338. Available from: DOI: 10.1101/2020.04.20.20073338
7. The Royal Society. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK. 2020 Aug 24; Available from: https://royalsociety.org/-/media/policy/projects/set-c/set-covid-19-R-estimates.pdf
8. Scientific Advisory Group for Emergencies. Scientific evidence supporting the government response to coronavirus (COVID-19) [Internet]. 2020. Available from: https://www.gov.uk/government/collections/scientific-evidence-supporting-the-government-response-to-coronavirus-covid-19
9. Funk S, Abbott S, Atkins B, Baguelin M, Baillie J, Birrell P, et al. Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. medRxiv. 2020 Jan 1;2020.11.11.20220962.

10. Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. Am J Epidemiol. 2004;160(6):509–16.
11. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci. 2007 Feb 22;274(1609):599–604.
12. Cori A, Ferguson NM, Fraser C, Cauchemez S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. Am J Epidemiol. 2013;178(9):1505–12.
13. Abbott S, Hellewell J, Sherratt K, Gostic K, Hickson J, Badr HS, et al. EpiNow2: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters [Internet]. 2020. Available from: DOI: 10.5281/zenodo.3957489
14. Keeling MJ, Dyson L, Guyver-Fletcher G, Holmes A, Semple MG, Investigators I, et al. Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number. medRxiv. 2020 Sep 29;2020.08.04.20163782.
15. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis. 2020 May;20(5):553–8.
16. Cori A, Donnelly CA, Dorigatti I, Ferguson NM, Fraser C, Garske T, et al. Key data for outbreak evaluation: building on the Ebola experience. Philos Trans R Soc B Biol Sci. 2017 May 26;372(1721):20160371.
17. Gostic KM, McGough L, Baskerville E, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. medRxiv. 2020 Jan 1;2020.06.18.20134858.
18. Department of Health and Social Care. Coronavirus (COVID-19) - Scaling up our testing programmes [Internet]. 2020 Apr. Available from: https://www.gov.uk/government/publications/coronavirus-covid-19-scaling-up-testing-programmes
19. Public Health England, NHSX. Coronavirus (COVID-19) in the UK [Internet]. 2020. Available from: https://coronavirus.data.gov.uk/about-data
20. Abbott S, Sherratt K, Bevan J, Gibbs H, Hellewell J, Munday J, et al. covidregionaldata: Subnational Data for the Covid-19 Outbreak [Internet]. 2020. Available from: DOI: 10.5281/zenodo.3957539
21. Public Health England. National COVID-19 surveillance reports [Internet]. GOV.UK; 2020. Available from: https://www.gov.uk/government/publications/national-covid-19-surveillance-reports
22. Office for National Statistics. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland
23. Office for National Statistics. Number of deaths in care homes notified to the Care Quality Commission, England [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/numberofdeathsincarehomesnotifiedtothecarequalitycommissionengland
24. The Health Foundation. COVID-19 policy tracker [Internet]. A timeline of national policy and health system responses to COVID-19 in England. 2020. Available from: https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker
25. Abbott S, Hellewell J, Hickson J, Munday J, Gostic K, Ellis P, et al. EpiNow2 v1.2.0: Estimate Real-Time Case Counts and Time-Varying Epidemiological Parameters [Internet]. 2020. Available from: DOI: 10.5281/zenodo.4088545
26. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: http://www.R-project.org/
27. Stan Development Team. RStan: the R interface to Stan [Internet]. 2020. Available from: http://mc-stan.org/
28. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance. 2020;25(17).
29. Riutort-Mayol G, Bürkner P-C, Andersen MR, Solin A, Vehtari A. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. ArXiv Prepr [Internet]. 2020 Apr 23;arXiv:2004.11408. Available from: https://arxiv.org/abs/2004.11408v1
30. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med [Internet]. 2020; Available from: https://doi.org/10.7326/M20-0504
31. Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data. 2020 Mar;7(1):106.
32. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 16,749 hospitalised UK patients with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol. medRxiv [Internet]. 2020 Apr 28; Available from: http://medrxiv.org/lookup/doi/10.1101/2020.04.23.20076042
33. World Health Organisation. Considerations in adjusting public health and social measures in the context of COVID-19 [Internet]. World Health Organisation; 2020 May. Report No.: WHO/2019-nCoV/Adjusting_PH_measures/2020. Available from: https://www.who.int/publications-detail-redirect/public-health-criteria-to-adjust-public-health-and-social-measures-in-the-context-of-covid-19
34. Sherratt K, Abbott S. Rt comparison by data source in the UK [Internet]. GitHub; 2020. (GitHub). Available from: https://github.com/epiforecasts/rt-comparison-uk-public
35. Acts of Parliament. The Health Protection (Coronavirus, Restrictions) (Leicester) Regulations 2020 [Internet]. 2020 No. 685 Jul 4, 2020. Available from: https://www.legislation.gov.uk/uksi/2020/685/
36. Acts of Parliament. The Health Protection (Coronavirus, Restrictions on Gatherings) (North of England) Regulations 2020 [Internet]. 2020 No. 828 Aug 5, 2020. Available from: https://www.legislation.gov.uk/uksi/2020/828
37. Smith DR, Duval A, Pouwels KB, Guillemot D, Fernandes J, Huynh B-T, et al. How best to use limited tests? Improving COVID-19 surveillance in long-term care [Internet]. 2020 Apr. (Scientific Advisory Group for Emergencies (SAGE)). Available from: http://medrxiv.org/lookup/doi/10.1101/2020.04.19.20071639
38. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis. 2020 Jun 1;20(6):669–77.
39. Levin AT, Meyerowitz-Katz G, Owusu-Boaitey N, Cochran KB, Walsh SP. Assessing the age specificity of infection fatality rates for Covid-19: Systematic review, meta-analysis, and public policy implications. medRxiv [Internet]. 2020 Jul 24; Available from: http://medrxiv.org/lookup/doi/10.1101/2020.07.23.20160895
40. Evans S, Agnew E, Vynnycky E, Robotham J. The impact of testing and infection prevention and control strategies on within-hospital transmission dynamics of COVID-19 in English hospitals. medRxiv [Internet]. 2020 May 20; Available from: https://www.medrxiv.org/content/10.1101/2020.05.12.20095562v2

41. Gordon AL, Goodman C, Achterberg W, Barker RO, Burns E, Hanratty B, et al. Commentary: COVID in care homes—challenges and dilemmas in healthcare delivery. Age Ageing. 2020 Aug 24;49(5):701–5.
42. Scientific Advisory Group for Emergencies. SAGE 33 minutes: Coronavirus (COVID-19) response, 5 May 2020 [Internet]. 2020 May. Available from: https://www.gov.uk/government/publications/sage-minutes-coronavirus-covid-19-response-5-may-2020
43. Guenther F, Bender A, Katz K, Kuechenhoff H, Hoehle M. Nowcasting the COVID-19 Pandemic in Bavaria. medRxiv [Internet]. 2020 Jun 28; Available from: http://medrxiv.org/lookup/doi/10.1101/2020.06.26.20140210
44. Pellis L, Scarabel F, Stage HB, Overton CE, Chappell LHK, Lythgoe KA, et al. Challenges in control of Covid-19: short doubling time and long delay to effect of interventions. Eprint ArXiv200400117 Q-Bio [Internet]. 2020 Mar 31; Available from: http://arxiv.org/abs/2004.00117
45. Department of Health and Social Care. COVID-19 testing data: methodology note [Internet]. 2020. Available from: https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note
46. Sam Abbott, Hickson J, Peter Ellis, Hamada S. Badr, Munday J, Jamie Allen, et al. National and subnational estimates of the time-varying reproduction number for Covid-19 [Internet]. 2020. Available from: https://github.com/epiforecasts/covid-rt-estimates
47. Sam Abbott, Joe Hickson, Peter Ellis, Hamada S. Badr, Jamie Allen, JD Munday, et al. Covid-19: National and Subnational estimates for the United Kingdom [Internet]. Epiforecasts. 2020. Available from: https://epiforecasts.io/covid/posts/national/united-kingdom/

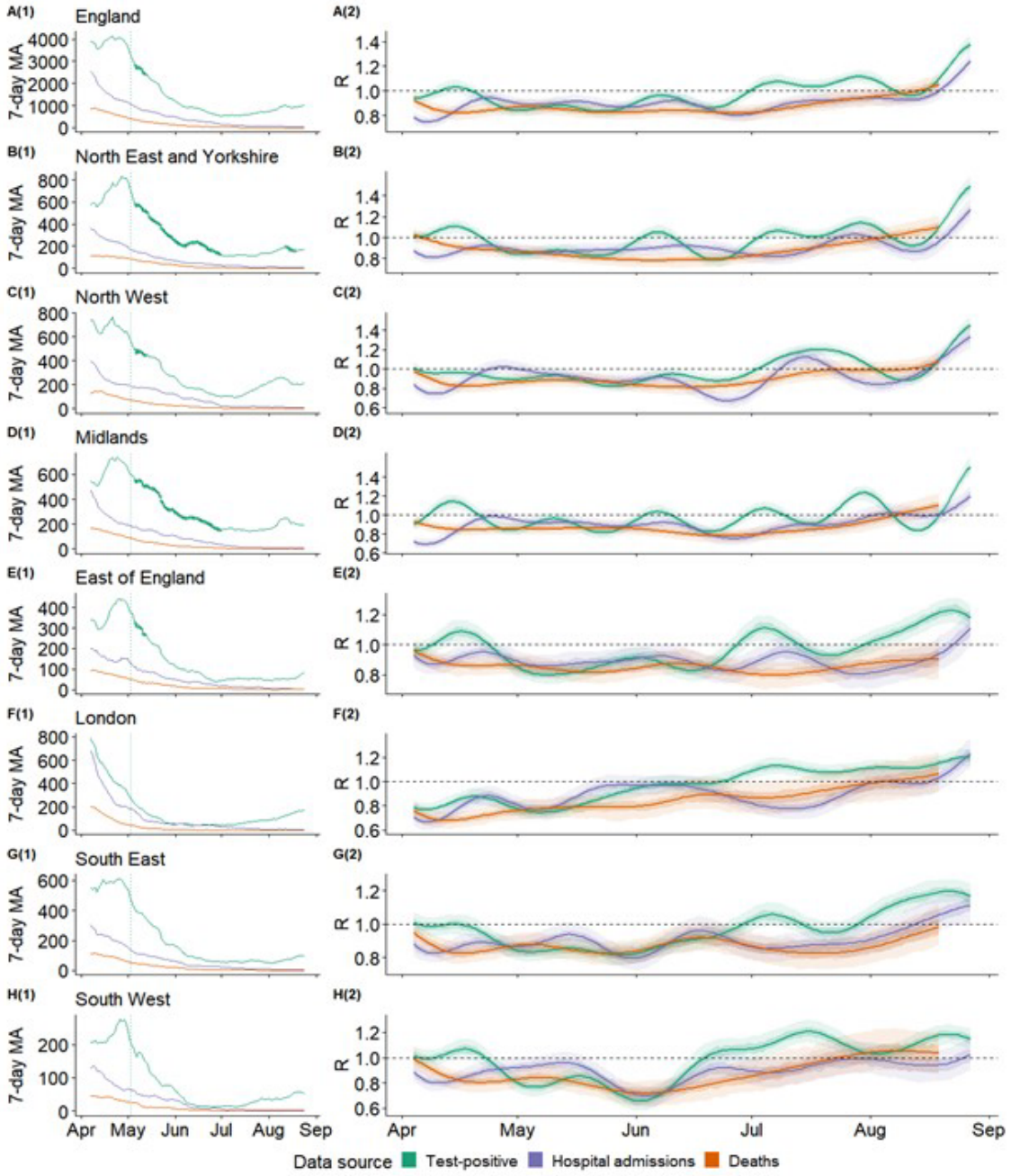

Figure 1. Epidemic dynamics across **(A)** England and **(B-H)** seven English National Health Service regions, 5th April through 27th August 2020. **A1-H1**: Daily counts of confirmed cases by data source, as centred seven day moving average. Counts marked with crosses indicate dates within weeks which averaged >5% test positivity (positive / all tests per week). Vertical dotted line indicates the start of national mass community testing on 3rd May. **A2-H2**: Estimates of Rt, (median, with 50% (darker shade) and 90% (lightest shade) credible interval), derived from each data source. Data sources include all test-positive cases, hospital admissions, and deaths with a positive test in the previous 28 days.

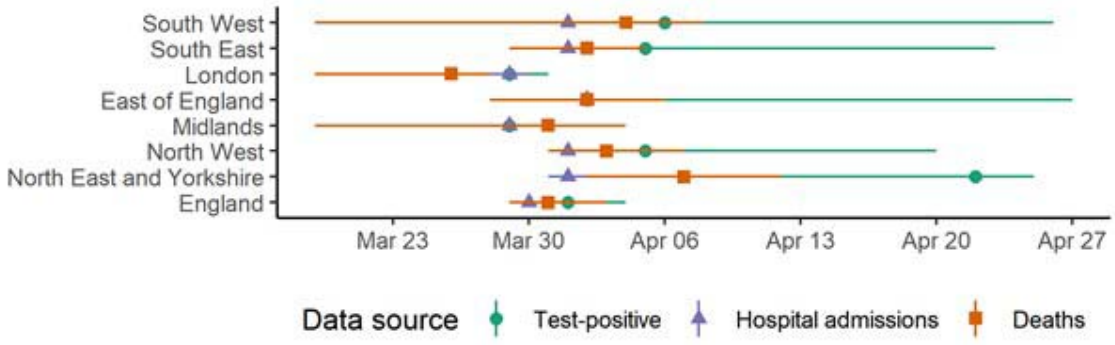

Figure 2. Dates on which Rt estimate crossed 1 after first epidemic peak, median and 90% credible interval, by data source for England and seven NHS regions.